SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction

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

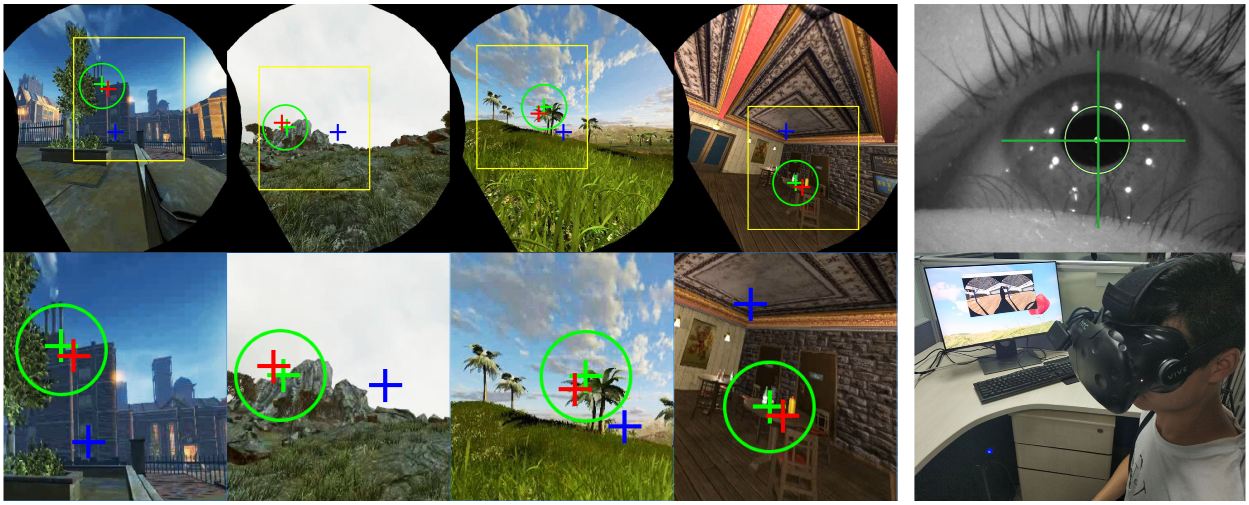

Fig. 1: Realtime gaze prediction performed using our eye-head coordination model. The quadruplet on the left demonstrates the gaze prediction results tested using different scenarios. The upper row shows the images captured from an HMD's screen, with zoomed-in view in the lower row. The ground truth of eye gaze in each scenario is marked using a green cross, the blue cross denotes the mean baseline, and the red one denotes our result. The green circle shows the foveal region with a 15 ° field of view. The figure of a user's eye gaze on the top-right illustrates that our goal is to predict realtime gaze position, and the bottom-right illustrates our experimental setup. From these results, our model has high accuracy when compared with the ground truth from the eye tracker.

Abstract— We present a novel, data-driven eye-head coordination model that can be used for realtime gaze prediction for immersive HMD-based applications without any external hardware or eye tracker. Our model (SGaze) is computed by generating a large dataset that corresponds to different users navigating in virtual worlds with different lighting conditions. We perform statistical analysis on the recorded data and observe a linear correlation between gaze positions and head rotation angular velocities. We also find that there exists a latency between eye movements and head movements. SGaze can work as a software-based realtime gaze predictor and we formulate a time related function between head movement and eye movement and use that for realtime gaze position prediction. We demonstrate the benefits of SGaze for gaze-contingent rendering and evaluate the results with a user study.

Index Terms—Eye-head coordination, gaze prediction, Pearsons correlation coefficient, eye tracking, saliency

1 Introduction

Virtual reality (VR) systems are used to explore 3D virtual worlds using multimodal interfaces. During navigation or exploration, a user may gaze at different objects of interest or in different directions. There is considerable work on the use of gaze information for eye movement-based interaction [35] and gaze-contingent rendering (or foveated rendering) [17, 28, 33, 34]. Gaze-contingent rendering is a technique that is used to improve the frame rate and is based on decreasing the rendering quality in the peripheral region while maintaining high fidelity in the foveal region. A hardware-based solution for computing the gaze position is based on using eye trackers, which are designed for realtime gaze tracking. There is considerable work on the development of eye-trackers and their integration with commercial head mounted devices (HMDs). However, as accessory equipment attached to an HMD, eye trackers can be relatively expensive and lack ease of use. In this paper, we investigate a complimentary software-driven approach for realtime gaze prediction.

Gaze prediction is different from prior works in human eye fixations or visual saliency, which predict a density map of eye fixations [10]. In contrast, realtime gaze prediction aims to predict a single gaze and the realtime requirements (60Hz or better) make it distinct from approaches that focus on saliency prediction. Realtime gaze prediction is a very challenging task due to multiple factors. First, the attention of the human visual system (HVS) is affected not only by the content of the scene but also by specific tasks performed in a virtual environment. The HVS has a two-component framework of attention: a bottom-up mechanism and a top-down mechanism [20]. The bottom-up mechanism is fast and it biases the observer's attention towards the scene's salient object. On the other hand, the top-down mechanism is slower and it directs attention under cognitive control [20]. It is difficult to predict gaze positions when both mechanisms are in effect. Second, a human's gaze can be considered a random behavior [2] independent of observation conditions. Different individuals may gaze at different positions even when they are provided with the same content, and an individual user's gaze behavior is not identical even if he or she performs a task repeatedly within the same scene. Furthermore, the existence of a saccade, a quick, simultaneous movement of both eyes

- *Sheng Li is the corresponding author.*
- *Zhiming Hu, Congyi Zhang, Sheng Li and Guoping Wang are at Peking University, China. E-mail: {jimmyhu | cyzh | lisheng | wgp}@pku.edu.cn.*
- *Dinesh Manocha is at University of Maryland, USA. E-mail: dm@cs.umd.edu.*

between two or more phases of fixation in the same direction, makes gaze prediction difficult. It is challenging to capture users' realtime gaze positions accurately when a saccade occurs. Finally, the realtime requirement imposes strict restrictions on the efficiency of gaze prediction methods. Many prior works on saliency prediction [4] mainly focus on the accuracy of the algorithms and do not provide guarantees in terms of realtime performance for VR applications.

There is considerable work on gaze-based interaction in different fields. At a broad level, prior interaction methods can be classified into four types: diagnostic (off-line measurement), active (selection), passive (foveated rendering), and expressive (gaze synthesis) [12]. In the context of VR applications, there is interest in active, passive, and expressive methods. The active techniques provide the gaze data as a streaming signal corresponding to the eye movement and are used for selection or similar commands as part of the interface. On the other hand, passive techniques are used to incorporate a change in the display like foveated rendering, but do not imply a specific user action. Expressive techniques are used for observations and are useful in terms of noticing the movement of the eyes of an avatar or a virtual agent.

Main Results: We present and evaluate a software-based solution for predicting realtime gaze position in an immersive VR system without any additional specialized hardware called *SGaze*. Our approach is based on a novel, data-driven method that models coordinated movements between eyes and the head called *eye-head coordination*. Moreover, we use this model (SGaze) for realtime prediction of gaze positions using the head movements and other factors. Our technique can possibly be used as a substitute for an eye tracker for exploring virtual worlds using HMDs, especially for passive gaze-based interactions. Our data-driven model is formulated based on the following procedures:

Build a dataset: We record data that can reflect the realtime viewing statuses of the observers. To this end, we conduct a study on 60 participants who are asked to freely navigate in 7 virtual scenes including both indoor and outdoor field scenes with different lighting conditions (See Fig. 2). Next, we build a dataset that records the gaze positions and the corresponding head poses of the participants. The realtime scenes viewed by the observers are also recorded. In this paper, we mainly focus on the free exploration of the virtual scene, i.e. no specific tasks or instructions are given to the participants when they are navigating using an HMD.

