Fovea Prediction Model in VR

Daniele Giunchi
University College London

Riccardo Bovo
Imperial College London

Nitesh Bhatia
Imperial College London

Thomas Heinis
Imperial College London

Anthony Steed
University College London, UK

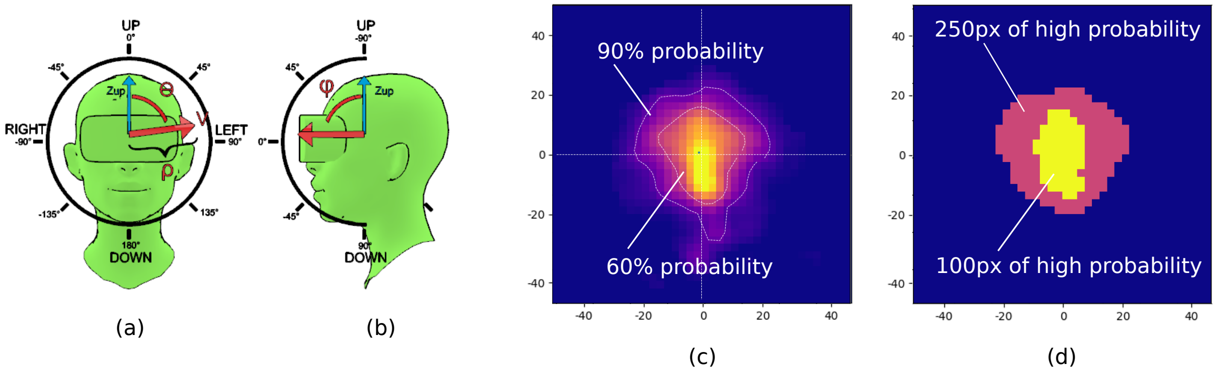


Figure 1: (a) Head rotation velocity and angle are measured in VR. (b) Vertical head tilt is also recorded. (c) Eye-gaze patterns are analyzed to show 90% and 60% likelihood areas, using head coordinates for longitude and latitude. (d) Analysis identifies two gaze-probability regions: the top 100 pixels (yellow) and the next 150 pixels (orange), aiding selective rendering.

Abstract

We propose a lightweight deep learning approach for gaze estimation representing the visual field as three distinct regions: fovea, near, and far peripheral. Each region is modelled using a gaze parameterization gaze regarding angle-magnitude, latitude, or a combination of angle-magnitude-latitude. We evaluated how accurately these representations can predict a user's gaze across the visual field when trained on data from VR headsets. Our experiments confirmed that the latitude model generates gaze predictions with superior accuracy with an average latency compatible with the demanding real-time functionalities of an untethered device. We generated an outperforming ensemble model with a comparable latency.

Keywords: neural networks, gaze prediction, visual attention

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing methodologies—Machine learning—Machine learning approaches—Neural networks

1 Introduction

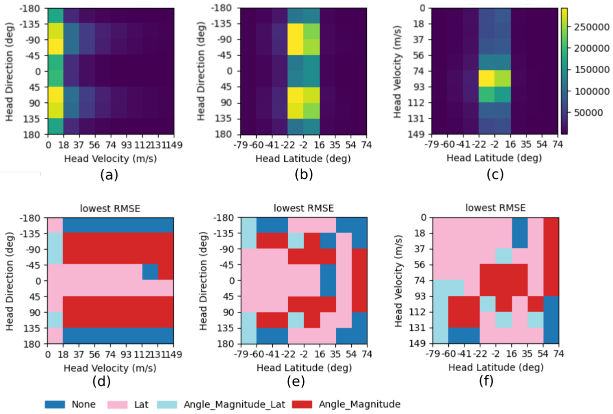
The human eye is designed to interpret visual signals with high spatial resolution in the centre of the retina (the fovea), and this resolution declines with eccentricity (angular distance from the centre), which is one of the essential characteristics of the human visual system. This functionality has created rendering or image compression models to minimise the data needed for display. Foveated rendering reduces computation and can be used in interactive 3D graphics for 2D displays or virtual reality headsets, video compression, and video reconstruction by in-painting, among other applications [3, 11, 10, 6] especially in virtual reality (VR) headset [4, 8, 9, 7]. This study explores deep-learning gaze-prediction models for fovea prediction in VR. It relies exclusively on head movements, making it suitable for real-time deployment on affordable HMDs. Our choice of the multi-perceptron (MLP) neural network (NN) is informed by its successful results in Bovo's work [2]. Our MLP architecture is a fully connected layer designed to learn the central fixation points and the outlines of probable fixation areas. These areas are predicted by such models as shown in Figure . Rather than pinpointing the precise location of a user's gaze, our model outlines the three regions of the visual fields: fovea, near peripheral, and far peripheral. This approach addresses the inherent limitations of head movements. It minimizes the visual footprint of head-based cues, which is particularly valuable in immersive scenarios where consumer-market HMDs do not include an eye-tracker for cost reasons. Our model does not depend on visual saliency and can be deployed in various environments and scenarios. Predicting the location of such areas in a virtual reality (VR) immersive scenario gives several advantages.

We created three distinct fovea estimation models characterized by the inputs: latitude, velocity angle-magnitude, and latitude and angle-magnitude, each designed to tackle the challenges inherent to accurate visual field prediction in extended reality (XR) environments. To gauge their efficacy, we conducted a comprehensive evaluation focusing on two pivotal metrics: accuracy and latency. We trained these models to the GIW [5] and panoramic 360° [1] datasets. Building upon these individual model evaluations, we created an ensemble model. This ensemble approach effectively harnessed the strengths of each model, forging a comprehensive solution that not only improved fovea prediction accuracy but also maintained a low latency value.

