P3C: Technical Report

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

STUDENT GROUP

Group 8

Muchen **He**	44638154	*mhe@ece.ubc.ca*
Kaseya **Xia**	27553304	*zxia@uvm.edu*
Beibei **Xiong**	13233747	*beibei971030@gmail.com*

INSTRUCTOR

Dr. Sidney Fels

Electrical and Computer Engineering, The University of British Columbia

Revision 1.3 — April 20, 2021

THE UNIVERSITY OF BRITISH COLUMBIA

Electrical and Computer Engineering

Contents

List of Figures

Preface

This document is intended for the instructor or students of the human-computer interface course CPEN 541 to learn and possibly reproduce the experimental setup and results. The document outlines in detail the research questions we are pursing, as well as the motivation. This document also describe the prototype and its experiments we developed to test our research questions along with our findings. Lastly, this document features the limitations as well as the future work (including what the prototype would ideally look like).

Please also refer to our other submitted work, including our conference paper, demonstrative video, and presentation slides.

We can be contacted via our emails (listed at the top of the document). All the code for this project can be found on GitHub: https://github.com/CPEN541-FutureGazer.

1 Introduction and Problem

Over the last year, we all observed and experienced first-hand at attempting to scavenge the productivity and work-ethics we once had while working from home while in a pandemic. And one of the most important work-related ritual is none other than meetings and hangouts.

As limited by the restriction and the isolation to curb the infectious virus, we sacrificed the ability to meet with people face-to-face (F2F) — whether that be getting work done, attending classes, discussing and brainstorming ideas, and even asking for help and or assistance from peers. Out of necessity, we all forced to use online web-video-conferencing (WVC) software such as Zoom, Skype, etc. as a substitute.

Despite these companies touting the technological advancements, the intuitive of its user-interfaces, the affordable cost of "free" to its users, they all had this one *glaring* (pun-intended) problem that none of these software truly provided in the absence of F2F meetings: that **eye-contact** and **gaze** are an essential part of body language and non-verbal communications in social situations.

Conventional WVC software evolved from 1-on-1 meetings, where the only other person on the screen is the one you talk to, like in a phone call. This has the few problems of modern, large meetings involving more people, since the other person's face is the only thing you look at and thus easier to make eye-contact and attention.

However, seen in Figure 1(a), as we add more participants to the meeting, instead of being an intimate experience, we are often presented with a gallery, somewhat like a bookshelf view of all the participants. Each participant is placed in a uniform cell as part of a grid. One could describe it as a jail-cell, lacking in connection, organicity, and coherency. Furthermore, because everyone is looking at their own screen, towards their camera, we get a mostly static view of all the participants looking out of the screen on the receiving end. Somewhat akin to have a large audience all staring at you — even if you're one among them passively listening.

2 Background and Prior Works

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.[1] Some also suggest WVC provides higher satisfaction scores than other tools and has become one of the most popular online teaching methods.[2,3]

However, in the study by,[4] 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.[5] Lastly, a study observes an increase in participants' pro-social behaviour when being watched by deceptive video conferencing manipulation.[6]

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.[7–9] 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.[10]

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.[11,12] Eye contact plays an important role in both in person and a WVC system.[13,14] 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.[12] 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[15] 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.[16,17] 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.[18–22] Those methods usually require a number 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.[23] 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.[24] 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[25] 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.[26] 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.[2627] found that without eye contact, 88% of the participants indicated they had trouble perceiving whom their partners were talking to.

3 Research Questions and Scope

This motivated us to explore alternative ways to represent participants in WVC applications to increase engagement, interactivity, and attention.

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.

We are set out to build a prototype WVC application to study the perceived effects of eye-contact, gaze, and head orientation in a virtual 3D space to make up the missing aspects from in-person F2F interactions.

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.

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.[28]

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.

The key metrics we want to observe in this project are: participant's attention, engagement, and the feeling of connectedness. 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**):

1. Can a person tell if they are being looked at in a WVC and how can 3D avatars be augmented to

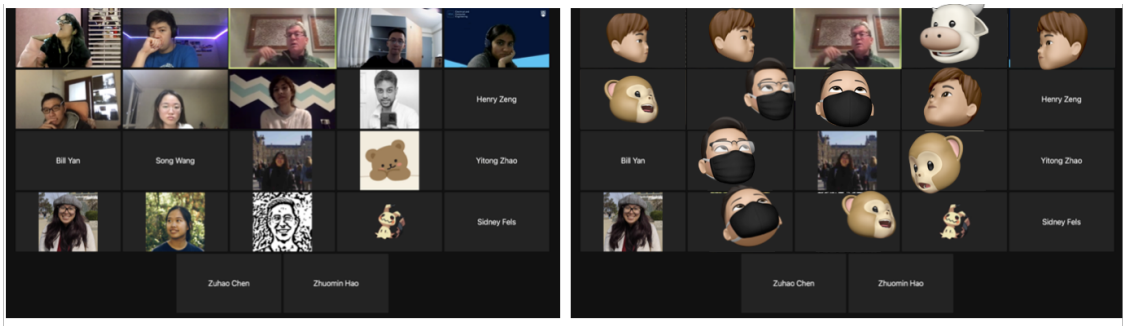


Figure 1: (a) Current WVC model: each participant stays in their grid and no eye contact interaction. (b) Proposed model: students can look at each other to create virtual eye-contact.

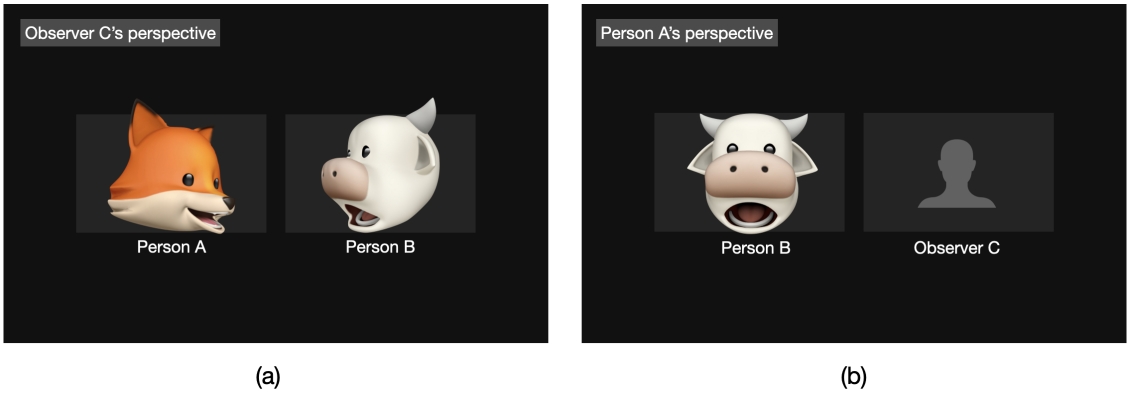


Figure 2: (a) Observer's perspective in a meeting room with person A (fox) and person B (cow) looking at each other. (b) Person A's perspective in the exact same meeting room at the exact same time.