Model eye-head coordination: We perform Pearsons correlation coefficient (PCC) analysis on the dataset and observe that there exists a range within which gaze positions have a strong linear correlation with head rotation angular velocities. We also find that there exists a latency between eye movements and head movements. Moreover, we analyze the existence of saccades in our gaze dataset and discover that long saccades seldom exist when users are immersed in virtual environments. Based on the analysis, we present our eye-head coordination model as a time related function between head movement and eye movements.

Predict gaze and application: Based on our eye-head coordination model, we propose a novel realtime gaze prediction method that combines realtime head poses obtained from a head tracker with other factors, including saliency maps of the scenes. We also propose some baselines and evaluation metrics for realtime gaze position prediction. We evaluate the performance of our method and the results show that our method performs significantly better than the baselines. Moreover, we use our model for gaze-contingent rendering and conduct a user study to validate the effectiveness of our approach. Our preliminary results from the user study indicate that our approach can be considered as an alternative to eye trackers in some immersive VR applications. We also evaluate our model's performance on simple tasks and the result shows that our model still outperforms these baselines.

Overall, our contributions include:

- We gather data and evaluate the stereoscopic vision interaction with an HMD in immersive virtual environments. We use this data to build an eye-head coordination model (SGaze) by analyzing the key components of users' gaze behaviors.
- We present a novel realtime gaze prediction method based on our eye-head coordination modeling.
- We present some baselines and evaluation metrics for realtime gaze prediction and validate our method by gaze-contingent rendering. Our data and metrics can be used for other analysis and applications.

2 Related Work

In this section, we give a brief overview of prior work on gaze prediction and eye-head coordination.

2.1 Gaze Prediction

Visual saliency prediction or gaze prediction is a well-studied field in computer vision and related areas. Many saliency models have been proposed in the last three decades. Inspired by the neuronal architecture of the early primate visual system, Itti et al. [21] proposed a classical visual attention model that computes saliency maps using multiscale image features. Realizing that scene context plays an important role in visual saliency, Oliva et al. [27] proposed a saliency model that takes contextual information into consideration. In general, most of the existing saliency prediction models are based on bottom-up approaches that only utilize low-level image features such as intensity, color, and orientation [6, 21]; top-down approaches, which consider high-level knowledge of the scene like specific tasks and context [5, 19]; or hybrid models. With recent advances in deep learning, many methods based on convolutional neural networks (CNNs) have been proposed and have achieved good performances on the saliency benchmarks [10, 22]. Most of the previous visual saliency models were designed for a single image. Apart from these models, researchers also investigated the saliency of stereo images [9, 18] and videos [11, 36]. However, in the field of virtual reality, there is limited work on visual saliency and gaze behavior. Sitzmann et al. [31] and Rai et al. [30] both explored the VR saliency in 360° static images. Xu et al. proposed a model for gaze prediction in dynamic 360° immersive videos [38]. These visual saliency prediction or gaze prediction methods generally predict a density map of eye fixations rather than predicting a single realtime gaze position. However, for some VR applications like gaze-contingent rendering and eye movement interaction, a user's realtime gaze position has more practical significance than a density map of eye fixations. In this paper, we present a realtime method for gaze position prediction using our data-driven head-eye coordination model.

2.2 Eye-Head Coordination

Eye-head coordination refers to coordinated movements between eyes and the head and has been investigated in the fields of cognitive science and neuroscience. In his seminal work, Yarbus [39] found that the eyes and the head move in coordination during gaze shifts and that there exists a connection between eye-head coordination and visual cognition. Nakashima and Shioiri [26] further revealed that there is interference with visual processing when head and eye directions are different because it takes time to modulate attention in this viewing condition. This study indicates that eye-head coordination is important for visual cognition and that humans achieve better cognitive performance when eye and head directions are coordinated during gaze shifts. Einhäuser et al. discovered the eye-head coordination of humans during free navigation of natural environments [13]. Other works [1, 40] revealed that there exists a latency in eye-head coordination and that eye movements usually happen before head movements.

Many studies on eye-head coordination concentrated on the relationship between the amplitude of head movements and the amplitude of gaze shifts, revealing that a head movement's amplitude is closely related to the gaze shift amplitude [14, 15]. Stahl [32] reported that when gaze shift amplitude is low, there exists an eye-only range in which gaze shifts are unaccompanied by head movements. Conversely, when gaze shift amplitude is high, there exists an eye-head range in which head movement amplitude has a linear relationship with gaze shift amplitude. Nakashima et al. [25] managed to improve the accuracy of saliency prediction by utilizing head direction. Our approach is motivated by these prior works and is based on collecting a large dataset to compute a correlation between realtime gaze position and realtime head movement.

There is considerable work on head tracking and current VR systems can measure head movement with high precision in realtime. Recent works by Sitzmann et al. [31] and Rai et al. [30] predicted saliency maps using head orientation for 360° images in VR. However, only head orientation is used for 360° panorama applications. In contrast, our approach targets finding a correlation between head movements and on-screen gaze positions.

2.3 Eye Tracking and Gaze Positions

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. An eye tracker is a device for measuring eye movements that can be used in research on the visual system, psychology, and psycho-linguistics, as an input device for human-computer interaction, and also in product design. Eye tracking technology has recently received more attention in VR systems. Gaze-contingent rendering [17, 28] requires the realtime gaze positions of the users to determine the central foveal region for rendering applications. Eye movement-based interactions in virtual reality also demand accurate realtime gaze positions to provide users with a good experience [24].

3 Gaze Data Collection

A key part of our data-driven model is gathering the data of a large number of participants freely navigating in virtual worlds. We use this data for analysis and for formulating our eye-head coordination model. In this section, we present the details of our data collection process.

Although there are already a few datasets that record users' gaze data in virtual reality, the stimuli of these datasets are either 360° images [30, 31] or 360° videos [38]. These datasets consist of head rotation information, and it is not clear whether these datasets can be used directly for users to freely explore and interact with the 3D virtual scenes. Furthermore, those datasets lack auxiliary information like 3D scenes viewed by the observers, information that is useful for realtime gaze analysis and prediction. Therefore, we have built a new dataset for realtime gaze prediction that is recorded from a VR HMD and eye tracker hardware. Our dataset also includes realtime 3D scenes viewed by the observers, their gaze data, and their head poses.

3.1 Stimuli

To collect users' realtime gaze data in virtual environments, participants were asked to freely explore 3D virtual scenes. Fig. 2 shows 7 virtual scenes used in our experiments with different lighting conditions. These test scenes include city, desert, forest, etc., which are common in VR applications. All scenes are static without any moving objects in them. To collect users' gaze data in different lighting conditions, we use seven different scenes and view them with bright lighting (Fig. 2, top) and dim lighting (Fig. 2, bottom). Compared with 360° images or 360° videos, our test scenes are represented using 3D models and participants can teleport themselves to any positions they want using the interactive controllers.

