DGaze: CNN-Based Gaze Prediction in Dynamic Scenes

Zhiming Hu, Sheng Li*, Congyi Zhang, Kangrui Yi, Guoping Wang*, Dinesh Manocha

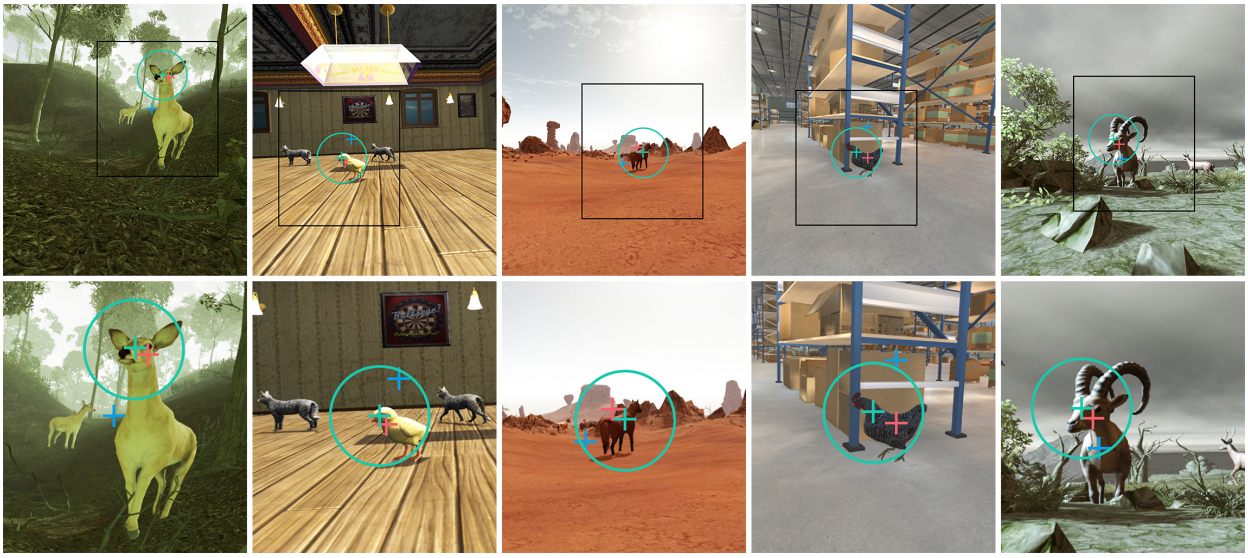

Fig. 1: Gaze prediction performance of our model (DGaze) in different scenes. The upper row shows some captured images from an HMD's screen and the lower row exhibits the corresponding enlarged view. The green cross denotes the ground truth of gaze position, the blue cross is generated using SGaze [21] algorithm, and the red cross represents the result of our novel algorithm, DGaze. The green circle illustrates the foveal region with radius 15°. In practice, our gaze prediction algorithm exhibits higher accuracy than prior method.

Abstract—We conduct novel analyses of users' gaze behaviors in dynamic virtual scenes and, based on our analyses, we present a novel CNN-based model called DGaze for gaze prediction in HMD-based applications. We first collect 43 users' eye tracking data in 5 dynamic scenes under free-viewing conditions. Next, we perform statistical analysis of our data and observe that dynamic object positions, head rotation velocities, and salient regions are correlated with users' gaze positions. Based on our analysis, we present a CNN-based model (DGaze) that combines object position sequence, head velocity sequence, and saliency features to predict users' gaze positions. Our model can be applied to predict not only realtime gaze positions but also gaze positions in the near future and can achieve better performance than prior method. In terms of realtime prediction, DGaze achieves a 22.0% improvement over prior method in dynamic scenes and obtains an improvement of 9.5% in static scenes, based on using the angular distance as the evaluation metric. We also propose a variant of our model called DGaze_ET that can be used to predict future gaze positions with higher precision by combining accurate past gaze data gathered using an eye tracker. We further analyze our CNN architecture and verify the effectiveness of each component in our model. We apply DGaze to gaze-contingent rendering and a game, and also present the evaluation results from a user study.