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.

4 The Prototype

This section describes the technical details regards to our implementation. We first outline the language and framework this prototype is developed on, then we describe the rendering, layout, and orientation calculation. Lastly we go over the configuration files, events, timed and randomized sequences, and integrated questionaries that aid us in user-experimentation.

Note: Since this project is not mainly focused on the usability study of this concept, we only implemented the minimum-viable-product to carry-out the user experiments due to time-constraints. We discuss how this prototype could be improved in future>prototype section.

4.1 Platform

The prototype is mostly developed in Processing[29] (a Java-based graphic-centric language mostly used in education, prototype, and visual arts) and Unity Engine (a popular free-to-use game engine). The Unity engine wraps around the Processing prototype to provide user-testing interface such as integrated questionaries and videos.

We chose Processing and Unity as they have strong support of 3D environment, video and audio playback, robust mouse input, ease-of-use, and quick prototype turn-around overheads. Both frameworks are also cross-platform and can execute on Windows, macOS, and Linux. However, as we also discuss in Section limitations section, recent macOS update added a *Gatekeeper* security feature that prevents un-sighed software from running[30] — thus complicating the user-testing process, as distributed prototype executables to the participants cannot be opened.

4.2 3D Environment

Figure 3 shows a typical 3D environment rendered in the prototype app. We base the user interface (UI) design — including colours, buttons, and layout — from existing WVC apps such as Zoom to present the participants with a familiar user experience. By doing so, we limit other factors that would interfere with our user experiments. However unlike traditional WVC apps, we replace the centre (where typically there would be a gallery or grid view of camera video feeds) with our prototype avatar representation.

From hereafter, we will refer to these visual representation of participants in the meeting as *Avatar Views*. As outlined earlier in 3, we propose two types of avatars to explore: 3D head avatars and 2D eye avatars.

For user experiments, we create *mock avatars* to replicate/simulate real meeting participants. This creates an illusion for the test participants as if they're real people to help us speed up the testing process without involving lots of people. However, as it is an illusion, it might affect how people perceive our prototype — as we discuss in Section 8.

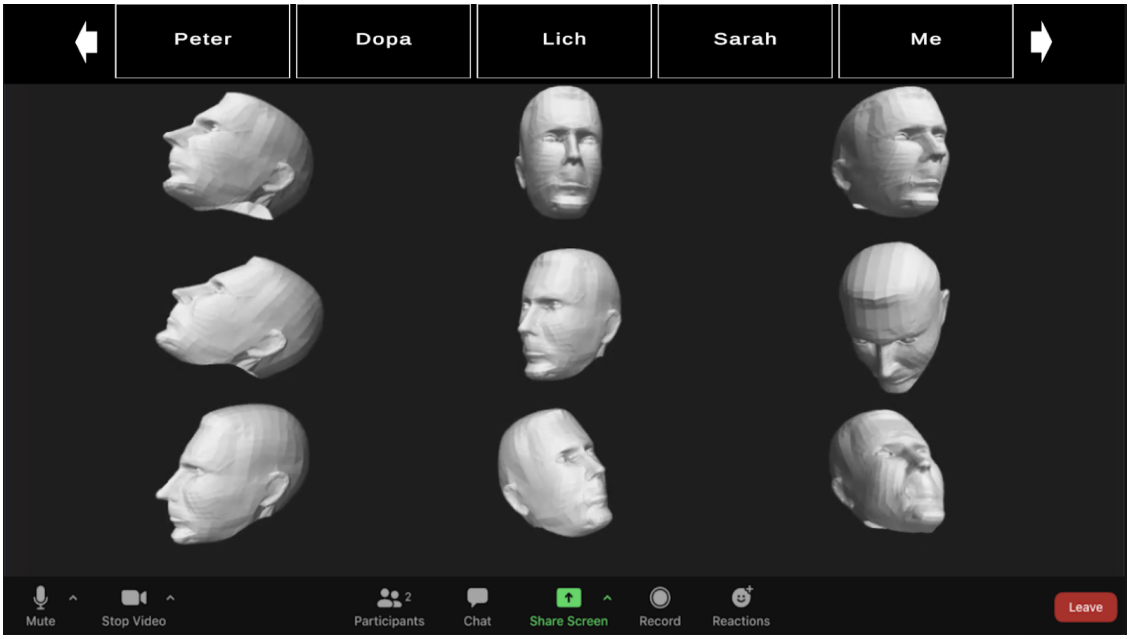


Figure 3: A typical view of the FutureGazer prototype app with the Unity interface

(a) 2D eyes avatar (b) 3D head avatar

Figure 4: 2D eyes and 3D heads are the two types of avatars in FutureGazer prototype