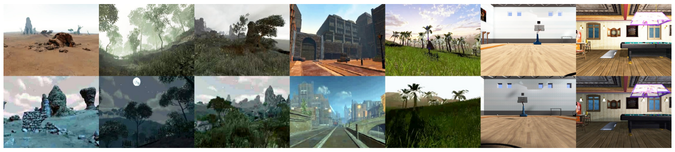

Fig. 2: Seven 3D virtual scenes used for data collection, including both indoor and outdoor scenes with different lighting conditions. Top: Bright lighting; Bottom: Dim lighting.

3.2 Participants

60 participants (35 male, 25 female, ages $18-36$) involved in our experiments. Each participant reported normal or corrected-to-normal vision and the eye tracker was calibrated for each user before he/she took part in our experiments.

Data item	Raw data	After processing
Gaze	On-screen point	Visual angle
Head pose	Rotation matrix	Rotation velocity & acceleration
Time	Time stamp	Calibrated time stamp

Table 1: Data structure and conversion performed on the raw gaze and head pose data along with the time.

3.3 System Details

In all our experiments, we use an HTC Vive head-mounted suit (including Vive controller for interaction) equipped with a 7invensun aGlass DK II embeddable eye tracker with an accuracy of 0.5° at a sampling frequency of 100 Hz. We record the head pose using HTC Vive's inertial measurement unit (IMU) at a sampling rate of 200 Hz with an accuracy up to 1.5 cm and jitter < 0.3 mm. We use the Unity game engine to display all the scenes and record the realtime scenes for the observers using a Bandicam screen-recorder at 60 fps. The CPU and GPU of our platform are an Intel(R) Xeon(R) E3-1230 v5 @ 3.40 GHz and an NVIDIA GeForce GTX 1060 6GB, respectively. The snapshot of the experiment setup is illustrated in the bottom-right of Fig. 1.

3.4 Procedure

At the beginning of the experiments, participants are given at least 3 minutes to get used to our experimental system. We set four start positions for each scene beforehand to help users fully explore our scenes. Participants can use the HTC Vive controllers to switch from the preset start positions to any position they like. During the experiments, no task is specified and the participants can freely navigate and observe the scenes without any instruction. Each observer performs 4 tests (2 different scenes randomly chosen from our 7 scenes with both bright and dim lighting) and each test lasts for at least 3 minutes. Each participant is provided with a pair of earplugs to avoid auditory disturbance. We record the realtime scenes viewed by the observers, the gaze data and the head pose information. Specifically, our dataset includes 240 pieces of data (one piece of data contains the records from one participant's test) and each piece of data includes 18,000 gaze positions (100Hz sampling rate), 36,000 head pose records (200Hz sampling rate) and 10,800 frames of scene screenshots (60Hz sampling rate). The time stamps of these data are also recorded and can help align them with each other.

For later analysis, we make some data conversions to the raw gaze and head pose data, as shown in Table 1.The raw gaze data are users' on-screen gaze positions ranging from (0,0) for top-left to (1,1) for bottom-right. We use a Cartesian coordinate for gaze by setting the origin to the screen center and orienting the X-axis from left to right and the Y-axis from bottom to top in the plane of the screen. We convert the raw gaze position to a visual angle (angle between the vector of the sight line and the normal of the screen) in the new coordinate system. The raw head pose data are the rotation matrices from the head position to the absolute tracking system, and they are converted into the rotation velocity and acceleration along the screen plane's X and Y axes.

4 SGaze: Eye-head Coordination Model

Recent works by Sitzmann et al. [31] and Rai et al. [30] introduced head pose data into saliency predictions for 360° images. Because their models are specialized for 360° images with only the head orientation information, it is not clear whether these datasets can be used to find a correlation between head movements and on-screen gaze positions. Since the head pose data can be easily accessed from the HMD's IMU, we use that information to formulate our eye-head coordination model.

An intuitive observation of the eye-head coordination in virtual reality is that the amplitudes of users' on-screen gaze positions have strong relationships with their heads' rotations, namely, the velocity and acceleration. In addition, the gaze behavior in VR is a complicated pattern that is affected by multiple factors including content, task, latency, etc. In this section, we present an eye-head coordination model to explain users' realtime gaze behaviors when they explore virtual environments in a time sequence:

$$
\begin{aligned}
x_g(t) &= \alpha_x \cdot v_{hx}(t+\Delta t_{x1}) + \beta_x \cdot a_{hx}(t) + F_x(t+\Delta t_{x2}) + G_x(t) + H_x(t), \\
y_g(t) &= \alpha_y \cdot v_{hy}(t+\Delta t_{y1}) + F_y(t+\Delta t_{y2}) + G_y(t) + H_y(t),
\end{aligned}
\tag{1}
$$

where x_g and y_g denote the coordinates of the gaze position; v_{hx} and v_{hy} are the head rotation velocities in horizontal and vertical directions, respectively; Δt_{x1} and Δt_{y1} are the latencies between eye movements and head movements in horizontal (X) and vertical (Y) directions, respectively; a_{hx} is the head rotation horizontal acceleration; F_x and F_y represent the influences from the exact VR content appearing in the HMD; Δt_{x2} and Δt_{y2} are the response latencies; G_x and G_y refer to the influences of specific tasks; H_x and H_y refer to the influences of all other factors, including the user's personal characteristics, behavioral habits, mental state, etc.; and α_x,α_y and β_x are influence coefficients that can be further solved by data fitting. We will discuss each entry based on the analysis of key factors from the biological movements and statistics in the rest of this section.

4.1 Head Movement

Eye-head coordination in virtual reality may be quite different at different head velocities. Recent work by Sitzmann et al. [31] has revealed that human gaze behaviors are different at low and high head velocities. The authors divided their data into two regions by setting a threshold velocity. However, acute head movements with extremely high velocities may be caused by sudden accidents, which may influence the eye-head coordination [16]. Therefore, to quantitatively analyze the eye-head coordination, we classify the domain of horizontal and vertical head rotation velocity into three separate regions for the eye-head coordination model: *Static*, *Intentional Move*, and *Sudden Move* (Fig. 3). The thresholds between those regions are denoted as v_{xh}, v_{xmin}, v_{xmax}, v_{yh}, v_{ymin}, v_{ymax}.

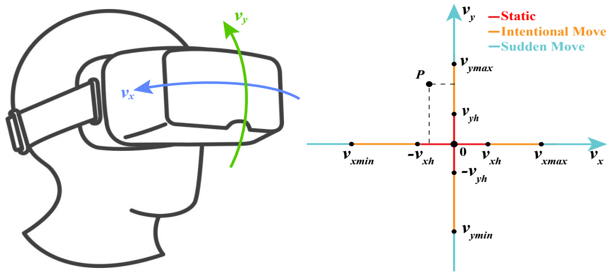

Fig. 3: Left: Head rotation angular velocity coordinate system on the HMD. Right: Segmentation of head rotation velocity into several regions for eye-head coordination. Horizontal velocity is segmented into 3 separate sub-regions: *Static* region (red), *Intentional Move* region (orange), and *Sudden Move* region (baby blue); vertical velocity is segmented as well, e.g., a gaze position P's horizontal velocity locates in *Static* region and its vertical velocity lies in *Intentional Move* region.

