A Gaze Prediction Model for Task-Oriented Virtual Reality

Konstantina Mammou and Katerina Mania
School of Electrical and Computer Engineering, Technical University of Crete, Greece

Contact: kmammou@tuc.gr , amania@tuc.gr, surreal.tuc.gr

INTRODUCTION

- **Gaze prediction in Virtual Reality (VR)** can assist or replace eye trackers in techniques such as **foveated rendering**, thus aiding in solving the **latency** issues they suffer from [1].
- The main **challenge** is the **dynamic** and **immersive nature** of **VR**, especially in **real-time**, **task-oriented** applications such as **games**.
- It is also difficult to create a model that can predict across **different tasks**, since that requires training with a huge amount of **diverse** data.
- Finally, **integration** into gaze-contingent rendering pipelines is challenging, as they require **fast** and **efficient** computations; any delay can disrupt the user's immersion.

This work investigates the role and potential of **temporal continuity** in enabling accurate predictions in **diverse task categories**. Our model reduces input complexity while maintaining **high prediction accuracy**. Evaluated on the **OpenNEEDS** dataset, it significantly outperforms baseline methods. The model demonstrates **strong potential** for **integration** into **gaze-based VR interactions** and **foveated rendering pipelines**.

RELATED WORK

- **DGaze** [2] achieves **real-time CNN-based gaze prediction** in dynamic scenes under free-viewing conditions but struggles with **task-oriented scenarios**.
- **FixationNet** [3] focuses on forecasting eye fixations during **specific visual search tasks in VR**. Consequently, this model cannot be directly applied to different tasks.

OVERVIEW

It is proven that temporal continuity [4] is particularly evident in task-oriented VR environments, **where gaze is frequently guided by specific goals** [5]. Thus, our model uses a relatively simple input only of sequences of **past frames and gaze points** to learn gaze behaviour patterns over time. The proposed architecture consists of three modules:

(1) the Image Sequence Module (ISM) to capture temporal motion features from consecutive frames,
(2) the Gaze Sequence Module (GSM) to learn temporal gaze patterns, and
(3) the Gaze Fusion Module (FM) that integrates both outputs to predict a single gaze point.

METHODOLOGY

• Dataset

We used the **OpenNEEDS** dataset [6], a newly published, large-scale, high frame rate, comprehensive and open-source dataset designed for gaze and interaction research in VR. Its key features are:

(a) 44 participants exploring two different (indoor and outdoor) environments,
(b) A variety of tasks: reading, throwing, drawing, aiming, shooting, object manipulation, and
(c) Continuous data samples, recorded for up to 5 minutes per participant.

• Pre-processing

Main steps:

(a) Normalization of frames to the [0,1] range,
(b) Conversion of gaze vectors to **2D visual angles** and normalization,
(c) Outlier removal using the **Interquartile Range Method (IQR)**,
(d) Creation of input sequences: each sequence consists of consecutive frames and their corresponding gaze points.

• Model architecture

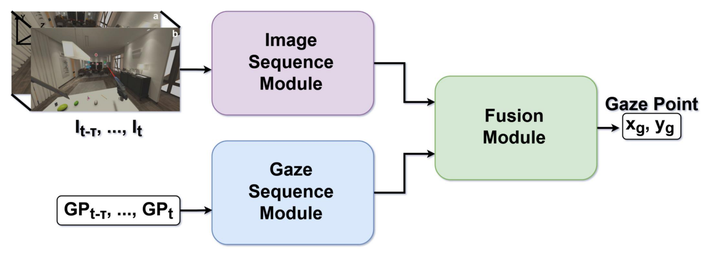

Figure 1: *Model Architecture.*

(1) ISM: It consists of **5 ConvLSTM2D** layers (ReLU), **4 MaxPooling** layers for dimensionality reduction, and 4 fully connected (**FC**) layers. A **dropout** layer prevents overfitting. The input is a sequence of 10 continuous frames.
(2) GSM: It processes a sequence of 10 continuous gaze points with **4 LSTM** layers (ReLU), **2 FC** layers, and a **dropout** layer.
(3) FM: It **merges** the ISM and GSM outputs using a **maximum** operation, followed by a **FC** and a **dropout** layer. The **final FC layer** (size 2, **sigmoid**) outputs the predicted gaze point (x_g, y_g), within the range [0,1].