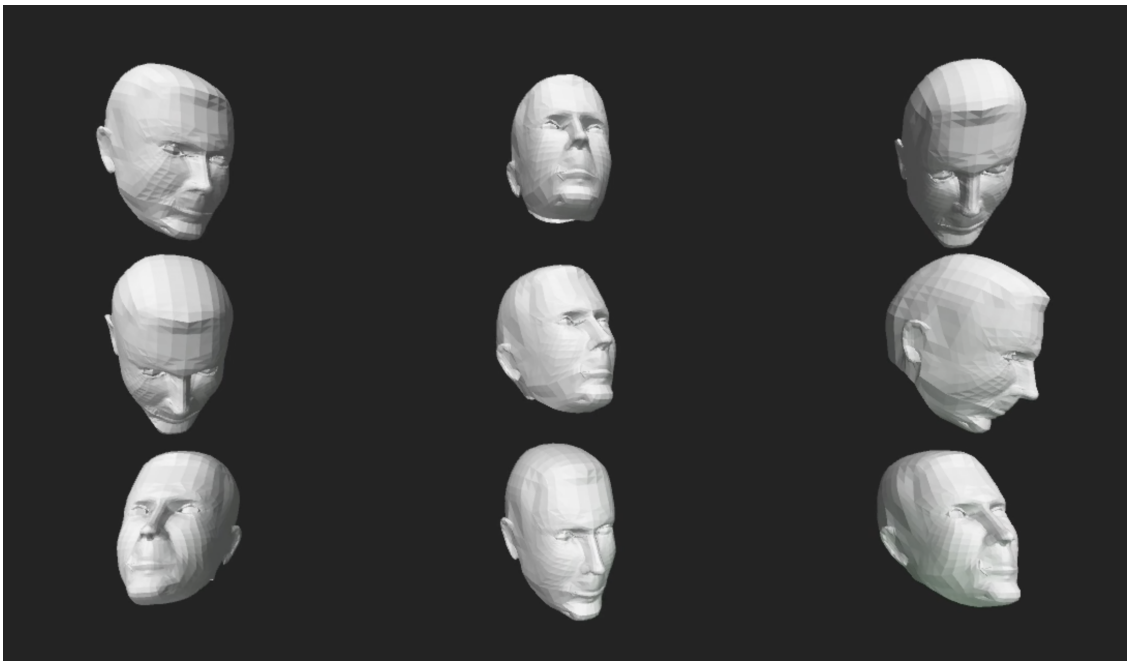


Figure 5: Nine head avatars fitted inside a grid view

4.2.1 Heads

In the 3D environment, the head avatars are loaded from a free-to-use .OBJ 3D model downloaded from *TurboSquid* with standard usage license.[31] We load the object into the 3D scene once, and use the same model instance for all other head avatar objects to save computation — since they're essentially the same 3D model, just redrawn with different transformations.

Due to time and budget constraints, we are limited to a low-polygon-count model of the head, with no texture, no rigging, no soft-body animation, and no physics simulation. The 3D model of the head, as shown in Figure 4b, has minimum features of a face. Despite the low-quality 3D models, the primitiveness of the 3D model helps maintaining a high rendering frame-rate while the prototype is running, even without writing custom shaders and performing fine-tuned optimizations — especially when there could potentially be up to 9 to 25 avatars being drawn concurrently on to the screen. This is important as we need the head avatars to transform and render in real-time as if they're real inputs in a WVC application. As we will discuss in Section 8, if given more budget and 3D talent, more work can be allocated to polishing the visual appeal of the avatars to be more inviting and friendly, and ultimately improve user-experience.

4.2.2 Eyes

The 2D eyes avatar was inspired from goggly eyes and novel desktop widgets (e.g. XEyes[32]) created as general amusement. However, we decided to explore this as an alternative to the head avatars to see whether if 3D and head orientation is required to deliver eye-contact and gaze hints to the meeting participants.

insert eye avatar view

insert eye avatar grid view

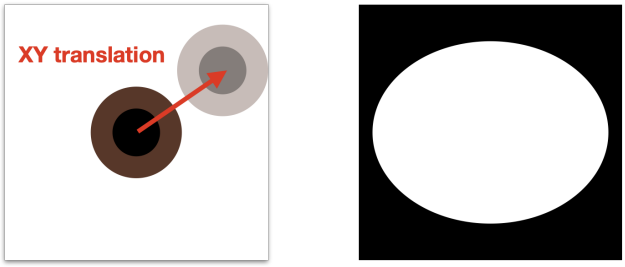


Figure 6: Breakdown of how an eye is rendered for eye avatars

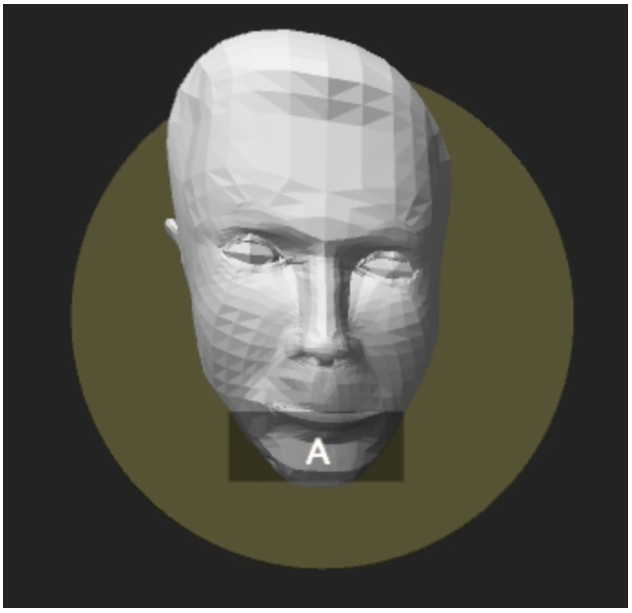


Figure 7: A Halo highlights avatars that are in focus, i.e. whoever is talking

With eyes avatars, instead of a 3D object transformed to orient a direction, we simply render a pair of pupils and irises on a white background as a sub-image/texture (Figure 6). Then depending on where the eye should be looking at, we rendering this sub-image with a transform that translates it in x or y direction (Figure 6). Finally, we create an eye mask that only only renders the eye itself (Figure 6).

The result is somewhat similar to a gallery view of meeting participants in a traditional WVC application, except with only the eyes and their gaze visually represented.

4.2.3 UI Elements

As discussed in Section 4.2, the 3D head avatars and 2D eye avatars are rendered at $z = 0$. In front of that, we render the UI elements such as the bottom menu bar, as well nameplates to aid in our experiments and to provide a familiar environment to the participants.

In Figure 7, when a participant corresponding to an avatar is talking, we have an option to activate a yellow halo behind the avatar for indication. To do this, each avatar base class has a flag `isFocused` that can be toggled. If this flag is on, then we draw this additional halo on a layer behind the avatars.

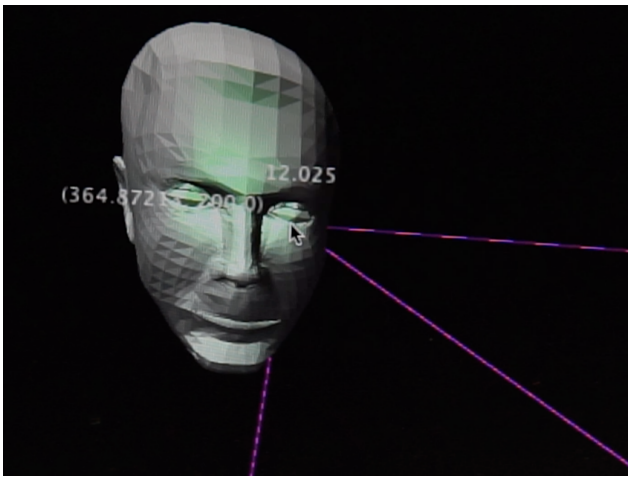


Figure 8: Point light highlights where the mouse is

While not used in our final experiments, we also added an option to highlight avatars in on the screen based on mouse cursor positions. Rather than drawing a bounding box around the avatar, we thought it was appropriate to instead model the mouse cursor effects as a point-source-light with some vibrant colour, such as seen in Figure 8. This option can be turned on via experiment configuration file (see Section 4.5).