Static region refers to the condition when the users move their heads slightly or not at all. In this region, users' gaze positions are mainly influenced by factors like the content of the scenes, users' mental statuses, etc., and are invulnerable for minor head movements. *Intentional Move* region stands for the circumstance when head movements are intentionally driven by the users themselves and we find that within this region, users' on-screen gaze positions have strong linear correlations with the rotation velocities of their heads. *Sudden Move* region is the region within which head movements are caused by sudden accidents and the correlation between gaze direction and head rotation velocity is small. Moreover, we will show that there exists a latency between eye movements and head movements, meaning that eye movements happen before head movements. Furthermore, a saccade, which refers to a fast eye movement, may violate our eye-head linear correlation due to the vestibulo-ocular reflex [23]. Since the scenes tested are static and the users are given no specific tasks, we confirm that saccades seldom occur in our dataset and they may happen more frequently when the head rotation velocity is relatively high.

4.2 Eye-Head Linear Correlation

Static region refers to the low velocity region where users' attentions are more likely to be influenced by the content of the scene and are weakly correlated with head movements [31]. To determine the threshold velocities v_{xh} and v_{yh} for horizontal and vertical *Static* regions, we initialize v_{xh} and v_{yh} to 0 and further increase them at a growth rate of $0.1°/s$. Then we calculate the Pearson's correlation coefficient (PCC) between the gaze positions and head rotation angular velocities and find that PCC increases with the magnitude of threshold velocity in the low velocity regions. PCC is an indicator that ranges from -1 (perfect negative linear correlation) to 1 (perfect positive linear correlation). We set a constraint that the PCC in *Static* region should be lower than 0.15 [8], which means that the eye-head linear correlation is very weak. Within this constraint, we get $v_{xh} = 0.5°/s$ and $v_{yh} = 0.2°/s$.

Sudden head movements seldom occur during locomotion [16, 29], which means that *Sudden Move* region should contain very little data and *Intentional Move* region should take a larger portion of the records over *Static* region. Therefore, to locate the boundary between *Intentional Move* and *Sudden Move* regions (Fig. 3), we compute the PCC along the horizontal direction (the same for the vertical case) in a series of intervals $\{[v_{x1}, v_{x2}] | v_{x1} < -v_{xh}, v_{x2} > v_{xh}\}$ under the constraint that each valid interval has to contain over 95% of all records outside *Static* region. Practically, we enumerate the lower and upper bounds of the intervals at every $0.1°/s$ and filter out the invalid intervals that cover less than 95% of the records, then find the optimal interval with the highest PCC. We finally confine the interval of *Intentional Move* region as $[-88.5°/s, -0.5°/s] \cup [0.5°/s, 83.8°/s]$ horizontally and $[-35.6°/s, -0.2°/s] \cup [0.2°/s, 36.0°/s]$ vertically.

Table 2 shows the distribution of data in different regions and we can see that most of the data lies in *Intentional Move* region. The PCCs between gaze positions and head rotation velocities in different regions are shown in Table 3, demonstrating that the linear correlation between gaze position and head rotation velocity in *Intentional Move* region is significantly stronger than in *Static* and *Sudden Move* regions, coinciding with our observations. Furthermore, the linear correlation in *Intentional Move* region is higher than the whole region, which validates the effectiveness of analyzing head velocity in divided regions rather than in the whole region.

	Static	Intentional	Sudden
Horizontal	5.55%	91.45%	3.00%
Vertical	4.54%	90.69%	4.77%

Table 2: Distribution of data in different regions. Most of the data lies in *Intentional Move* region and only a small portion of the data lies in *Static* and *Sudden Move* regions.

	Static	Intentional	Sudden	Whole
PCC(v_x)	0.1345	0.5883	0.1511	0.5641
PCC(v_y)	0.1484	0.4969	-0.0906	0.4132

Table 3: The PCCs between gaze position and head rotation velocity in different regions. The PCC in *Intentional Move* region is significantly larger than it is in *Static* and *Sudden Move* regions and is better than the whole region. The correlations in *Static* and *Sudden Move* regions are rather small.

We also calculate the PCC between gaze position and head rotation acceleration and get 0.1134 in the horizontal direction and 0.0132 in the vertical direction. This result shows that there exists a weak linear correlation in the horizontal direction and thus we include horizontal head rotation acceleration in our model (Equation 1).

4.3 Eye-Head Latency

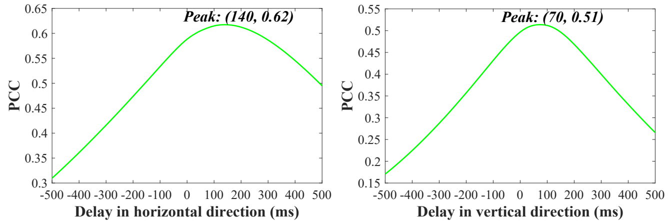

Fig. 4: Pearson's correlation coefficients between gaze positions and delayed head rotation angular velocities in *Intentional Move* regions of the horizontal (left) and vertical (right) directions, respectively. Each graph with a peak value of PCC demonstrates the exact latency between the gaze movement and the head rotation.

Many prior works [1, 37, 40] have revealed that there exists a latency in eye-head coordination. Eye movements usually happen before head movements [37, 40] and the latency varies with different head velocity regions [1]. This suggests that the eye-head coordination model may be more accurate when taking the eye-head latency into consideration. To validate this, we insert a continuously changing time delay between gaze positions and head rotation velocities in the horizontal and vertical *Intentional Move* regions and calculate a series of PCCs between them. From the graphs in Fig. 4, we can see that the eye-head coordination performs better when the eye movements are properly delayed, and this observation agrees with previous findings that eye movements usually precede head movements [37, 40]. The latency at which the PCC reaches its peak (Fig. 4) lies in the range proposed by Biguer et al. [1], which is about $20-200ms$. It varies in horizontal and vertical directions (Fig. 4), which aligns with previous findings that eye-head coordination differs in horizontal and vertical directions [14]. In fact, there is no constant latency between eye movements and head movements because it may change with gaze direction [1]. For simplicity, we consider the delay time at the PCC's peak point (Fig. 4) as the underlying latency and introduce delay constants into our model (Equation 1) by setting $\Delta t_{x1} = 140ms$ and $\Delta t_{y1} = 70ms$.

4.4 Saccade Analysis

There generally exists a special gaze pattern called a saccade, which refers to a quick, simultaneous movement of both eyes in the same direction between two or more phases of fixation. When saccades occur, gaze positions change rapidly and are quite difficult to predict. To analyze saccades in our dataset, we set a threshold v_s for the gaze speed. A saccade starts when the gaze speed exceeds v_s and it ends when gaze speed falls below v_s. We set $v_s = 75°/s$ as [14] and extract the vertical and horizontal saccades. We gather statistics of the amplitude of each saccade, which is expressed using the angular distance between the saccade's start point and end point (Fig. 5, top). We can see that the amplitudes of most of the saccades are not larger than 5°, which means these saccades will have little impact on applications like gaze-contingent rendering where the central foveal radius is set to 15° [28]. Furthermore, the durations of most of the saccades are very short (Fig. 5, bottom) and the total durations of horizontal and vertical saccades take up only 1.44% and 1.06% of the total gaze duration, respectively. We also analyze the frequency of saccades in different head velocity regions (Table 4) and observe that saccades are more likely to occur in *Sudden Move* region, which is only a small portion of the whole dataset (Table 2). The frequency of saccades is very low across all regions (Table 4), ensuring that gaze prediction using our dataset is practicable to some extent.