The model is trained using **Mean Absolute Error (MAE)** as the loss function and the Adam optimizer with an initial learning rate of 0.001. Training was performed on Google Colab using the **NVIDIA L4 Tensor Core GPU**, with a batch size of 64 for 10 epochs.

RESULTS

Following the approach of Hu et al.[2], the model was evaluated based on its **prediction error**, **recall rate** (to assess its potential for integration into foveated rendering pipelines), and **runtime** performance. Two baselines — **center and mean**— were defined for comparison. For the recall rate, a foveal radius of **15 degrees**, centered at the ground truth gaze point, was applied.

The model achieved a **low median error** with a **narrow interquartile range**, indicating **robustness**, **accuracy**, and **consistency** (Figure **2**).
It demonstrated a **66.43%** and **63.08% improvement** over the **center** and **mean** baselines, respectively. Additionally, the model significantly outperformed the baselines in terms of **recall rate**, achieving values suitable for practical applications (Table **1**). Observing the visualised results (Figure **3**), we notice that the predictions follow the pattern of the ground truth closely, even without achieving perfect accuracy. However, the **average runtime** of approximately 150 ms remains a notable **limitation**, affecting its viability for real-time use.

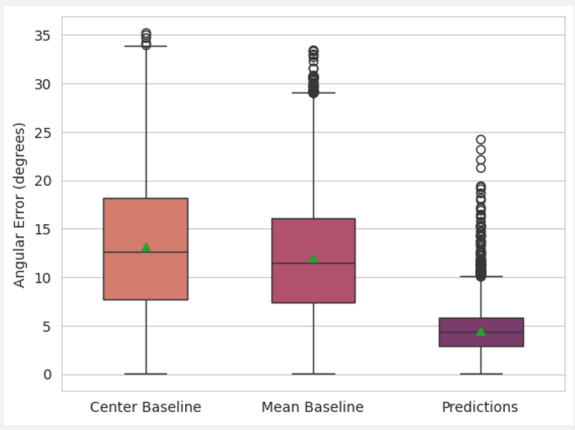

Figure 2: *Angular Error comparison between our model and the baselines.*

	Center	Mean	Model
Mean recall rate	60.8%	69.46%	99.87%

Table 1: *Recall rates of our model and the baselines.*

