Eye Fixation Forecasting in Task-Oriented Virtual Reality

Zhiming Hu*
Peking University

ABSTRACT

In immersive virtual reality (VR), users' visual attention is crucial for many important applications, including VR content design, gaze-based interaction, and gaze-contingent rendering. Especially, information on users' future eye fixations is key for intelligent user interfaces and has significant relevance for many areas, such as visual attention enhancement, dynamic event triggering, and human-computer interaction. However, previous works typically focused on free-viewing conditions and paid less attention to task-oriented attention. This paper aims at forecasting users' eye fixations in task-oriented virtual reality. To this end, a VR eye tracking dataset that corresponds to different users performing a visual search task in immersive virtual environments is built. A comprehensive analysis of users' eye fixations is performed based on the collected data. The analysis reveals that eye fixations are correlated with users' historical gaze positions, task-related objects, saliency information of the VR content, and head rotation velocities. Based on this analysis, a novel learning-based model is proposed to forecast users' eye fixations in the near future in immersive virtual environments.

Index Terms: Fixation forecasting—Visual attention—Visual search—Eye tracking; Deep learning—Convolutional neural network—Virtual reality

1 INTRODUCTION

Immersive virtual reality (VR) provides users with high sense of presence and has become an important 3D user interface. Human visual attention in immersive VR is key for many important applications, such as VR content design [10], VR content compression [10], gaze-based interaction [8], gaze guidance, and gaze-contingent rendering [4, 5]. Especially, information on users' future eye fixations is valuable for intelligent user interfaces and has significant relevance for plenty of areas, such as visual attention enhancement, pre-computation of gaze-contingent rendering [4], dynamic event triggering, and human-computer interaction. However, prior works typically focused on free-viewing conditions (no specific task) [1,4,5,10] and few works have studied task-oriented situations, which are more challenging but also more practically relevant [3, 6].

To address the limitations of existing methods, a learning-based model is proposed to forecast users' eye fixations in task-oriented virtual environments. Specifically, this work focuses on visual search, which is a frequent and important routine behavior in people's daily life, e.g. when looking for a friend in a crowd, trying to find your smartphone, or searching for food in the fridge. Visual search has become an active area of vision research in the past few decades [11]. However, most of the findings about visual search are derived from 2D viewing conditions while visual search in immersive virtual reality has not been fully explored.

This research starts with building an eye tracking dataset that corresponds to different users performing a visual search task in immersive virtual environments. The characteristics of users' eye fixations are analyzed and the results reveal that human eye fixations are correlated with users' historical gaze positions, task-related objects, saliency information of the VR content, as well as head rotation velocities. Based on the analysis, a novel learning-based model is proposed to forecast users' eye fixations in the near future in immersive virtual environments.

This work makes the following contributions:

- A novel learning-based model for forecasting eye fixations in task-oriented virtual reality;
- A comprehensive analysis of users' visual attention during visual search in VR;
- A new task-oriented VR eye tracking dataset.

2 RELATED WORK

2.1 Computational Modeling of Visual Attention

Computational modeling of visual attention is an active area of vision research and many visual attention models have been proposed in the past few decades. Specifically, models of visual attention can be classified into bottom-up models and top-down models [6]. Bottom-up models are geared to free-viewing conditions, in which subjects are asked to freely observe the stimuli, e.g. images and videos, and are assigned no specific task. These models employ low-level image features such as intensity, contrast, color, and orientation to predict human visual attention [7]. In contrast, top-down models aim at task-oriented situations, where users' visual attention is influenced by a specific task. Top-down models predict visual attention by utilizing high-level image features like specific tasks and scene context [9]. Compared with free-viewing conditions and bottom-up models, the research on task-oriented situations and top-down models is relatively limited for the reason that task-oriented situations are much more challenging than free-viewing conditions. Given that task-oriented situations are more practically relevant, this work focuses on task-oriented virtual environments and presents a novel top-down visual attention model.

2.2 Gaze Prediction in Virtual Reality

Gaze prediction in virtual reality has also been explored by many researchers in recent years. Sitzmann et al. focused on 360° images [10]. They analysed the characteristics of users' gaze behaviors and adapted existing saliency predictors to predict saliency maps of the scenes. Xu et al. presented research on 360° videos [12]. They proposed a model to predict gaze displacement by extracting features from 360° video frames. Hu et al. conducted a comprehensive analysis of human gaze behaviors in immersive virtual environments and proposed an eye-head coordination model to predict users' real-time gaze positions in static virtual scenes [5]. Recently, Hu et al. proposed a learning-based model to predict users' gaze positions in dynamic virtual scenes [4]. However, existing methods on gaze prediction in VR are typically derived from free-viewing conditions and their performances will deteriorate when applied to task-oriented situations [4]. Considering the limitations of existing methods, this research proposes a novel learning-based model to forecast eye fixations in task-oriented virtual reality.