5 Prediction Method

In this section, we use our eye-head coordination model (Equation 1) for realtime gaze prediction. Since eye-head coordination is distinct in different head rotation velocity regions, we treat each region in a different manner to predict gaze position $\tilde{x}_g$ in the horizontal direction and $\tilde{y}_g$ in the vertical direction.

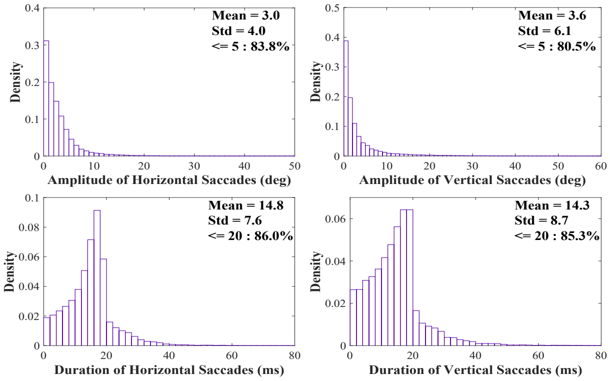

Fig. 5: Saccades in our dataset. Top: Amplitudes of horizontal and vertical saccades. Bottom: Durations of horizontal and vertical saccades. Most of the saccades are very short with amplitudes <= 5° and durations <=20 ms. This demonstrates that saccades in our dataset have little impact on the eye-head linear correlation.

	Static	Intentional	Sudden
Horizontal	0.68%	1.40%	3.85%
Vertical	1.05%	1.00%	2.04%

Table 4: The frequency of saccades in different head velocity regions. This demonstrates that saccades occur more frequently in *Sudden Move* region than in *Static* and *Intentional Move* regions.

When the velocity lies in *Static* region, we predict gaze positions in the corresponding direction using

$$\tilde{x}_g = E_x, \quad \tilde{y}_g = E_y, \tag{2}$$

where E_x and E_y are the expectations of horizontal and vertical gaze positions, respectively, and the corresponding head velocities of which fall into this region.

When the velocity is in *Intentional Move* region, we present a method that can adaptively determine the effect of head rotation velocity and predict the head velocity in the near future using

$$\begin{aligned} \tilde{x}_g =& \alpha_x \cdot \tilde{v}_{hx}(t+\Delta t_x) + \beta_x \cdot a_{hx} + b_x \cdot x_S + c_x, \\ \tilde{y}_g =& \alpha_y \cdot \tilde{v}_{hy}(t+\Delta t_y) + b_y \cdot y_S + c_y, \end{aligned} \tag{3}$$

where $\tilde{v}_{hx}$ and $\tilde{v}_{hy}$ are the predicted head rotation angular velocities, x_S and y_S are the salient positions, a_{hx} is the head rotation angular acceleration in the X axis, and c_x and c_y are the combinations of other factors. All the coefficients (α_x,α_y,β_x,b_x,b_y,c_x,c_y) in Equation 3 could be obtained by data fitting, which will be discussed in detail later.

When the velocity is in *Sudden Move* region, the gaze positions can be predicted by the mean values of former predictions using

$$\tilde{x}_g = \mu(\tilde{X}_g), \quad \tilde{y}_g = \mu(\tilde{Y}_g), \tag{4}$$

where $\mu(\tilde{X}_g)$ and $\mu(\tilde{Y}_g)$ are the average gaze positions predicted during the past 600 ms.

5.1 Static Region

When the velocity falls into *Static* region, the eye-head linear correlation is not obvious, according to Table 3. We propose utilizing the expectation of gaze positions in this region as our prediction (Equation 2). We calculate the statistical mean of the gaze positions in this region and obtain $-0.05°$ in the X axis and $-1.83°$ in the Y axis. We herein set $E_x = -0.05°$ and $E_y = -1.83°$ for practical use.

5.2 Intentional Move Region

Although the eye-head linear correlation is conspicuous in *Intentional Move* region, the magnitude of the influence will vary for different individuals in different scenes. To solve this problem, we propose a method to evaluate the impact adaptively. In addition, eye movement happens before head movement (Sect. 4.3) and an eye-head latency is introduced in our model (Equation 1). To predict realtime gaze positions, we first predict the head rotation velocity. Furthermore, the influences of content and horizontal head rotation acceleration are also considered.

To explore the impact of head rotation velocity on a realtime gaze position, we transform Equation 1 to a simplified form:

$$\begin{aligned} x_g &= \alpha_x \cdot v_{hx} + \gamma_x, \\ y_g &= \alpha_y \cdot v_{hy} + \gamma_y, \end{aligned} \tag{5}$$

where x_g and y_g are the gaze positions, α_x and α_y stand for the magnitude of the head velocity's influence on gaze position, and γ_x and γ_y are the combinations of other entries. We employ the simplified model (Equation 5) on every single piece of data (single user's data in a single scene) separately and evaluate α_x and α_y by linear fitting. As a result, we find that α_x and α_y fluctuate a lot among different pieces of data. Thus, a more reasonable way of employing our model is to determine α_x and α_y adaptively.

We observe that the greater the variation in head velocity, the less the gaze positions are influenced by head movement, which means the magnitude of the head velocity's influence may be negatively related to the standard deviation of the velocity. To validate this important underlying mechanism, we utilize PCC to analyze their correlation and get -0.6596 in the horizontal direction and -0.5878 in the vertical direction. The result corresponds with our observation. We herein adaptively compute α_x and α_y in Equation 3 using

$$\begin{aligned} \alpha_x &= k_x \cdot \sigma(v_{hx}) + \ell_x, \\ \alpha_y &= k_y \cdot \sigma(v_{hy}) + \ell_y, \end{aligned} \tag{6}$$

where α_x and α_y are the influences of head velocity and $\sigma(v_{hx})$ and $\sigma(v_{hy})$ are the standard deviations of the velocities. For realtime prediction, we utilize the standard deviation of velocity data in the past 10 seconds. All the coefficients (k_x, k_y, ℓ_x, ℓ_y) in Equation 6 are obtained by data fitting, which will be discussed below.

Analysis in Sect. 4.3 has revealed that head movement lags behind eye movement. Thus, our method for predicting gaze position should predict head velocity first. With this in mind, we assume that head motion accelerates in a short period of time and then we predict the next head velocities using this relationship:

$$\begin{aligned} \tilde{v}_{hx}(t+\Delta t_x) &= v_{hx}(t) + \mu(a_{hx}) \cdot \Delta t_x, \\ \tilde{v}_{hy}(t+\Delta t_y) &= v_{hy}(t) + \mu(a_{hy}) \cdot \Delta t_y, \end{aligned} \tag{7}$$

where $\tilde{v}_{hx}$ and $\tilde{v}_{hy}$ are the predicted velocities, v_{hx} and v_{hy} are the real velocities that are recorded from IMU, Δt_x and Δt_y are the time intervals between the predicted velocities and the current velocities, and $\mu(a_{hx})$ and $\mu(a_{hy})$ are the average accelerations in the past 20 *ms*. Δt_x and Δt_y are obtained by data fitting.

We also take the impact from the content of an HMD image into consideration when predicting gaze positions. The saliency information of the scenes will facilitate the prediction. Since extracting a saliency map is usually time-consuming and cannot fulfill the requirement of realtime computation [3], we only calculate the saliency maps within the region where gaze positions frequently occur. We analyze the distribution of gaze positions and find that most (99.8%) of the data lies in the central region with a radius of 35°. Thus, we only utilize the saliency map of the central 35° region. Since the scenes are static and there is no sudden change in the content, we extract saliency maps only on sparse key frames to improve the computational efficiency. More specifically, we extract the realtime scenes every 250*ms* as key frames and calculate the saliency maps inside the central region using the state-of-the-art SAM-ResNet saliency predictor [10].