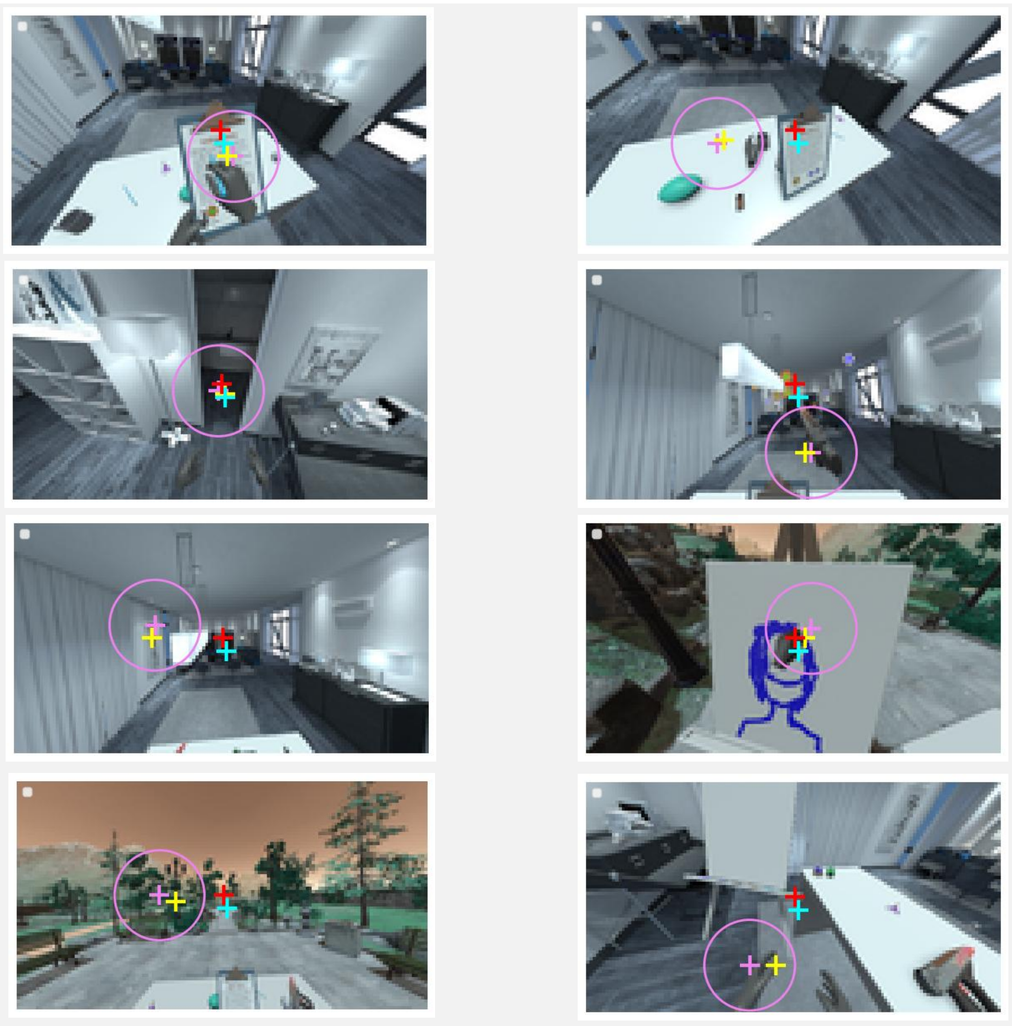

Figure 3: *The purple cross denotes the ground truth gaze position, with the circle illustrating the foveal region with radius 15 degrees. The yellow cross represents the prediction of our model, the red cross shows the center baseline and the blue the mean baseline.*

Acknowledgment
This work has been supported by the Horizon Europe Research & Innovation Programme under Grant agreement N. 101092612 (Social and hUman ceNtered XR - SUN project).

REFERENCES

[1] ARABADZHIYSKA E., TURSUN C., SEIDEL H.-P., DIDYK P.: Practical saccade prediction for head-mounted displays: Towards a comprehensive model. ACM Trans. Appl. Percept. 20, 1 (Jan. 2023).
[2] HU Z., LI S., ZHANG C., YI K., WANG G., MANOCHA D.: Dgaze: Cnn-based gaze prediction in dynamic scenes. IEEE transactions on visualization and computer graphics 26, 5 (2020)
[3] HU Z., BULLING A., LI S., WANG G.: Fixationnet: Forecasting eye fixations in task-oriented virtual environments. IEEE Trans actions on Visualization and Computer Graphics 27, 5 (2021).
[4] HU Z., LI S., GAI M.: Temporal continuity of visual attention for future gaze prediction in immersive virtual reality. Virtual Reality & Intelligent Hardware 2, 2 (2020).
[5] KOULIERIS G. A., DRETTAKIS G., CUNNINGHAM D., MANIA K.: Gaze prediction using machine learning for dynamic stereo manipulation in games. In 2016 IEEE Virtual Reality (VR) (2016).
[6] EMERY K. J., ZANNOLI M., WARREN J., XIAO L., TA LATHI S. S.: Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In ACM Symposium on Eye Tracking Research and Applications (2021)