Breaking the Mona Lisa Effect: Enhancing Eye Contact with Virtual Humans in 2D Display Environments*

Sunghun Jung[1], Junyeong Kum[1] and Myungho Lee[1,*]

[1]Pusan National University, Republic of Korea

Abstract

With the rise of large language models, virtual humans (VHs) are becoming increasingly prevalent as conversational service agents across various domains. Consequently, kiosks equipped with large 2D displays have become the primary platform for deploying VH systems. However, VHs presented on 2D displays introduce the Mona Lisa effect, which hinders precise eye contact interaction with users. In this paper, we propose a straightforward method to alleviate the ambiguity in VHs' eye gaze direction. To evaluate the effectiveness of our method, we conducted an experiment involving 30 participants. The results revealed a statistically significant improvement in gaze perception accuracy within 2D VH systems when employing our approach. This improvement has the potential to enhance user engagement and the overall quality of human-virtual human interactions.

Keywords

virtual human, eye contact, reduced Mona Lisa effect, visual cue

1. Introduction

Recently, virtual humans (VHs) have been utilized in various fields, including education, counseling, and entertainment. This is partially attributed to VH modeling tools[1] that have simplified the creation of VHs resembling real people, and the availability of large-scale language models such as GPT-3 [1], enabling natural dialogue. Furthermore, advancements in computer vision AI technology have facilitated the distinction of individuals, user tracking, and analysis of emotions and other essential interaction-related information [2].

Ongoing studies are utilizing these technologies to interact with VHs [3, 4], and the use of VH kiosk systems is increasing in various domains, such as banks, government offices, and museums. While VH research projects typically incorporate specialized devices for the 3D representation of VHs, those service kiosks often solely employ a 2D display (see NVIDIA Omniverse Avatar [2]), potentially imposing an issue of delivering non-verbal communication cues requiring 3D perception.

Previous research has demonstrated that people have a tendency to expect human-like interactions from VHs. In such interactions, verbal expressions hold significance, while non-verbal communication cues like gestures, eye contact, and facial expressions are also crucial [5, 6]. Among these non-verbal cues, eye contact plays a vital role in expressing interest, concentration, facilitating dialogue in multiparty conversations, and enabling smooth turn-taking [7].

However, the representation of 3D VHs on a 2D display gives rise to various optical illusions, including the Mona Lisa effect, which poses challenges for establishing eye contact during communication [8, 9]. The Mona Lisa effect occurs when a user positioned within 5 degrees to the left or right of the VH on the 2D display perceives the VH as making eye contact [9]. Unlike real-life conversations occurring in a 3D space, where individuals can accurately perceive each other's gaze even in confined areas, the presence of the Mona Lisa effect can disrupt the accurate perception of eye contact between users and VHs in 2D kiosk systems, particularly when users are in a confined space [8]. As a result, rendering VHs in such a 2D kiosk system has the potential to reduce user engagement in conversations and overall satisfaction with the VH systems.

In this paper, we propose a technique to mitigate the Mona Lisa effect in 2D display environments without the need for specialized equipment such as a parallax barrier or stereoscopic glasses. Our method involves subtly rotating the virtual camera, which significantly improves the perception of the gaze direction of a VH displayed on a 2D screen. The rest of the paper discusses related work, details our proposed method, and presents an experiment conducted to validate the effectiveness of our approach.

APMAR'23: The 15th Asia-Pacific Workshop on Mixed and Augmented Reality, Aug. 18-19, 2023, Taipei, Taiwan

*Corresponding author.

sunghun@pusan.ac.kr (S. Jung); junegold12@pusan.ac.kr (J. Kum); myungho.lee@pnu.edu (M. Lee)

0009-0007-0987-7190 (S. Jung); 0000-0003-3943-3463 (J. Kum); 0000-0002-9421-8566 (M. Lee)

[1]https://www.reallusion.com/character-creator/

[2]https://www.nvidia.com/en-us/on-demand/session/gtcfall21-d31017/

2. Related Works

While we are unaware of previous efforts directly addressing the mitigation of the Mona Lisa effect, some relevant areas of work are summarized in this section.

The Mona Lisa effect is a phenomenon in which individuals perceive that the gaze of a static image, depicting a person or animal, is following them. This effect is particularly noticeable within an angle of 5 degrees in both the left and right directions [9]. Moubayed and Beskow [8] conducted a study to investigate the impact of the Mona Lisa effect on gaze perception. They compared a face projected onto a 3D object with a face presented in a 2D format. Interestingly, while both conditions involved eye movements that followed the users, the results revealed a significant decrease in accuracy in the 2D condition.