To explore the correlation between saliency maps and realtime gaze positions, we divide the saliency maps into 28×28 subregions (the side length of each subregion is 2.5°) and calculate the average saliency value for each subregion (Fig. 6). The center of the subregion with the maximal average saliency value is regarded as the salient position for the next 250*ms*. We evaluate the PCC between salient positions and gaze positions and find that there exists a correlation between them: 0.0745 in the horizontal direction and 0.0723 in the vertical direction. The salient positions are linearly independent (PCC$<$ 0.025) with head rotation velocities and head rotation accelerations and thus we introduce salient positions in Equation 3.

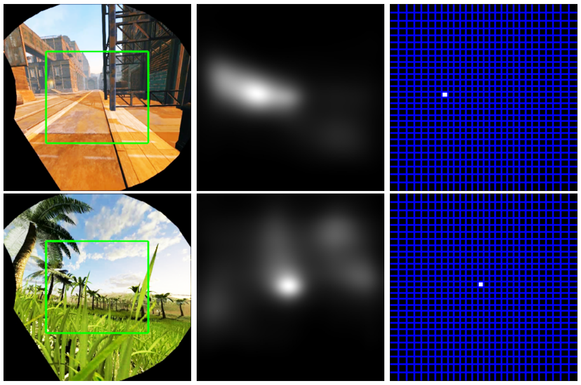

Fig. 6: We highlight the salient positions computed by our method. Left: Test scenes; Middle: Saliency maps of the central 35° region (i.e. the green rectangles in the scenes); Right: Subregions in a saliency map. The salient positions are the centers of the subregions with maximum average saliency values (i.e. the white grids in the subregions).

Since head rotation acceleration contributes a linear correlation with gaze position in the horizontal direction (Sect. 4.2), we also add them into our method (Equation 3).

The data gathered from 5 scenes out of the 7 scenes (around 77% of the total data) are used to train our model and the remaining 2 scenes (23% of the total data, including an indoor scene and an outdoor scene) are used for testing. We solve all the coefficients in Equation 3 using the least squares method and get these parameters: $k_x = -0.0015$, $\ell_x = 0.2491$, $\Delta t_x = 0.1480$, $\beta_x = 0.0006$, $b_x = 0.0344$, $c_x = 0.1777$, $k_y = -0.0053$, $\ell_y = 0.5293$, $\Delta t_y = 0.0304$, $b_y = 0.0503$, and $c_y = -2.5249$.

5.3 Sudden Move Region

When head velocity is in *Sudden Move* region, the acute head movements in this region are mainly caused by sudden accidents, which may break the eye-head coordination. Since saccades seldom occur in our dataset (Sect. 4.4), a reasonable prediction strategy is to keep the gaze position stable relative to its precedent positions. Whereupon, we compute the gaze positions according to previous predictions using Equation 4.

6 Results

Evaluation on the performance of our model is presented in this section. We first propose some baselines and evaluation metrics for the task of realtime gaze prediction. Then we demonstrate that our model performs best when compared with the baselines. We conduct an ablation study to validate the effectiveness of each component in our model and we prove the effectiveness of our region division strategy. We also apply our model to gaze-contingent rendering and verify the effectiveness of our model through a user study. We finally evaluate our model's performance in a task-oriented situation and find that our model still performs better than the baselines.

6.1 Baselines and Evaluation Metrics

6.1.1 Baselines

For the task of realtime gaze prediction, we introduce some baselines for evaluation. Recent work has indicated that observers tend to gaze at the center of an image, irrespective of its content [7]. The center bias has also been revealed in virtual reality applications [31]. Therefore, a simple but meaningful baseline is to use the screen center as the gaze position. In addition, we also utilize the statistical mean of all the gaze positions as another baseline, which is $(0.03^\circ, -2.34^\circ)$ from statistics. To evaluate the performance of utilizing only saliency information, salient position (Sect. 5.2) is also included as one of the baselines.

6.1.2 Evaluation Metrics

There are many existing evaluation metrics for saliency prediction, but evaluation metrics for the task of realtime gaze prediction have not been well studied. Therefore, we present a few evaluation metrics for realtime gaze prediction. Since our goal is to figure out the users' realtime gaze positions, our evaluation measure is set as the angular distance (visual angle) between the predicted gaze position and the ground truth gaze position. In addition, for many applications like gaze-contingent rendering, the gaze region plays a more important role than a single gaze position. In this case, we use both a precision rate, the proportion of the overlapped region at the predicted gaze region, and a recall rate, the proportion of the overlapped region at the ground truth gaze region, as our evaluation metrics.

6.2 Model Evaluation

6.2.1 Performance Evaluation

Angular Distance: To evaluate the performance of our model, we first apply it to the test data to compute the predicted gaze positions. We calculate the mean value and standard deviation of angular distances between the ground truth gaze positions and the predicted gaze positions and indicate them in Table 5, which shows that our model performs better than the baselines in terms of both mean value and standard deviation. We also calculate the cumulative distribution function (CDF) of the angular distances for performance evaluation (Fig. 7 left). The higher the CDF curve, the better the performance. The result shows that our model achieves the best performance in terms of CDF. We can see that the angular distances of our model are already very small compared with the field of view of the HMD, which is 110° for our device. Fig. 1 also highlights the prediction results using our model.

	Ours	Mean	Center	Saliency
Mean	8.52°	10.93°	11.16°	21.23°
Std	5.66°	6.43°	6.44°	12.10°

Table 5: Comparison of angular distances between our model and the baselines. Our model performs best in terms of both mean value and standard deviation.

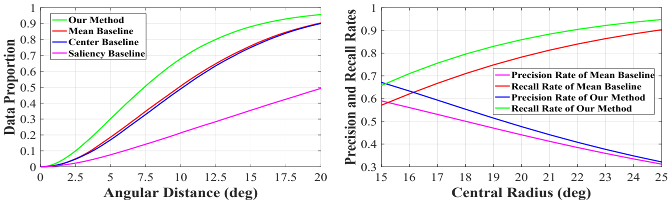

Fig. 7: Comparison on performance. Left: Cumulative distribution function of the angular distances. Right: Precision and recall rates of our model and the mean baseline at different central radii. Our model outperforms these baselines in terms of the CDF, precision rate, and recall rate. Our recall rate at the central radius of 20° reaches 85.92%, which can satisfy the requirements of many VR applications.

Precision and Recall Rates: To calculate the precision and recall rates, we should first determine the gaze region. We set the central radius to 15°, as used in gaze-contingent rendering [28], and treat the central region of the gaze position as the gaze region. For specific tasks like gaze-contingent rendering, the recall rate is generally the main concern and thus we enlarge the central radius of the predicted gaze region (central radius of the ground truth will remain unchanged) to obtain a better recall rate. The right of Fig. 7 shows that the precision and recall rates of our model are both better than the mean baseline (the best one according to Table 5) at different central radii. Therefore, we suggest an appropriate central radius with 20° that can make the recall rate of our model reach 85.92%, and the specification of such a gaze region with a high recall rate can be useful for many applications.