4.3 View Modes

To simplify the behaviour of mock avatars, we have three pre-programmed modes for the avatars to follow:

1. **NORMAL**: the avatar is in normal mode, it will follow and track a given set of target coordinates (see Section 4.4.1).
2. **STARE**: the avatar should stare at the active participant by gazing directly outwards from the screen.
3. **RANDOM**: the avatar randomly looks around as if they are distracted. Current implementation of the random mode uses a series of Perlin noise functions to approximate how real head moves around randomly.

Note that this would only apply to the mock avatars used in user experiments to simulate real people.

During user experiments, we can dynamically and programmatically change the modes of the avatars to simulate whether a person in the meeting is paying attention or not paying attention, and looking at the test subject participant, or anywhere else on the screen.

4.4 View Calculation

This section talks about the math and algorithms created to compute the target coordinates — the screen-space coordinates of where that avatar should be looking at. As well as the mapping between the target coordinates to a rotation (for head avatars) or a translation (for eye avatars) transformations.

4.4.1 Target Coordinates

Each avatar (*View* base class) has attributes targetX and targetY which corresponds to where the avatar should directly look at in *normal* mode. These attributes are not used in *random* and *stare* modes.

For example, if the avatar is set to track the mouse cursor's screen position, then we trivially set:

```
avatar.targetX = mouseX
avatar.targetY = mouseY
```

If an avatar A is set to look at another avatar B, we can set the target coordinates of A to the spatial coordinates of B:

```
A.targetX = B.x
A.targetY = B.y
```

4.4.2 Rotation Calculation

Once an avatar knows *where* to look (i.e. the target coordinates), it needs to compute *how* to look in that direction (i.e. compute the corresponding transforms required to show the correct visual representation).

For the 2D eye avatars, this mapping from target coordinates to transformation is a simple 2D translation $(\Delta x, \Delta y)$, as seen in Figure 6. First, a difference vector $\vec{d}$ is computed from the avatar's local origin to the target coordinates. We then scale it by some factor s to control the sensitivity of this translation. Finally, we constrain the magnitude of $s\vec{d}$ to the radius of the eyes to ensure the pupil do not go off the eye, creating a white eye.

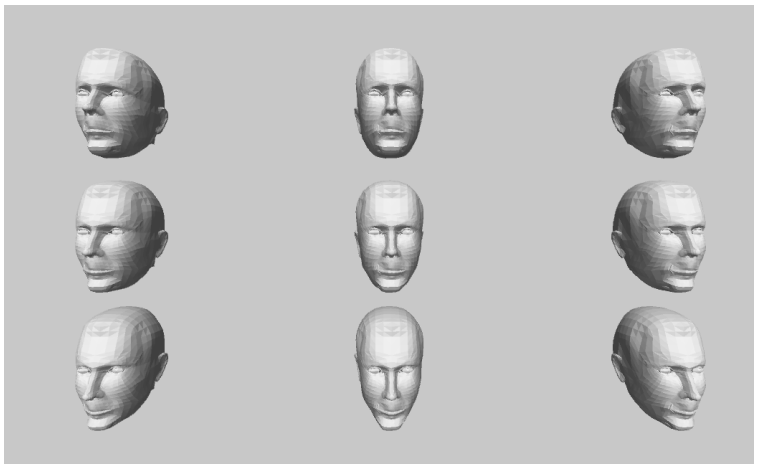
For the 3D head avatars, the mapping is instead from the target coordinates to a set of rotations along the x, y, and z-axis. The simplest way to do this is to draw a 3D vector from the head avatar origin to the target coordinates. Then using trigonometry on the vector, we can find all angles corresponds to the x, y, and z rotations.

Notice, that up to this point, we have not established the third, z-component, of the target coordinates. This is because it depends on what the head avatar should be looking at. If the avatar is looking at another avatar, we leave $z = 0$ since all avatars are on the $z = 0$ plane. However, if the avatar were to follow the mouse or stare at the user, then we must set $z = z'$, where z' is the z-offset of the camera. The difference between with vs. Without this z-correction can be seen in Figure 9.

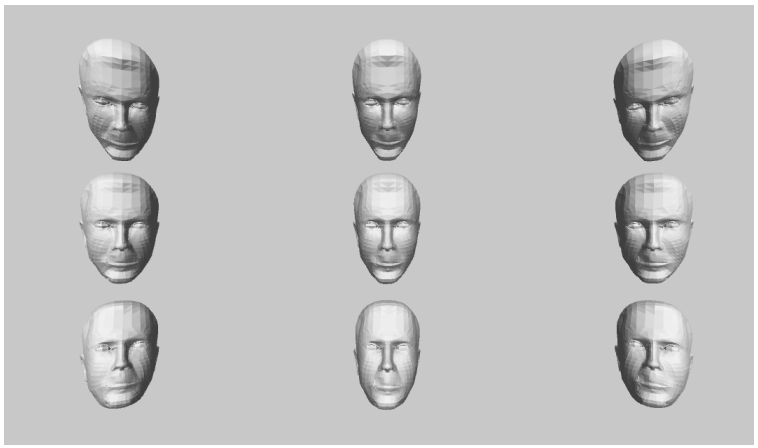
4.4.3 Realism Approximations

For both eye and head avatars, instead of applying the transformation in the renders according to he target coordinates instantaneously, we use a linear-interpolation function to smoothen the motion and simulate a more natural response — a response with mass, inertia, and a sense of reaction time delay that would exist in real people.

We also added an option to make the head wobble controlled by some noise to make the avatars feel less robotic and more organic. If a 3D head avatar's isFocused flag is set on (such as when the participant the avatar belongs to is speaking), the wobble amplitude is slightly increased to approximate the extra motion due to mouth movements.



(a) A grid of head avatars without rotation offset correction



(b) A grid of head avatars with rotation offset correction

Figure 9: Grid of heads require some rotation offset to stare out of the screen

4.5 Configuration Files

To facilitate a series of automated, consistent, yet randomly generated user experiment scenarios, we developed a configuration framework for our prototype. Each experiment setup (see Section 5) is contained inside a config.json file. Each file contains all the experiment parameters such as number of avatars, type of avatars, names, and the sequence of modes, target coordinates, etc.

4.5.1 Initialization

The file can be loaded at initialization-time of the experiment. Upon which, all the parameters in the configuration file is read, and the defined avatars are populated in the scene.

4.5.2 Events

The sequence of state changes in the prototype user experiment is controlled by *events*. These events are defined in the configuration files as a single array that represents a timeline. The different types of events supported are:

- Change mode
- Change avatar target
- Set focus flags
- Play sound
- Stop experiment

For more details regarding events, please refer to the source code.

4.6 Unity Engine Wrapper

To further facilitate user testing, we wrapped the Processing prototype in a Unity engine user interface, where participants are guided without the constant supervision from us. Each experiment setup is sequenced in order and appropriate questionaries are prompted in between each setup. Finally, the questionnaire responses are automatically logged in participants' computer, making it easy for review.

