Mutual gaze detection and estimation: towards human-robot interaction

Vidya Somashekarappa, Christine Howes and Asad Sayeed
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{vidya.somashekarappa,christine.howes,asad.sayeed}@gu.se

Abstract

Mutual gaze is an important part of social interaction and the perception of others emotional states and essential for establishing joint attention. It has been described as "the most powerful mode of developing a communicative link between humans". Although gaze estimation and prediction using machine learning and computer vision is extensively studied in images and videos for automation, there is a lack of research in dialogue and interaction. In this study, we estimate gaze using a PyTorch-based model with the goal of implementing human-like mutual gaze in a robot.

1 Introduction

Eye gaze supports and augments other social behaviours such as speech and gesture, and mental states or cognitive effort can substantially influence gaze. Since speech is a dominant mode of communication in human interactions, it is not feasible to separate gaze from speech in face-to-face human-human dialogue, and we should not do so for human-robot dialogue either. Researchers have shown that gaze improves speech-based interactions, e.g., disambiguating object references, maintaining engagement, conversation and narration, guiding attention, managing partners, influencing turn-taking (Kaiser et al., 2003; Rapp et al., 2021; Somashekarappa et al., 2021)

1.1 Mutual Gaze in Human Interaction

Mutual gaze occurs from birth when infants gaze at their caregivers. The field of vision of the newborns is approximately the distance required to make eye contact when held by an adult (Stern et al., 1985) and they prefer to look at faces over stimuli that engage them in mutual attention.

A study investigated if mutual gaze would induce feeling of romantic love. Subjects who gazed at their partners' eyes and whose partner was gazing back reported significantly higher feelings of affection, dispositional love and liking (Farroni et al., 2002).

1.2 Mutual Gaze in Human-Robot Interaction

In everyday situations, gaze is not only reactive, but also anticipates and predictes others' behaviour. In such scenarios, gaze is highly informative about intentions and upcoming decisions. An investigation into whether a humanoid robot's mutual or averted gaze influenced how people strategically reason in social decision making, after playing a strategic game with the robot iCub, revealed that participants were slower to respond when iCub established mutual attention before the decision. When people are sensitive to the mutual gaze of an artificial agent, they feel more engaged with the robot (Belkaid et al., 2021).

Robot gaze acts as a strong social signal for humans, modulating response times and decision threshold, promoting neural synchronization, and influencing choice strategies and sensitivity to outcomes. This has strong implications for robotics and clinical applications for all contexts involving human-robot interactions.

2 Aims of the study

- Estimate mutual gaze using neural networks
- Investigate effect of mutual gaze on agreement and disagreement in interaction.
- Understand the uncanny valley effect caused by eerie mutual attention.

3 Gaze Estimation

Gaze estimation aims to predict where the person is looking at by estimating the horizontal and vertical coordinates of the gaze target on a 2-D screen. Deep learning has revolutionised many computer

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

Figure 1: P1 gaze on P2

Figure 2: P2 gaze on P1

vision tasks; however, there is still a lack of guidelines for designing algorithms for gaze estimation in interaction. The GHI corpus (Lavia et al., 2018; Somashekarappa et al., 2020) has speech and gaze annotations in dyadic dialogues which was used in this study for automatic gaze detection.

The main approaches for gaze estimation that currently exist are deep learning-based (Cheng et al., 2021), headmap activated multimodal gaze estimation (Sinha et al., 2021), robust CNN model (Abdelrahman et al., 2022), and U-Net style multistream gaze estimation (synthetic to real transfer learning) (Mahmud et al., 2022). The benchmark datasets are MPIIGaze, Eyediap and UTMultiview.

3.1 Eye gaze classifier

The dataset contains 24 videos, capturing the frontal view of each participant, thus containing two different videos for each session. The videos were recorded at 30 high-definition frames per second.

The vector features were extracted from each video by the PyTorch implementation of MPII face gaze for AlexNet and ResNet14 [1]. For facial landmark detection, a pretrained dlib model was used. The processed video provided landmarks, head pose, projected points of the 3D face model, and a face bounding box. Every frame of the video containing gaze estimation coordinates was then extracted and time stamped. The mutual gaze incident is determined based on the overlapping averted gaze.

[1] https://github.com/hysts/pytorch mpiigaze

3.2 Mutual gaze during agreement and disagreement

Consistent with previous research, we noted that the participants looked at their partner more when listening than speaking. The magnitude of this listening-speaking difference depended on agreement condition, disagreement (but not agreement) exacerbated the maintaining mutual gaze, particularly by averting gaze.

4 Discussion

It is tempting to assume that perfectly matching robot gaze behaviors to human gaze behaviors will elicit identical responses from people, but this is not always the case. Several studies suggest that gaze from robots is interpreted differently than gaze from humans. In general, it is difficult to compare robot gaze to human gaze directly, because while robot gaze can be infinitely controlled, human gaze tends to have small, unpredictable variations.

Once the conversation has begun, conversational fluidity is managed as much by the absence of mutual gaze as by its presence. Virtual agents using gaze aversions for these conversational functions are more successful at regulating the conversational flow and elicit greater disclosure from people than agents that do not perform gaze aversions or perform gaze aversions at inappropriate times (Andrist et al., 2013). Expressive robots could take advantage of these fine-grained gaze behaviors to convey mental states—for example, when they are thinking, when they are waiting for a response, or when they are experiencing difficulty—in a natural and human-like way. For future work, we will conduct a blink estimation study as well as an interactive human-robot experiment.

Acknowledgements

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. 2022. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*.

Sean Andrist, Bilge Mutlu, and Michael Gleicher. 2013. Conversational gaze aversion for virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 249–262. Springer.

Marwen Belkaid, Kyveli Kompatsiari, Davide De Tommaso, Ingrid Zablith, and Agnieszka Wykowska. 2021. Mutual gaze with a robot affects human neural activity and delays decision-making processes. *Science Robotics*, 6(58):eabc5044.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*.

Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H Johnson. 2002. Eye contact detection in humans from birth. *Proceedings of the National academy of sciences*, 99(14):9602–9605.

Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 12–19.

Lisa Lavia, Harry J. Witchel, Francesco Aletta, Jochen Steffens, André Fiebig, Jian Kang, Christine Howes, and Patrick G. T. Healey. 2018. Non-participant observation methods for soundscape design and urban planning. In Francesco Aletta and Jieling Xiao, editors, *Handbook of Research on Perception-Driven Approaches to Urban Assessment and Design*. IGI Global.

Zunayed Mahmud, Paul Hungler, and Ali Etemad. 2022. Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning. *arXiv preprint arXiv:2206.09256*.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.

Neelabh Sinha, Michal Balazia, and François Bremond. 2021. Flame: Facial landmark heatmap activated multimodal gaze estimation. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 759–765.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2021. A deep gaze into social and referential interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Daniel Stern, Lynne Hofer, Wendy Haft, John Dore, Tiffany M Field, and Nathan A Fox. 1985. Social perception in infants.