In response to the Mona Lisa effect, researchers have proposed various methods to enhance eye contact interaction with VHs in 2D display environments. Some of these approaches involve modifying devices and applying content-based techniques to achieve a more natural eye contact experience. One such method includes the addition of an assistive device that rotates the display to align with the user's position, enabling a more natural and immersive eye contact interaction [10, 11]. The use of dual display also has been proposed as a means to mitigate the Mona Lisa effect [12]. They employed two displays, one depicting only the eyes and the other featuring the VH, to demonstrate that perceiving the character from a different angle reduced the Mona Lisa effect. However, these methods have the drawback of requiring specialized equipment or additional displays. Another approach involves leveraging visual effects within the content to create a sense of natural eye contact. A particular study showcased how VHs can establish eye contact with users in passing or standing scenarios by employing head rotation and changes in pupil position [13]. While these studies focused on the naturalness of eye contact rather than reducing the Mona Lisa effect, in this paper, we focused on ways to reduce the Mona Lisa effect using visual cues.

3. Proposed Method

Our method is based on Fish Tank VR, proposed by Ware et al. [14]. In Fish Tank VR, the position and rotation of the viewpoint are continuously updated to match the user's head position, so that the stereo image in the 2D display appears to be 3D. In our proposed method, we provide a visual cue by simply controlling the rotation of the main camera of Unity Engine, which displays the 3D virtual environment.

We employed YOLO8 [15] with RealSense Depth Camera D455 to recognize participants' faces. The estimated position and depth data were used to calculate the user's real-world position. However, the position of their faces can vary depending on their sitting posture or height, even though we controlled the position of the participants' seats during the experiment. To reduce the variability due to these differences, we limited the camera rotation value to a maximum of 1 degree left and right, respectively. Therefore, the yaw angle of the camera increased up to 1 degree, following the participant's face position. Beyond that, the yaw angle remains at 1 degree while VH's gaze continues to follow the participant. It should be noted that the camera position was fixed.

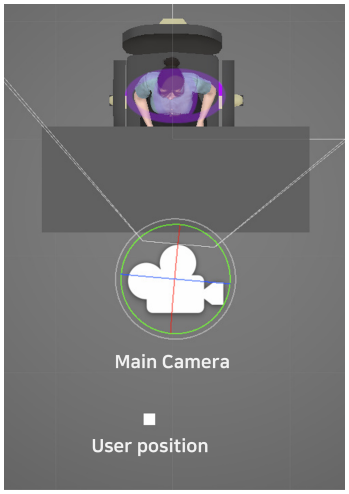


Figure 1: The top view in the Unity 3D virtual environment in the 5-degree condition with our proposed method, when VH looking at the left participant

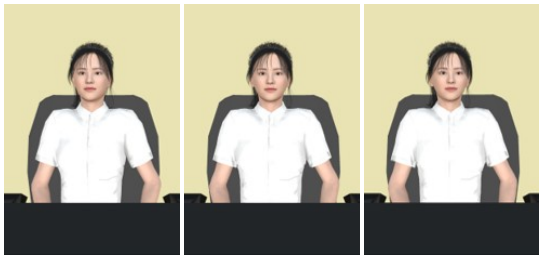
Figure 1 shows the top view in the Unity 3D virtual environment when the VH looks at the left participant. To clearly demonstrate the rotation of the main camera, we adjusted the camera angle to 5 degrees.

4. Experiment

This experiment examines whether our proposed method can improve participants' gaze recognition accuracy in a narrow range where the Mona Lisa effect can occur. In other words, when a VH on a 2D display looks at one of two participants located within a 5-degree, we check whether the participants can recognize whom a VH is looking at.

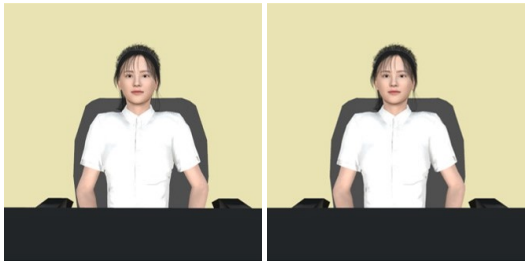
4.1. Method

We used a within-subjects design to examine the effectiveness of our proposed method. The experiment was conducted at 5 and 15-degree angles, with all participants first experiencing the 15-degree condition. In both condi-



(a) (b) (c)