5 Experiment Setup

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

5.1 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.

5.2 System Requriements

To run the experiments, we recommend the following system requirements:

- **Operating System**: Windows, macOS, or Linux in 64-bit
- **System Memory**: At least 4 GB
- **CPU**: At least Intel i3 or above
- **GPU**: At least Intel integrated graphics (e.g. Iris Plus Graphics 645) or above
- **Microphone**: yes
- **Zoom**: installed

5.3 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.

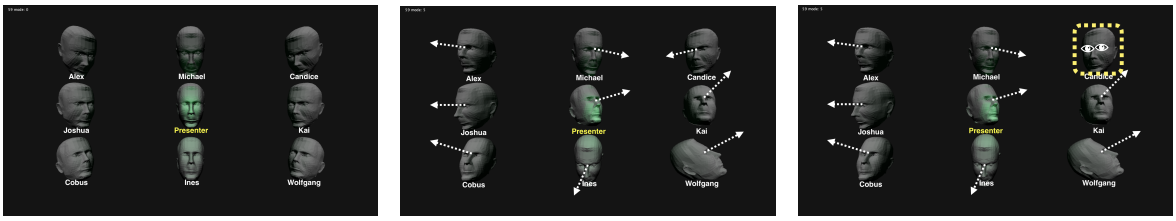
Experiment 3 (E3, H3) involves involves the participant to speak in a room of mock-avatars. In this case we explore **RQ3** and attempt to gauge how participants feel, including nervousness, focus, and engagement with the audience using our prototype.

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.



(a) (b) (c)

Figure 10: (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.

5.3.1 Experiment E1

Begin by setting up nine mock-avatars in the WVC window, each with a unique name as seen in Figure 10(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 10, head avatars are used, but as mentioned in 5.3, 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 10(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 10(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.

5.3.2 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 5.3, 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.

5.3.3 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.

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.

5.3.4 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.[5] Inspired by the body sheets as a method to collect user responses in La Delfa et al.'s work in Drone Chi,[33] 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 11 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 11, of the relationship matrix, to indicate which avatar is engaging with which. We may also ask the participant to annotate each arrow

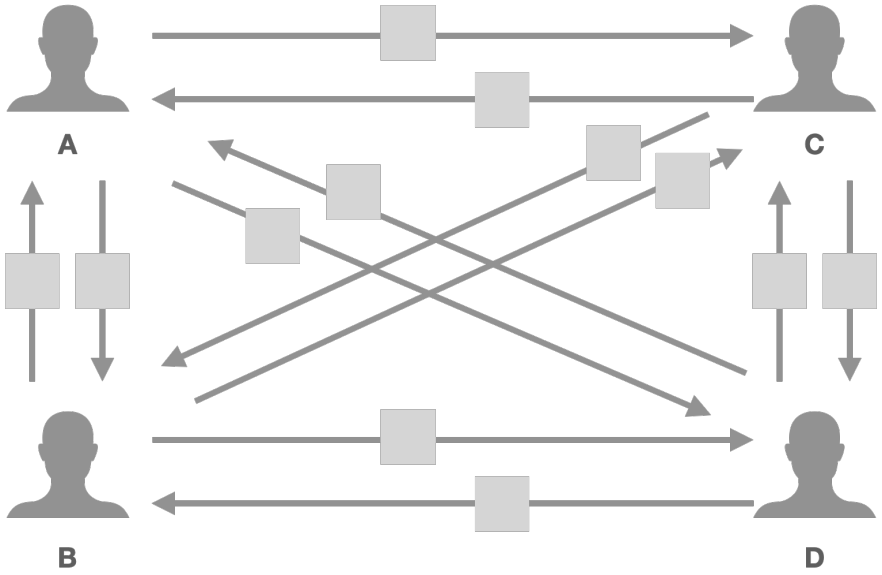


Figure 11: The diagram we ask the participant to fill out which corresponds to the relationship matrix $\mathbf{P}$

with a confidence score (0.0 - 1.0). These scores can be normalized and compared with the probability matrix $\mathbf{P}$ 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.

6 Evaluation & 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 12 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.

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 13.

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

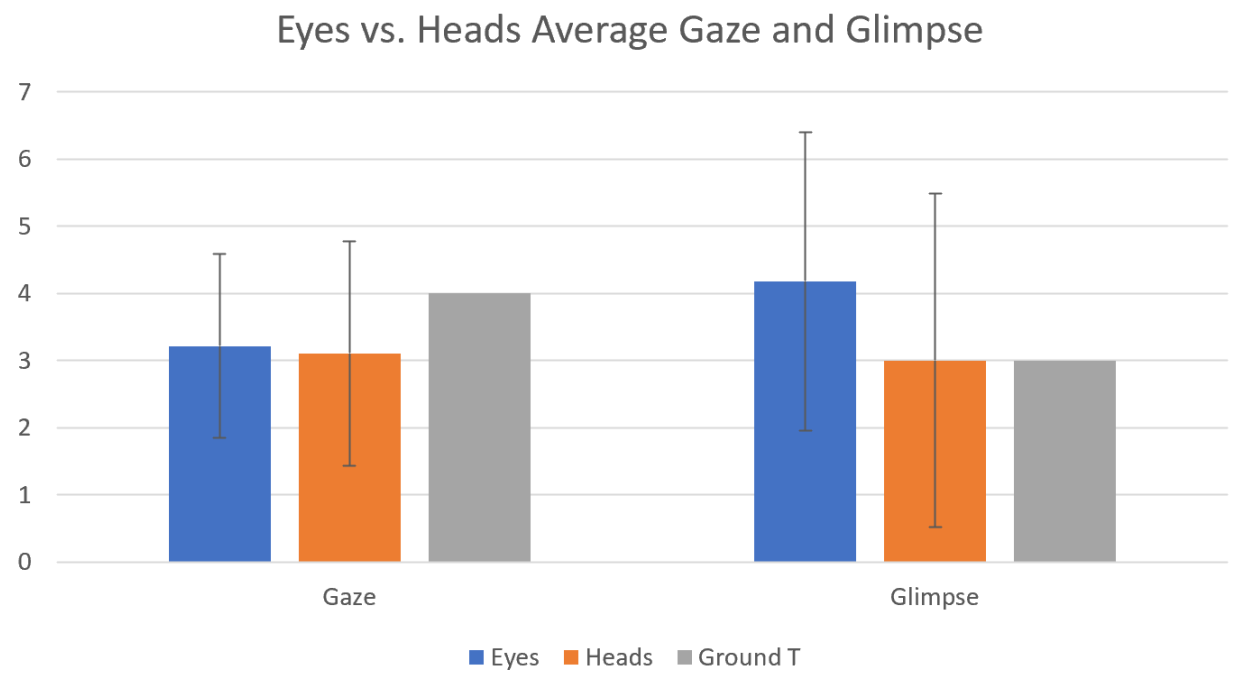


Figure 12: The eye model and head model gaze and glimpse times comparison with ground truth