Index Terms—Gaze prediction, convolutional neural network, eye tracking, dynamic scene, gaze-contingent rendering, virtual reality

1 Introduction

With the development of virtual reality (VR) technology, users' gaze information in VR becomes increasingly important and can be used for different applications, including VR content design [39], VR content compression [39], eye movement-based interaction [14, 29, 34, 43], gaze behavior analysis [2, 4, 20, 21, 39], gaze-contingent rendering (or foveated rendering) [18, 33, 41, 42], etc. Currently, the most common solution for eye tracking is based on hardware-based eye trackers. Eye trackers are accurate and can be integrated with head mounted devices (HMDs). However, as an accessory equipment of an HMD, an eye tracker can be costly and lacks ease of use. Moreover, eye trackers are mainly used to measure the current gaze direction, and they cannot directly predict the future gaze position. Therefore, there is considerable interest in developing alternate methods for eye tracking and gaze prediction [21, 39, 45].

Human gaze behaviors in dynamic scenes are more intricate than that in static scenes because users' attention will be attracted not only by static stimuli but also by moving stimuli [1, 17]. In VR applications, dynamic scenes are much more common than static scenes. However, currently, many studies have focused on static scenes [21, 36, 39] while dynamic scenes have not been well explored. Therefore, it is very meaningful to analyze human gaze behaviors in dynamic scenes. One

- *Zhiming Hu, Sheng Li, Kangrui Yi, Guoping Wang are at Peking University, China. E-mail: {jimmyhu | lisheng | yikangrui | wgp}@pku.edu.cn.*
- *Congyi Zhang is at both The University of Hong Kong, China and Peking University, China. E-mail: cyzh@hku.hk.*
- *Dinesh Manocha is at University of Maryland, USA. E-mail: dm@cs.umd.edu.*
- *Sheng Li and Guoping Wang are the corresponding authors.*

- *Project URL: cranehzm.github.io/DGaze*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

of our goals is to analyze the gaze behavior for dynamic scene and use that behavior to design better prediction algorithms.

Our model is formulated based on the following procedures:

Gaze Data Collection: We record users' eye tracking data in dynamic scenes to analyze their gaze behaviors. Specifically, a total of 43 participants are asked to freely explore 5 dynamic virtual scenes including both indoor and outdoor scenes (See Fig. 2) and the participants are given no specific task. We collect users' data from the eye tracker to build a dataset. Our dataset includes 86 pieces of data from 43 users and each piece of data contains about 18,000 gaze positions, 18,000 object positions, 36,000 head velocities, and 10,800 frames of scene screenshots. The dataset has been released for public usage[1].

Gaze Behavior Analysis: We further analyze users' gaze behaviors based on our dataset. Specifically, we perform Spearman's rank correlation analysis [30] on the dataset and find that users' gaze positions are correlated with their head rotation velocities and the positions of dynamic objects. We also extract the saliency maps of the images viewed by observers and find that gaze positions correspond with salient regions. The above results reveal that head velocity, dynamic object position, and saliency feature facilitate the task of gaze prediction. The effectiveness of each component is further validated in our ablation study (Sect. 6.4). In addition, we analyze the distribution of users' gaze data and find that most of the gaze positions lie in the central region of the HMD's screen.

Novel CNN-Based Model: Based on our analyses, we propose a CNN-based model (DGaze) to predict gaze positions. Our model consists of three modules: sequence encoder module, saliency encoder module, and gaze prediction module. The sequence encoder module employs a 1D CNN layer to encode the object position sequence and the head velocity sequence. The saliency encoder module applies the state-of-the-art SAM-ResNet saliency predictor [13] to extract saliency features of the past images and utilizes a fully connected (FC) layer to encode the features. The gaze prediction module combines the outputs of the 2 modules and employs 3 FC layers to predict gaze positions. When an eye tracker is available, the sequence of past gaze data can also be used in the sequence encoder module to improve our model's performance. We refer to this revised model as DGaze_ET as it utilizes the data gathered using an eye tracker. We utilize 60% of our dataset (1045654 data points) to train the model and employ the remaining data (698304 data points) to test. The effectiveness of our CNN architecture is further validated (Sect. 6.4). The source code of our model and the pre-trained models are publicly available[1].