*e-mail: jimmyhu@pku.edu.cn

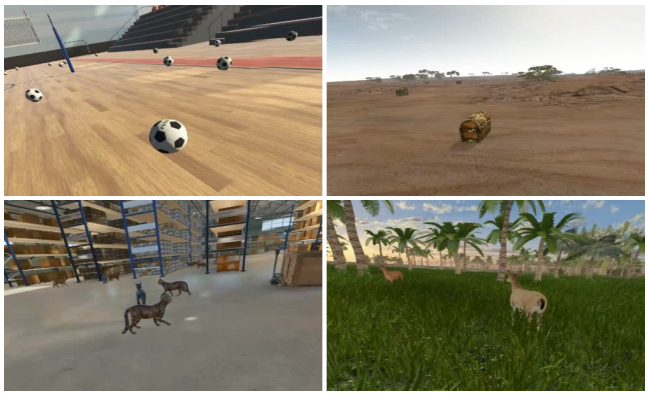

Figure 1: Four immersive virtual environments used in the data collection process, containing two static scenes (top) and two dynamic scenes (bottom).

3 Current Research

3.1 Gaze Data Collection

To collect users' gaze data, four immersive virtual environments were utilized as the stimuli (Fig. 1), containing two static scenes (a gym and a desert) and two dynamic scenes (a warehouse and a tropical island). The static scenes contain some static objects, i.e. chests and footballs, while the dynamic scenes have some dynamic animals, i.e. deer and cats that are set to wander in the environments in a random manner. Participants were asked to perform a visual search task, i.e. searching for a particular type of object, in the virtual environment. Participants' gaze data was collected using an eye tracker; their head rotation velocities were obtained from HTC Vive; the scene content viewed by the observers was recorded by a screen-recorder; the information on the task-related objects was collected using a Unity script.

3.2 Fixation Analysis

Based on the collected data, a comprehensive analysis is performed to reveal the characteristics of users' task-oriented visual attention in immersive virtual reality. Users' fixation positions were extracted from the raw gaze data and the distribution of the fixation positions was analysed. The result shows that most of users' eye fixations lie in the central region of the screen. The correlations between users' eye fixations and other factors were also analysed. The results indicate that the fixation positions are highly correlated with users' historical gaze positions, as well as with task-related objects, saliency information of the VR content, and users' head rotation velocities. The analysis provides meaningful insights for establishing fixation prediction models.

3.3 Fixation Forecasting Model and Results

Based on the analysis, a novel learning-based model was proposed to forecast eye fixations in task-oriented virtual environments [2]. This model consists of a feature extraction network, which extracts features from VR images, historical gaze data, task-related data, and head data, and a fixation prediction network, which employs the extracted features to forecast users' eye fixations.

Extensive experiments were conducted to evaluate the performance of the proposed model. A cross-user evaluation and a cross-scene evaluation were employed to compare the proposed model with the state-of-the-art method [4] both on the collected data and a free-viewing VR eye tracking dataset [4]. The experimental results demonstrate that the proposed model outperforms the state-of-the-art method in both task-oriented situations and free-viewing conditions by a large margin.

4 Future Research

This research aims at forecasting human eye fixations in task-oriented virtual environments. However, the current work only focuses on a visual search task and only reveals the influences of previous gaze, scene content, task-related objects, and head movements. Considering these limitations, there is still plenty of room to improve this work:

- **Other Factors:** Human eye fixations may also have correlations with other factors, such as sound, users' mental states, users' gestures, and users' behavioral habits. Taking these factors into consideration may help derive a more precise fixation forecasting model.
- **Other Tasks:** Different tasks require different task-specific gaze behaviors. Exploring users' gaze behaviors when performing other tasks, e.g. text editing or assembly task, is an interesting avenue for future work.
- **Application of the Model:** The information on future eye fixations is crucial for intelligent user interfaces. Applying the proposed fixation forecasting model to intelligent systems is of great significance for many relevant areas.
- **Other Systems:** This research is currently limited to immersive VR systems and it has the potential to be converted to other systems such as augmented reality system and mixed reality system.

References

[1] Z. Hu. Gaze analysis and prediction in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020.

[2] Z. Hu, A. Bulling, S. Li, and G. Wang. Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[3] Z. Hu, S. Li, and M. Gai. Temporal continuity of visual attention for future gaze prediction in immersive virtual reality. *Virtual Reality & Intelligent Hardware*, 2020.

[4] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 2020.

[5] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE transactions on visualization and computer graphics*, 25(5):2002–2010, 2019.

[6] L. Itti. *Models of bottom-up and top-down visual attention*. PhD thesis, California Institute of Technology, 2000.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[8] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014.

[9] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2007.

[10] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics (IEEE VR 2018)*, 24(4):1633–1642, 4 2018.

[11] J. M. Wolfe and T. S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017.

[12] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360 immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5333–5342, 2018.