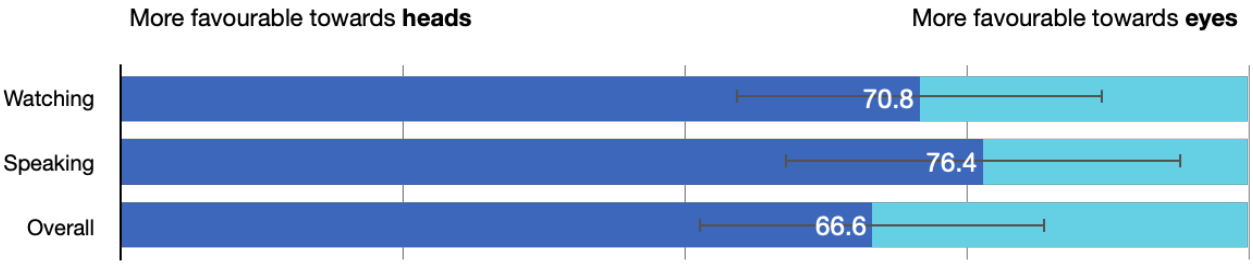


Figure 13: The attention distribution of heads and eyes in watching, speaking and overall tasks

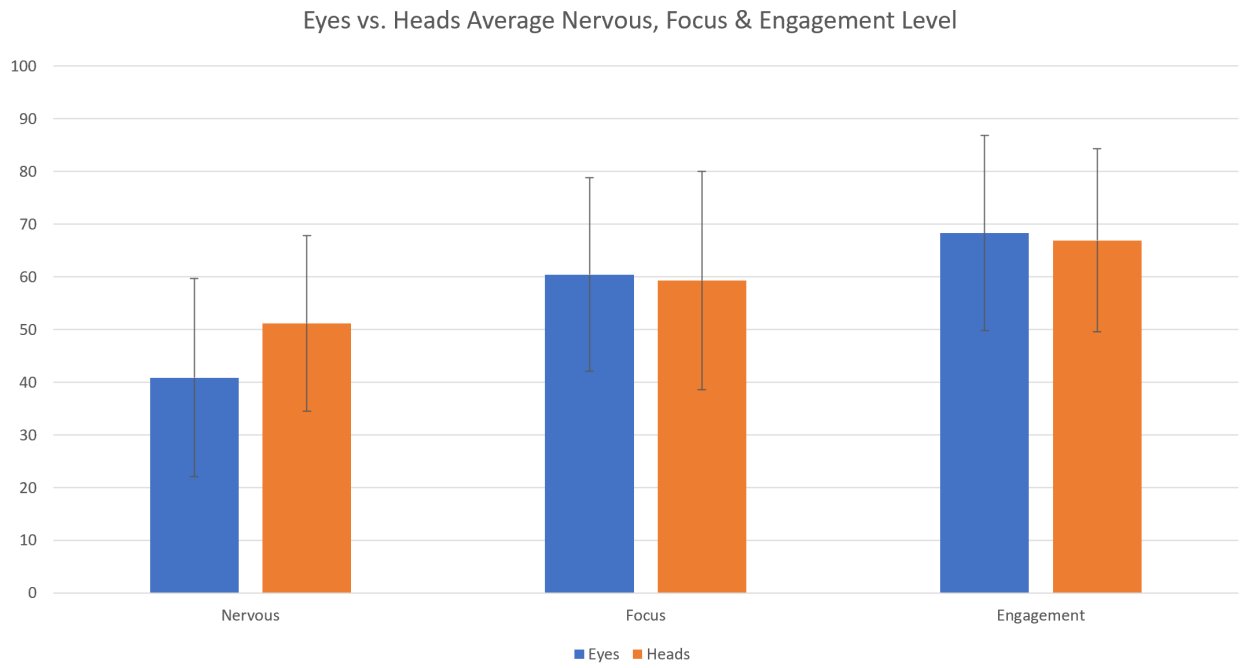


Figure 14: Nervousness, focus, and engagement levels reported by the participants as they are asked to speak/present in a room of mock-avatar listeners (E3, H3)

and engagement level compared with traditional WVC when using the eye model and the head model. As shown in Figure 14, 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 11 were more closely related to the dynamics in the dialog, rather than eye contact or gaze.

Figure 15 shows the mean-squared error (MSE) of the relationship matrix response. Eventhough users reported having an easier time seeing using head avatars, they were not any better at correctly perciving the interaction and relationships. There average increase in misperceptions/errors when switched to heads; and according to user responses, this could be attributed to head avatars being more distracting. But it's worth noting that the best-case of MSE decreases when switching from eyes to heads.

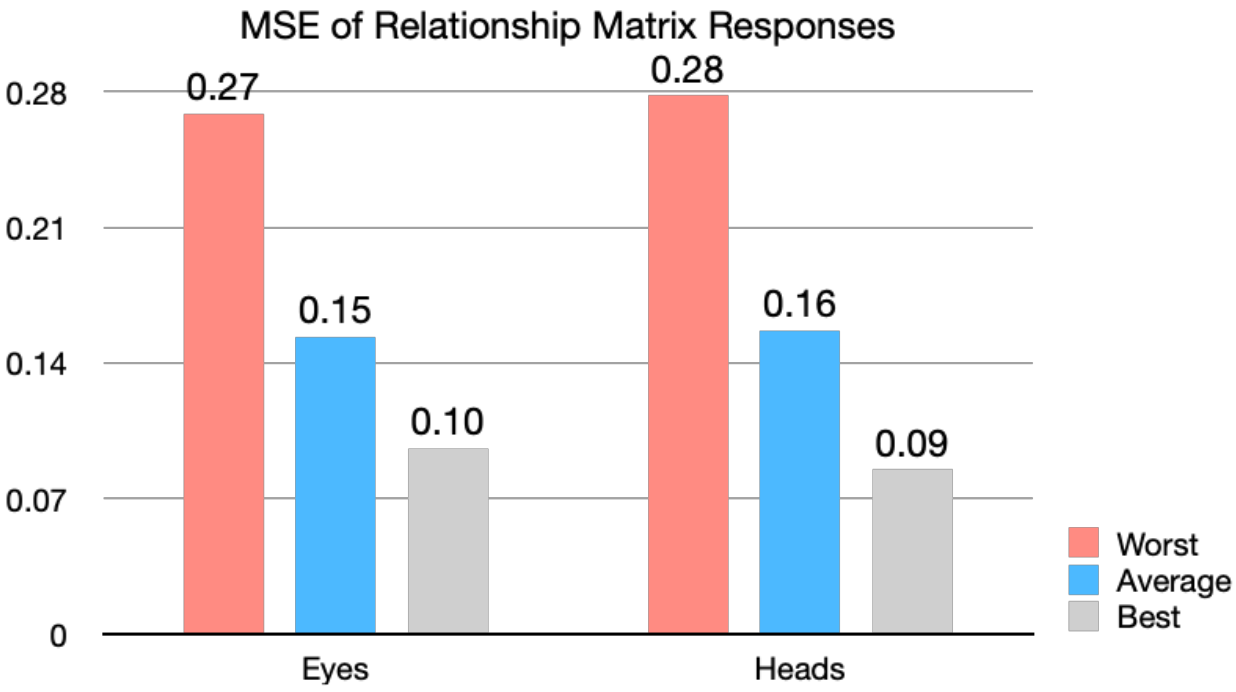


Figure 15: Mean-squared error of relationship matrix responses (E4, H4)