2 Fovea Prediction Models

Our model library included three variations of the fovea prediction. We developed the "Angle + Magnitude" model, which accepts in input the head velocity angle and magnitude. In contrast, the "Latitude" model used the pitch information from head movement. The

. alt.



(a) (b) (c)

(d) (e) (f)

alt

Figure 2: (a)(b)(c) plots depicting the density of samples across combinations of head rotational direction (θ) velocity magnitude (ρ) and head latitude (ϕ). (e)(f)(g) plots depicting the best performing model cross combinations of head rotational direction (θ) velocity magnitude (ρ) and head latitude (ϕ).

”Latitude + Angle + Magnitude” model, uses all three parameters. The model employed is an MLP neural network tailored for gaze estimation. We introduced the head’s pitch parameter compared to Bovo’s study [2] referred to as ”latitude”(ϕ). The latitude, as a factor in our predictive model, played a crucial role in improving the accuracy of gaze prediction. Latitude directly influences the alignment of the user’s gaze and, subsequently, their foveal vision. By incorporating latitude into our model, we could account for variations in how users tilt their heads when exploring virtual environments, leading to more precise and robust fovea predictions. Our research explored two distinct predictions that can be made using this model. The first one is a probability model of the vision field, which characterizes the likelihood of a user’s gaze falling within the foveal region at a given time. Secondly, we aim to estimate the pixel area associated with the foveal region and near the peripheral area. This pixel area prediction is a valuable tool for controlling the allocation of computing resources. We used the pixel area prediction method to assign a fixed number of pixels for the foveal region and near the peripheral area. This approach allows efficient resource consumption management while maintaining a high accuracy in gaze prediction. It optimizes computational efficiency and bandwidth allocation, ensuring users’ seamless and immersive experience.

3 Results

In our study, we comprehensively evaluated three distinct gaze estimation models in the context of virtual reality. We assessed their performance using an average fixation map (AFM) derived from the collective gaze data. The visual field was divided into a grid structure of 16x16, totalling 256 cells. We determined the most suitable gaze estimation model for each cell, ultimately generating a map highlighting the best-performing model across the visual space. Our assessment considered three key parameters: head velocity, head latitude, and head direction, allowing us to create maps that provide insights into the model’s effectiveness under different conditions. Additionally, we incorporated a weighted evaluation method to account for the non-uniform distribution of samples across the visual space. This approach ensured that the evaluation results accurately reflected real-world scenarios, where some cells might contain a greater number of samples than others. This comprehensive evaluation showed that the latitude-based model outperformed the other models, showcasing its effectiveness in estimating fovea behaviour. Furthermore, we examined the latency of all models. We determined that they were compatible with low-latency processing, making them suitable for deployment on mobile devices and enhancing the overall user experience in virtual reality applications.

4 Conclusion

Our study has presented a lightweight deep learning approach for gaze estimation in virtual reality, focusing on three distinct regions of the visual field: fovea, near peripheral, and far peripheral. We evaluated the effectiveness of three fovea prediction models, each tailored to address the unique challenges of gaze estimation in extended reality environments. Our extensive evaluation was based on key parameters such as head velocity, head latitude, and head direction, and we employed an average fixation map (AFM) to compare the models’ performance. To ensure the reliability of our evaluation, we implemented a weighted approach that considered the non-uniform distribution of samples across the visual space, addressing real-world scenarios where certain areas may contain more data. Our findings demonstrated that the latitude-based model consistently outperformed the other models, emphasizing its accuracy in predicting foveal gaze behaviour. Furthermore, we assessed the latency of all models. We found them compatible with low-latency processing, making them well-suited for deployment on mobile devices and enhancing the user experience in virtual reality applications.

References

[1] I. Agtzidis, M. Startsev, and M. Dorr. 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, p. 1007–1015. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3343031.3350947 1

[2] R. Bovo, D. Giunchi, L. Sidenmark, H. Gellersen, E. Costanza, and T. Heinis. Real-time head-based deep-learning model for gaze probability regions in collaborative vr. In *2022 Symposium on Eye Tracking Research and Applications*, pp. 1–8, 2022. 1, 2

[3] W. S. Geisler and J. S. Perry. Real-time foveated multiresolution system for low-bandwidth video communication. *Human Vision and Electronic Imaging III*, 3299:294–305, 1998. doi: 10.1117/12.320120 1

[4] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3D graphics. *ACM Transactions on Graphics*, 31(6):1–10, 2012. doi: 10.1145/2366145.2366183 1

[5] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):2539, 2020. 1

[6] S. Lee, M. S. Pattichis, and A. C. Bovik. Foveated video compression with optimal rate control. *IEEE Transactions on Image Processing*, 10(7):977–992, 2001. doi: 10.1109/83.931092 1

[7] X. Meng, R. Du, M. Zwicker, and A. Varshney. Kernel Foveated Rendering. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1):1–20, 2018. doi: 10.1145/3203199 1

[8] A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Benty, A. Lefohn, and D. Luebke. Perceptually-based foveated virtual reality. *ACM SIGGRAPH 2016 Emerging Technologies, SIGGRAPH 2016*, pp. 7–8, 2016. doi: 10.1145/2929464.2929472 1

[9] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):1–12, 2016. doi: 10.1145/2980179.2980246 1

[10] S. Rimac-Drlje, G. Martinović, and B. Zovko-Cihlar. Foveation-based content adaptive structural similarity index. *International Conference on Systems, Signals, and Image Processing*, (165):401–404, 2011. 1

[11] O. Wiedemann, V. Hosu, H. Lin, and D. Saupe. Foveated Video Coding for Real-Time Streaming Applications. *2020 12th International Conference on Quality of Multimedia Experience, QoMEX 2020*, 2020. doi: 10.1109/QoMEX48832.2020.9123080 1