6.2.2 Ablation Study

To evaluate the effectiveness of each component in our model, we perform an ablation study by removing one component at a time and retraining the ablated models. Since our model mainly works in *Intentional Move* region, we test the ablated models only when the horizontal and vertical head velocities are both in *Intentional Move* region. Table 6 shows the performances of the ablated models. We observe that each component in our model indeed contributes to gaze prediction. Moreover, the head velocity component plays the most important role in our eye-head coordination model.

In the previous analysis (Sect. 4.2), we classify the head velocities into three regions and only train our model (Equation 3) in *Intentional Move* region. To validate the effectiveness of this classification, we further retrain our model for the entire region without classification to evaluate its performance. In terms of angular distance, the newly trained model obtains a mean value of 8.80° and a standard deviation of 6.07°, which infers that it cannot outperform our primordial model when compared with our results shown in Table 5. These results demonstrate the effectiveness of our strategy of using region classification.

The above results also manifest the sufficiency of PCC, which can be employed to determine the boundaries of the velocity regions and to single out such factors (head velocity, acceleration, etc.) that may facilitate the task of gaze prediction in our modeling stage. Other than PCC, there still exist some rank correlation coefficients (e.g., Spearman's rank correlation coefficient, Kendall rank correlation coefficient, etc.) that can also be used to measure the correlation between two variables. However, rank correlation coefficients are used to measure the monotonic correlation between two variables, while PCC can measure a specific correlation (linear correlation). Since our goal is to compute the correlation between gaze positions and other factors efficiently to formulate the eye-head coordination model, PCC is more suitable for completing such a task than rank correlation coefficients.

6.2.3 Runtime Performance

Our model utilizes head pose information and saliency maps from the scenes. The head pose is obtained through the IMU equipped on the HMD. Thus, the head pose information can be accessed in a very short time (less than 1 *ms*). Calculation of the saliency map can take considerable time, so we only compute the saliency maps of the central regions of the sampled scenes (Sect. 5.2). The average running time of our method for predicting a single gaze position is 4.5 *ms* and this result shows that our model can be employed in realtime applications.

6.3 User Study

6.3.1 Gaze-Contingent Rendering

To further demonstrate the usefulness of our model, we apply our model to gaze-contingent rendering (or foveated rendering) techniques [17, 28]. Gaze-contingent rendering decreases rendering quality in the peripheral region while maintaining high fidelity in the foveal region. We set the central radius of the foveal region to 20° and the width of the blending border to 60 pixels. Fig. 8 exhibits the result of our gaze-contingent rendering. Other than our model, we also utilize the gaze data from an eye tracker as the ground truth and use the statistical mean (the best baseline according to Table 5) as our baseline.

We conduct a user study to evaluate the availability of our model when applied to gaze-contingent rendering. In the experiment, each participant was asked to freely explore two randomly-assigned scenes with both bright and dim lighting (i.e. 4 tests in total), similar to the process of data collection. In each test, each participant should make comparisons twice (but in no fixed order): our model vs. ground

	Ours	w/o Acceleration	w/o Std	w/o Saliency	w/o Latency	w/o Velocity	Mean Baseline
Mean	8.42°	8.46°	8.46°	8.48°	8.49°	10.85°	10.96°
Std	5.63°	5.65°	5.65°	5.66°	5.65°	6.36°	6.42°

Table 6: Angular distances of the ablated models. Each component in our model helps improve our model's accuracy. In addition, the head velocity component plays the most important role in our model and this is why we call our model an eye-head coordination model.

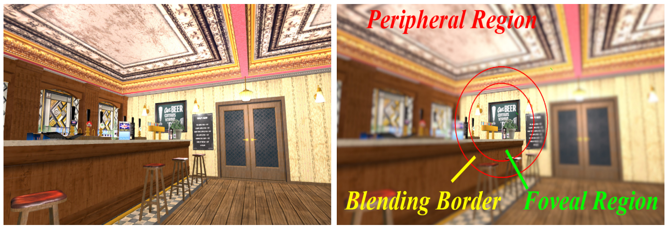

Fig. 8: Application of our gaze prediction to gaze-contingent rendering. Left: Rendered in normal mode; Right: Rendered using gaze-contingent rendering mode. Gaze-contingent rendering decreases rendering quality in the peripheral region while maintaining high fidelity in the foveal region. The inner circle is the foveal region rendered with high quality, the outer region is the peripheral region rendered with low quality and the blending border denotes the transitional region.

truth, and our model vs. mean baseline. To make the user study objective and fair, the competitors involved in a comparison present randomly, without any specific order. It generally takes around 2 minutes per comparison in a test, and the exploration using one side in the comparison will last for 1 minute before switching to the other. To collect users' responses, we run a two-alternative forced choice (2AFC) test that requires the participants to indicate which one they prefer with higher quality from these choices.

A total of 15 participants (12 male and 3 female, ages 18-30) who had never been involved in the data collection experiment participated in the next user study. We collected all the responses and found that the ground truth is preferred by 51.67% of the total responses and there is no significant difference between our model and the ground truth (t-test, $p > 0.75$). When comparing our model with the baseline, we find that 71.67% of the total responses preferred the results from our model, and this result is statistically significant (t-test, $p < 0.01$).

These results show that our model can be a good alternative to using eye trackers in immersive VR applications.

6.3.2 Performance on Simple Task

In task-oriented applications like games, users' gaze behaviors may be seriously influenced by the tasks assigned to them. Although our method is designed for free exploration of virtual environments in which users are given no specific tasks, it's worthy applying our model to such task-oriented scenarios and evaluating its performance.

We conducted the following experiments to analyze the users' gaze behaviors when a simple task was assigned to them. Specifically, users were asked to explore two scenes and complete the tasks assigned to them as shown in Fig. 9. In the gym scenario with many pieces of sports equipment (basketball, volleyball, tennis ball, etc.) randomly placed, the users were asked to find them and choose their favorite ones. In the city scenario with many trees along the street, the users were asked to count the number of trees during their exploration. Each exploration lasts for at least 2 minutes and 12 participants (9 male, 3 female, ages 18-25) participate in the experiment. We collect these data fro later analysis in the same way as described in Sect. 3.

We utilize the statistical mean of the newly collected gaze positions as the mean baseline, which is $(0.57°, -2.14°)$, and evaluate our already trained model on the newly collected data. We can see from Table 7 that our model still performs best when compared with the baselines and this result further proves the effectiveness of our model. However, due to the influence of the tasks, our model shows a smaller improvement relative to the best baseline when compared with the no-task situations (Table 5); the relative improvement of the angular distance decreases from 22.0% to 14.2%. This result inspires us to propose a new eye-head coordination model that takes specific tasks into consideration for task-oriented situations in our future work.

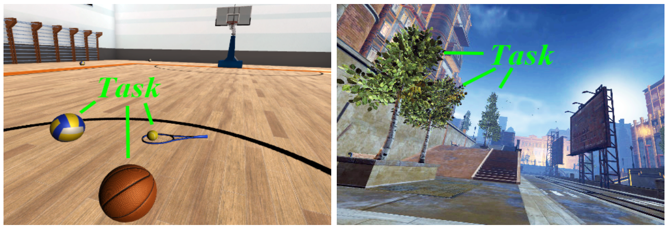

Fig. 9: Simple task-oriented scenarios. Left: Users were asked to look for the sports equipment (basketball, volleyball, tennis ball, etc.) and choose their favorite one in a gym. Right: Users were asked to count the number of trees during their exploration in a urban area.