Finally, Figure 16 shows the magnitude of the user responses between eye avatars and head avatars. There is a slight increase in percived attention amongst avatars when using heads. However, this change is very insignificant and could be result of variance/noise.

Generally, the relationship matrix study is not conclusive, and more scenarios and test participants is needed.

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.

7 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.

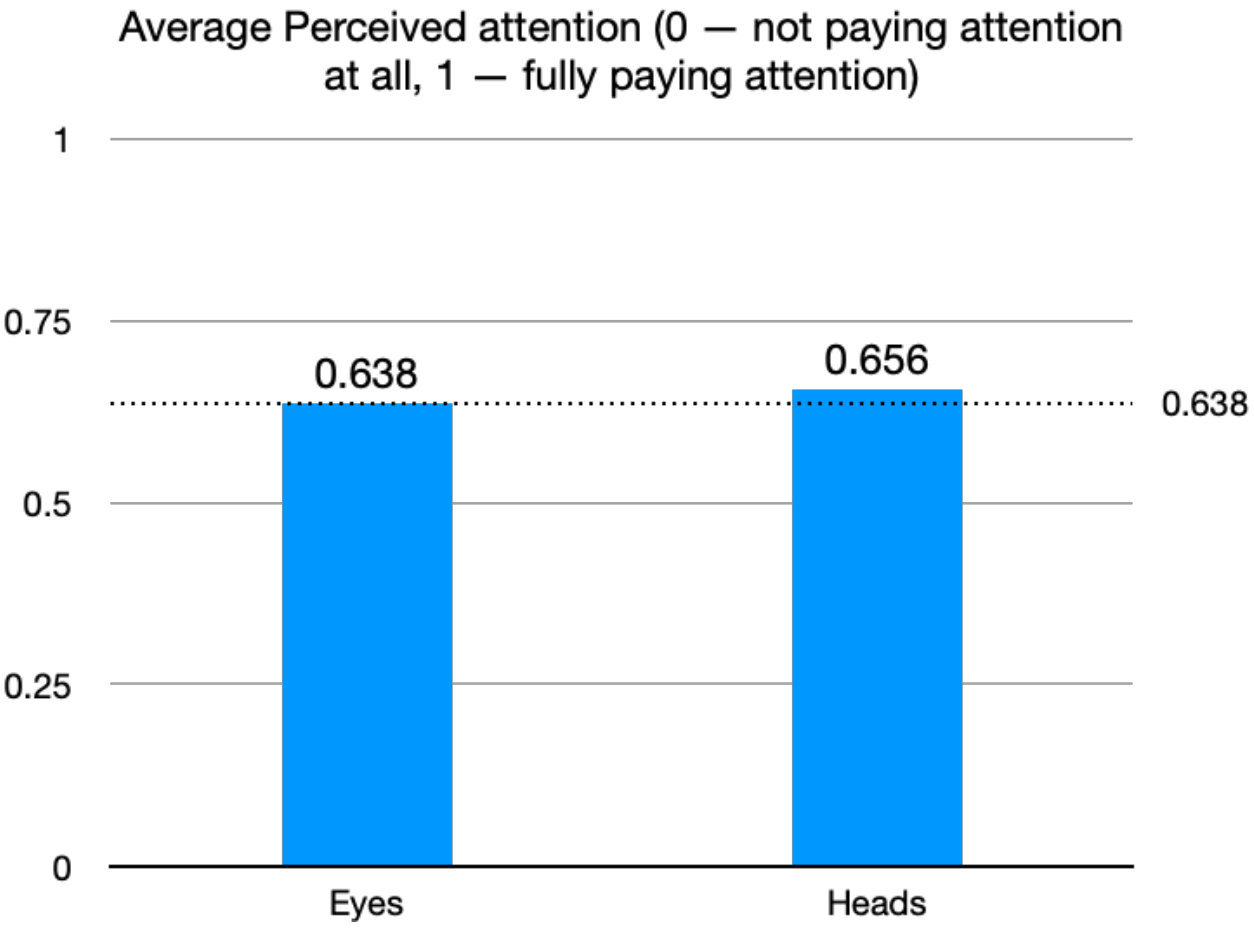


Figure 16: Normalized mean magnitude of the responses (higher means more percieved attention) (E3, H3)

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.

8 Limitations & 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.*"

Overall, the project documented in this report is only a single component of the entire stack that is required to create a fully-functional app. Other key components that is not mentioned here (since they are not relevant to the research question) include networking using web sockets, computer graphic optimizations, encryption and security, privacy, data transmission, gaze tracking, etc. Figure 17 shows the proposed data flow and modules of a completed FutureGazer app.

9 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.