Figure 2: VH looking (a) a left participant, (b) the center, and (c) a right participant in 15-degree condition



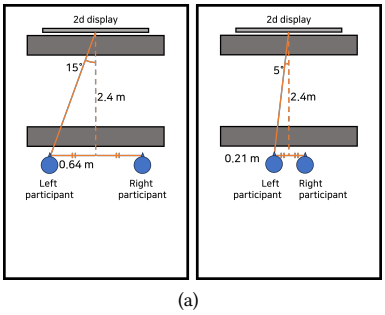
(a) (b)

Figure 3: VH looking (a) a left participant with and (b) without camera rotation in 5-degree condition

tions, the virtual human (VH) appropriately rotated her spine, head, and eyes to match the position of the target it was looking at (Figure 2 and Figure 3).

The 15-degree condition was conducted to verify that the participants understood the experiment by assessing the VH's gaze outside the range where the Mona Lisa effect occurs. There were no other visual cues except for the VH's gesture. Then, the 5-degree condition was conducted to examine our proposed method in the narrow range where the Mona Lisa effect could occur. In both conditions, the VH looked at the left participant, the middle between both, or the right participant. A set of 30 gaze judgment tasks consisting of ten of each gaze direction was used for the 15-degree condition, while in the 5-degree condition, we used a total of 60 gaze judgment tasks. In the latter condition, our proposed method was applied for half the time (see Figure 3(a)), and the other half was not (Figure 3(b)) The order of VH's gaze directions was counterbalanced and randomized. Additionally, five VHs were used in the experiment, and the order was also randomized. For the eye gaze of the VH, we exploited the Wizard of Oz paradigm, where an experimenter in another space controlled the position of the target for the VH to look at using a GUI.

4.2. Environment



(a)

(b)

Figure 4: (a) The experimental setup and the schematics of the experimental space and (b) actual experimental space

We organized our experimental environment with a 65-inch TV, two tables, and two chairs (Fig 4(b)). The distance between the 2D display and the center of the two participants was 2.4 meters. Additionally, the participants sat at the same distance from side to side based on the VH; 0.64 meters in the 15-degree condition and 0.21 meters in the 5-degree condition (Fig 4(a)). A desk was placed in front of the participants so that they could fill out a questionnaire with their smartphones.

We used the Unity game engine [3] to render the VH and an office-like environment on the 2D display. For the VHs, we used rigged 3D human models created by Character Creator 3 [4]. To eliminate the gender effect of the VH, we made all five VHs female. We modified the appearance of the VHs to have different facial features, hairstyles, and clothes. The VH was placed behind a table with a monitor and other office supplies. In the real experimental space, we placed an actual desk in front of the 2D display. We used Final IK [5] for natural gaze behavior.

[3] https://unity.com
[4] https://www.reallusion.com/character-creator/
[5] https://root-motion.com/

4.3. Participants

Due to the experimental configuration, we recruited two participants at a time. In cases where only one participant was available, one of the experimenters pretended to be the participant and joined the experiment. The data submitted by the experimenter were excluded from the analysis. We recruited 30 participants (13 males and 17 females) from a local university. The average age of the participants was 22.4 (SD=2.10). Before the experiment, an experimenter tested the participants' dominant eye through a simple test (right eye 18, left eye 12).

4.4. Questionnaire

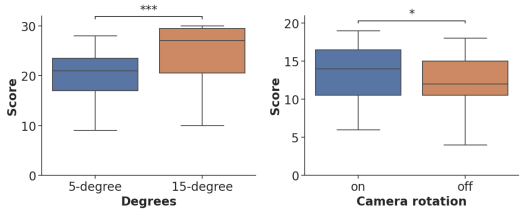
The pre-questionnaire included questions about demographics and the dominant eye. During the experiment, the participants selected whether they thought the VH was looking at them, the center, or the other participant. We counted the correct responses.

4.5. Procedure

After the participants completed the pre-questionnaire, the experimenter briefly explained the experiment. The participants were then seated in the appropriate seats based on the conditions, and the experiment began.

After a 3-second black screen, the VH was rendered on the display, looking either at the left participant, the center, or the right participant, depending on the condition. After both participants answered where they thought the VH was looking, the experimenter pressed the next button on the GUI to activate the black screen. The direction the VH looked at was indicated on the GUI in a predetermined order. While the black screen was displayed, the experimenter clicked on the location of the target for the VH to look at in order. The VH and the main camera rotation were controlled in Unity automatically. For each condition, we repeated this process 30 times.