	Ours	Mean	Center	Saliency
Mean	8.99°	10.48°	10.69°	18.49°
Std	5.76°	6.00°	6.03°	13.11°

Table 7: Comparisons of angular distances between our model and the baselines for the simple tasks. Our model still outperforms others in terms of both mean value and standard deviation.

7 Conclusion, Limitations, and Future Work

We present a novel, data-driven eye-head coordination model (SGaze) that can be used for realtime gaze prediction for immersive HMD-based applications. Our approach does not require any special hardware (e.g. an eye tracker) and is based on dataset collection and statistical techniques. In particular, we observe that there exists a range within which gaze positions have a strong linear correlation with head rotation angular velocities. We propose an eye-head coordination model using statistical analysis. We have evaluated our model for realtime gaze prediction and gaze-contingent rendering and the preliminary results are promising. Our model also outperforms these baselines when applied to such scenarios with simple tasks.

Our approach has some limitations. Our dataset is constructed through each user performing a free exploration of the virtual scenes in a passive way, as classified in prior work [12], without performing any specific tasks. Therefore, our data-driven model does not necessarily handle other types of VR applications well. The mechanism of gaze behavior in virtual reality is intricate and many aspects in the model (e.g., H_x and H_y shown in Equation 1) have not been explored thoroughly in our analysis. In terms of context, the saliency from realtime images can only give expression to 2D space. However, users generally interact directly with 3D objects in an immersive VR environment. It may help improve our model to explore the relationship between gaze behavior and a 3D representation of a scene (e.g., a depth map). Furthermore, the influences from sound and dynamic objects in a scenario have not been taken into consideration. In practice, sound plays an important role in an immersive environment and dynamic scenes can result in a different gaze behavior. Moreover, we can use other techniques such as deep learning or other statistical methods (e.g., rank correlation) to derive more accurate models and evaluate their performance for general VR applications. Since our model is built on a relative head rotation velocity coordinate system (Fig. 3 left), any other kinds of head pose tracking devices that can convert the head pose data into such a relative coordinate system with high accuracy have the potential to be applied to gaze prediction.

Our model also offers the potential to improve existing saliency models, although our model aims at predicting realtime gaze positions rather than post-processing saliency maps like Sitzmann et al's method [31]. After collecting an adequate number of users' realtime exploration data (head velocity, acceleration, etc.) for the same scene, our model can be employed to predict realtime gaze positions and these gaze positions will also form a saliency map after a long period of data gathering, which can benefit existing saliency models due to our model's high accuracy in realtime conditions. However, this approach of generating saliency maps using our model will be time-consuming and is not as elegant as Sitzmann et al's method [31]. Improving saliency models efficiently using our model is also an interesting avenue for future work.

Acknowledgments

We would like to thank anonymous reviewers for their valuable comments, Mr. Kangrui Yi for rendering on HTC Vive, and all the participants for the experiments. We would also thank 7invensun for their support on the eyetracking device. This project was supported by the National Key R&D Program of China (No.2017YFB0203002, No.2017YFB1002700) and National Natural Science Foundation of China (No.61632003, No.61661146002, No.61631001).

References

[1] B. Biguer, M. Jeannerod, and C. Prablanc. The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental brain research*, 46(2):301–304, 1982.

[2] G. Boccignone and M. Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004.

[3] A. Borji, M. M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, Dec 2015.

[4] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[5] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 470–477. IEEE, 2012.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[7] A. D. Clarke and B. W. Tatler. Deriving an appropriate baseline for describing fixation behaviour. *Vision research*, 102:41–51, 2014.

[8] J. Cohen. Statistical power analysis for the behavioral sciences. 2nd, 1988.

[9] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016.

[10] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 2018.

[11] X. Cui, Q. Liu, and D. Metaxas. Temporal spectral residual: fast motion saliency detection. In *Proceedings of the 17th ACM international conference on Multimedia*, pp. 617–620. ACM, 2009.

[12] A. T. Duchowski. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics*, 73:59–69, 2018.

[13] W. Einhäuser, F. Schumann, S. Bardins, K. Bartl, G. Böning, E. Schneider, and P. König. Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems*, 18(3):267–297, 2007.

[14] Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, and S. Shioiri. Eye-head coordination for visual cognitive processing. *PloS one*, 10(3):e0121035, 2015.

[15] E. G. Freedman. Coordination of the eyes and head during visual orienting. *Experimental brain research*, 190(4):369, 2008.

[16] G. E. Grossman, R. J. Leigh, L. Abel, D. J. Lanska, and S. Thurston. Frequency and velocity of rotational head perturbations during locomotion. *Experimental brain research*, 70(3):470–476, 1988.

[17] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3d graphics. *ACM Trans. Graph.*, 31(6):164:1–164:10, Nov. 2012.

[18] F. Guo, J. Shen, and X. Li. Learning to detect stereo saliency. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pp. 1–6. IEEE, 2014.

[19] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pp. 545–552, 2007.

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

[21] 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.

[22] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge. Understanding low-and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision*, pp. 4799–4808, 2017.

[23] V. Laurutis and D. Robinson. The vestibulo-ocular reflex during human saccadic eye movements. *The Journal of Physiology*, 373(1):209–233, 1986.

[24] P. Majaranta. *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies*. IGI Global, 2011.

[25] R. Nakashima, Y. Fang, Y. Hatori, A. Hiratani, K. Matsumiya, I. Kuriki, and S. Shioiri. Saliency-based gaze prediction based on head direction. *Vision research*, 117:59–66, 2015.

[26] R. Nakashima and S. Shioiri. Why do we move our head to look at an object in our peripheral region? lateral viewing interferes with attentive search. *PloS one*, 9(3):e92284, 2014.

[27] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *Image processing, 2003. icip 2003. proceedings. 2003 international conference on*, vol. 1, pp. I–253. IEEE, 2003.

[28] 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 Trans. Graph.*, 35(6):179:1–179:12, Nov. 2016.

[29] T. Pozzo, A. Berthoz, and L. Lefort. Head stabilization during various locomotor tasks in humans. *Experimental brain research*, 82(1):97–106, 1990.

[30] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 205–210. ACM, 2017.

[31] 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, April 2018.

[32] J. S. Stahl. Amplitude of human head movements associated with horizontal saccades. *Experimental brain research*, 126(1):41–54, 1999.

[33] N. T. Swafford, D. Cosker, and K. Mitchell. Latency aware foveated rendering in unreal engine 4. In *Proceedings of the 12th European Conference on Visual Media Production*, p. 17. ACM, 2015.

[34] N. T. Swafford, J. A. Iglesias-Guitian, C. Koniaris, B. Moon, D. Cosker, and K. Mitchell. User, metric, and computational evaluation of foveated rendering methods. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 7–14. ACM, 2016.

[35] V. Tanriverdi and R. J. K. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pp. 265–272. ACM, New York, NY, USA, 2000.

[36] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[37] D. Whittington, M.-C. Hepp-Reymond, and W. Flood. Eye and head movements to auditory targets. *Experimental Brain Research*, 41(3-4):358–363, 1981.

[38] 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.

[39] A. Yarbus. Eye movements and vision. 1967. *New York*, 1967.

[40] W. H. Zangemeister and L. Stark. Active head rotations and eye-head coordination. *Annals of the New York Academy of Sciences*, 374(1):540–559, 1981.