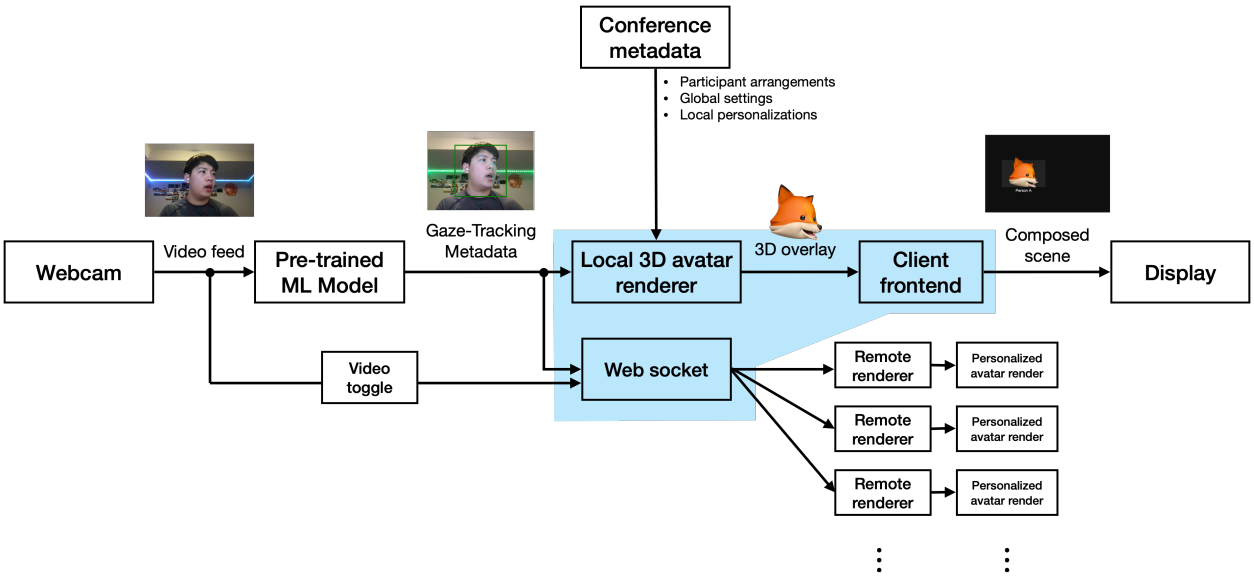


Figure 17: Data flow and block diagram of an ideal fully-functional FutureGazer WVC application

References

[1] H. Al-Samarraie, "A scoping review of videoconferencing systems in higher education: Learning paradigms, opportunities, and challenges," *International Review of Research in Open and Distance Learning*, vol. 20, 07 2019.

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

[3] J. J. Roth and S. Pierce, MariBrewer, "Performance and satisfaction of resident and distance students in videoconference courses," *Journal of Criminal Justice Education*, vol. 31, no. 2, pp. 296–310, 2020.

[4] D. Doggett and A. Mark, "The videoconferencing classroom: What do students think?," *Architectural and Manufacturing Sciences Faculty Publications*, p. 3, 2008.

[5] J. K. Hietanen, "Affective eye contact: an integrative review," *Frontiers in psychology*, vol. 9, p. 1587, 2018.

[6] R. Cañigueral and A. F. d. C. Hamilton, "Being watched: Effects of an audience on eye gaze and prosocial behaviour," *Acta psychologica*, vol. 195, pp. 50–63, 2019.

[7] A. T. Duchowski and A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2017.

[8] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding*, vol. 98, no. 1, pp. 4–24, 2005.

[9] T. Ohno, N. Mukawa, and A. Yoshikawa, "Freegaze: a gaze tracking system for everyday gaze interaction," in *Proceedings of the 2002 symposium on Eye tracking research and applications*, pp. 125–132.

[10] NVidia, "Gans improve video conferencing with maxine," 2020.

[11] C. Macrae, B. Hood, A. Milne, A. Rowe, and M. Mason, "Are you looking at me? eye gaze and person perception," *Psychological science*, vol. 13, pp. 460–4, 10 2002.

[12] M. Chen, "Leveraging the asymmetric sensitivity of eye contact for videoconference," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, (New York, NY, USA), p. 49–56, Association for Computing Machinery, 2002.

[13] N. Mukawa, T. Oka, K. Arai, and M. Yuasa, "What is connected by mutual gaze? user's behavior in video-mediated communication," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, (New York, NY, USA), p. 1677–1680, Association for Computing Machinery, 2005.

[14] D. M. Grayson and A. F. Monk, "Are you looking at me? eye contact and desktop video conferencing," *ACM Trans. Comput.-Hum. Interact.*, vol. 10, p. 221–243, Sept. 2003.

[15] R. Stokes, "Human factors and appearance design considerations of the mod ii picturephone® station set," *IEEE Transactions on Communication Technology*, vol. 17, no. 2, pp. 318–323, 1969.

[16] K.-I. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita, "Multiparty videoconferencing at virtual social distance: Majic design," in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, (New York, NY, USA), p. 385–393, Association for Computing Machinery, 1994.

[17] H. Ishii and M. Kobayashi, "Clearboard: A seamless medium for shared drawing and conversation with eye contact," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, (New York, NY, USA), p. 525–532, Association for Computing Machinery, 1992.

[18] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, (USA), p. 369–374, ACM Press/Addison-Wesley Publishing Co., 2000.

[19] W. Matusik and H. Pfister, "3d tv: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," *ACM Trans. Graph.*, vol. 23, p. 814–824, Aug. 2004.

[20] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, (New York, NY, USA), p. 600–608, Association for Computing Machinery, 2004.

[21] B. Petit, J.-D. Lesage, C. Ménier, J. Allard, J.-S. Franco, B. Raffin, E. Boyer, and F. Faure, "Multi-camera real-time 3d modeling for telepresence and remote collaboration," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, 08 2010.

[22] C. Kuster, T. Popa, C. Zach, C. Gotsman, and M. Gross, "FreeCam: A Hybrid Camera System for Interactive Free-Viewpoint Video," in *Vision, Modeling, and Visualization (2011)* (P. Eisert, J. Hornegger, and K. Polthier, eds.), The Eurographics Association, 2011.

[23] C. Kuster, T. Popa, J.-C. Bazin, C. Gotsman, and M. Gross, "Gaze correction for home video conferencing," *ACM Trans. Graph.*, vol. 31, Nov. 2012.

[24] M. Argyle, M. Cook, and D. Cramer, "Gaze and mutual gaze," *British Journal of Psychiatry*, vol. 165, no. 6, p. 848–850, 1994.

[25] E. A. Isaacs and J. C. Tang, "What video can and can't do for collaboration: A case study," in *Proceedings of the First ACM International Conference on Multimedia*, MULTIMEDIA '93, (New York, NY, USA), p. 199–206, Association for Computing Machinery, 1993.

[26] A. J. Sellen, "Remote conversations: The effects of mediating talk with technology," *Hum.-Comput. Interact.*, vol. 10, p. 401–444, Dec. 1995.

[27] R. Vertegaal, G. Veer, and H. Vons, "Effects of gaze on multiparty mediated communication," 12 2000.

[28] P. Hart, L. Svenning, and J. Ruchinskas, "From face-to-face meeting to video teleconferencing: Potential shifts in the meeting genre," *Management Communication Quarterly*, vol. 8, no. 4, pp. 395–423, 1995.

[29] P. Foundation, "Processing," 2020.

[30] Apple, "Using gatekeeper in macos deployments," 2020.

[31] TurboSquid, "Turbosquid 3d model license," 2020.

[32] coralw, "Xeyes," 2016.

[33] J. La Delfa, M. A. Baytas, R. Patibanda, H. Ngari, R. A. Khot, and F. Mueller, "Drone chi: Somaesthetic human-drone interaction," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.