Evaluation of DGaze: We utilize the angular distance between the ground truth and the predicted gaze position as the evaluation metric (Sect. 6.1.2). DGaze achieves good realtime prediction performance on both dynamic (an improvement of 22.0% over prior method) and static datasets (9.5% improvement). DGaze can also be applied to predict gaze positions at any specified future time ($0 \sim 1000$ *ms*). When an eye tracker is provided, DGaze_ET can take advantage of accurate past gaze data to achieve higher performance. We apply DGaze to gaze-contingent rendering and conduct a two-alternative forced choice (2AFC) test to verify its effectiveness. Compared with prior method, DGaze is preferred by 68.9% of the users' responses and the result is statistically significant. We also apply DGaze to a game to evaluate its effectiveness and the result shows that DGaze outperforms prior method by 14.3% in terms of realtime prediction.

Overall, our contributions include:

- A novel CNN-based model (DGaze) for realtime and future gaze prediction in immersive HMD-based applications with high accuracy.
- Analyses of human gaze behaviors in dynamic virtual scenes that provide insights into formulating gaze prediction models.
- A dataset that contains 43 users' eye tracking data that is gathered using five dynamic scenes.

[1] cranehzm.github.io/DGaze

2 Related Work

In this section, we give a brief overview of prior works on gaze prediction, gaze behavior analysis, and the applications of eye tracking technology.

2.1 Gaze Prediction

In the area of vision research, gaze prediction or visual saliency prediction has been well-studied and many gaze prediction models have been proposed in the last three decades. Generally, most of the existing models are based on bottom-up approaches, top-down approaches, or hybrid approaches. Bottom-up approaches employ low-level image features such as intensity, color, and orientation to predict visual attention [10, 24] while top-down approaches take high-level features of the scene like specific tasks and context into consideration [8, 19]. Hybrid approaches combine low-level features and high-level features to obtain better performance [7, 32]. Recently, with advances in deep learning, many deep learning-based models have been proposed and have achieved good performances [13, 26].

However, there is limited work on gaze prediction in the area of virtual reality. Sitzmann et al. [39] and Rai et al. [36] both focused on saliency in 360° static images. They collected users' eye tracking data and predicted saliency maps of the scenes. Xu et al. [45] built a dataset that contains observers' gaze data in dynamic 360° videos and proposed a deep learning-based model for gaze displacement prediction. Koulieris et al. [25] predicted gaze object categories in a video game. Our approach is also related to the recent work on SGaze [21]. Hu et al. gathered the data of a large number of participants freely exploring static virtual scenes and proposed an eye-head coordination model called SGaze for predicting users' realtime gaze positions. In contrast with Hu et al.'s work, we focus on dynamic scenes and predict not only realtime gaze positions but also future gaze positions.

2.2 Gaze Behavior Analysis

There is some work on the analysis of human gaze behaviors. Itti [23] reported that human gaze behaviors are controlled by both a bottom-up mechanism and a top-down mechanism, which means visual attention is influenced not only by the content humans see but also by the tasks assigned to them. Pinto et al. [35] further revealed that the two mechanisms are independent. Baloh et al. [3] discovered many differences between horizontal and vertical eye movements and Rottach et al. [37] showed evidence for independent feedback control of horizontal and vertical saccades. Brockmann et al. [9] and Boccignone et al. [6] focused on gaze shifts and they modeled gaze shifts as a stochastic process. Franconeri et al. [17] revealed that moving stimuli capture visual attention and they further reported that looming stimuli capture attention while receding stimuli do not attract attention. The phenomenon that motion attracts attention was also observed in Abrams et al.'s work [1]. Yarbus [46] found that the eyes and the head move in coordination during gaze shifts and Einhäuser et al. [15] discovered human eye-head coordination in natural exploration. Nakashima et al. [31] proposed a model to improve the accuracy of saliency prediction by utilizing head direction.

The characteristics of human gaze behaviors in virtual reality have also been explored. Sitzmann et al. [39] revealed that there exists a latitudinal equator bias during users' exploration of 360° images. Xu et al. [45] explored human's gaze behaviors in dynamic 360° videos and found that users' gaze positions coincide with salient regions and moving objects. Hu et al. [21] revealed that, in static virtual scenes, users' gaze positions are correlated with their head rotation velocities and there exists a latency between eye movements and head movements. Based on the above-mentioned works, to explore huamn gaze behaviors in dynamic virtual scenes, we record the image sequences viewed by the observers, users' gaze positions, users' head rotation velocities, and the positions of dynamic objects for analysis.

2.3 Application of Eye Tracking

In VR systems, eye tracking technology has gained importance and it can be applied to many aspects. Tanriverdi et al. [43] utilized eye tracking technology as an interaction tool in virtual environments and