5. Result



(a) (b)

Figure 5: Each graph shows the score in (a) 5-degree and 15-degree conditions and in (b) camera rotation on / off in 5-degree condition.

We found statistically significant results by using the one-way ANOVA. First, we compared the 15-degree condition and the 5-degree condition (Figure 5(a)). The results showed that participants were more accurate at guessing where the VH was looking in the 15-degree condition, outside the Mona Lisa effect ($p < 0.000$).

Additionally, to assess the effectiveness of our proposed method, we used a dataset from the 5-degree condition for analysis (Figure 5(b)). In this result, we only utilized the dataset when the VH looked at the left or right participant since there was no camera rotation when the VH looked at the center. The participants were more accurate at guessing where the VH was looking in the narrow range when our proposed method was applied ($p = 0.033$).

6. Discussion&Conclusion

In this study, we found that the inclusion of supplementary visual cues effectively diminished the Mona Lisa effect. Not only did participants recognize eye contact as a distinguishing feature of the VH, but the subtle camera rotation also introduced subtle disparities between the desk and chair-like background. Building upon these findings, we anticipate that incorporating multiple visual cues will enhance the likelihood of accurately perceiving eye contact.

For instance, displaying the complete body of the VH can provide insight into the character's body orientation, while incorporating lines or patterns into the background can aid in establishing eye contact. These additional cues contribute to a more comprehensive visual context that facilitates the perception of eye contact.

However, a notable limitation of this experiment is its focus on two user cases centered around a specific basis, casting doubt on whether the significant effect observed would be consistent with a larger number of users or under non-central conditions. Consequently, further investigation is needed to determine whether these limitations can be overcome and whether this method remains effective for more than three users.

Additionally, our experiments solely considered gaze. Further tests are required to ascertain whether these minute differences in eye contact can be perceived during a dialogue with a Virtual Human (VH). To enable these experiments, we need to equip the VH with the ability to recognize user speech and maintain eye contact with the correct user during conversation. Our objective is to test whether this enhancement enables users to detect these subtle differences.

Acknowledgments

This research is supported Year 2021 Culture Technology

R&D Program by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency(Project Number: R2021040269)

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] D. Bohus, E. Horvitz, Facilitating multiparty dialog with gaze, gesture, and speech, in: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, 2010, pp. 1–8.

[3] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al., Simsensei kiosk: A virtual human interviewer for healthcare decision support, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014, pp. 1061–1068.

[4] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, H. Yan, More than just a pretty face: conversational protocols and the affordances of embodiment, Knowledge-based systems 14 (2001) 55–64.

[5] J. Cassell, K. R. Thorisson, The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents, Applied Artificial Intelligence 13 (1999) 519–538.

[6] D. Aneja, R. Hoegen, D. McDuff, M. Czerwinski, Understanding conversational and expressive style in a multimodal embodied conversational agent, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–10.

[7] C. Oertel, M. Włodarczak, J. Edlund, P. Wagner, J. Gustafson, Gaze patterns in turn-taking, in: Thirteenth annual conference of the international speech communication association, 2012.

[8] S. A. Moubayed, J. Edlund, J. Beskow, Taming mona lisa: communicating gaze faithfully in 2d and 3d facial projections, ACM Transactions on Interactive Intelligent Systems (TiiS) 1 (2012) 1–25.

[9] E. Boyarskaya, A. Sebastian, T. Bauermann, H. Hecht, O. Tüscher, The mona lisa effect: Neural correlates of centered and off-centered gaze, Human brain mapping 36 (2015) 619–632.

[10] K. Otsuka, Mmspace: Kinetically-augmented telepresence for small group-to-group conversations. in 2016 ieee virtual reality (vr), online), DOI 10 (2016) 19–28.

[11] M. Vázquez, Y. Milkessa, M. M. Li, N. Govil, Gaze by semi-virtual robotic heads: Effects of eye and head motion, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 11065–11071.

[12] H. Mitake, T. Ichii, K. Tateishi, S. Hasegawa, Wide viewing angle fine planar image display without the mona lisa effect (2019).

[13] H.-H. Wu, H. Mitake, S. Hasegawa, Eye-gaze control of virtual agents compensating mona lisa effect, HAI (2018).

[14] C. Ware, K. Arthur, K. S. Booth, Fish tank virtual reality, in: Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, 1993, pp. 37–42.

[15] G. Jocher, A. Chaurasia, J. Qiu, YOLO by Ultralytics, 2023. URL: https://github.com/ultralytics/ultralytics.