EHTask: Recognizing User Tasks from Eye and Head Movements in Immersive Virtual Reality

Zhiming Hu, Andreas Bulling, Sheng Li*, *Member, IEEE,* Guoping Wang

Abstract—Understanding human visual attention in immersive virtual reality (VR) is crucial for many important applications, including gaze prediction, gaze guidance, and gaze-contingent rendering. However, previous works on visual attention analysis typically only explored one specific VR task and paid less attention to the differences between different tasks. Moreover, existing task recognition methods typically focused on 2D viewing conditions and only explored the effectiveness of human eye movements. We first collect eye and head movements of 30 participants performing four tasks, i.e. *Free viewing*, *Visual search*, *Saliency*, and *Track*, in 15 360-degree VR videos. Using this dataset, we analyze the patterns of human eye and head movements and reveal significant differences across different tasks in terms of fixation duration, saccade amplitude, head rotation velocity, and eye-head coordination. We then propose *EHTask* – a novel learning-based method that employs eye and head movements to recognize user tasks in VR. We show that our method significantly outperforms the state-of-the-art methods derived from 2D viewing conditions both on our dataset (accuracy of 84.4% *vs.* 62.8%) and on a real-world dataset (61.9% *vs.* 44.1%). As such, our work provides meaningful insights into human visual attention under different VR tasks and guides future work on recognizing user tasks in VR.

Index Terms—Visual Attention, Task Recognition, Eye Movements, Head Movements, Deep Learning, Virtual Reality

1 Introduction

Understanding human visual attention in immersive virtual reality (VR) is crucial for many important applications, including gaze prediction [1–3], VR content design [4], gaze guidance [5], and gaze-contingent rendering [2, 3]. However, previous works on visual attention analysis typically only explored one specific VR task, e.g. *Free viewing* task [2–4, 6] or *Visual search* task [1]. Similarly, existing VR datasets typically only cover one VR task, making it hard to analyze and compare performance across tasks. Analyzing and comparing human behaviour across different tasks helps to better understand the mechanisms of human visual attention in VR [1, 7] and to develop related VR applications, such as gaze guidance [5] or gaze-contingent rendering [2, 3]. In addition, it helps to build better predictive models of visual attention [1–4] and to derive models recognizing user tasks from visual attention [8–11].

Human visual attention is strongly influenced by the specific task a user is performing [1, 7, 10]. Investigating the effect of task on visual attention in immersive virtual reality is of great significance for the emerging research area of task recognition in VR [7]. In his seminal work [12], Yarbus analyzed human gaze positions in seven different visual tasks and found that their eye movement patterns were significantly different. Inspired by Yarbus' work, many researchers focused on the *inverse Yarbus process*, i.e. recognizing user tasks from eye movement patterns [8, 9, 13–19]. Task recognition methods have many important applications in the areas of virtual reality, augmented reality (AR), and mixed reality (MR), collectively referred to as XR, including adaptive virtual environment design [7], low-friction predictive interfaces [16, 20], and attention-aware intelligent systems [21]. Specifically, virtual environments can provide a user with dynamic and adaptive experiences according to the specific task the user is performing [7]. XR systems have the potential to alleviate a user's burden of interaction by recognizing user tasks and interaction goals and providing convenience for completing the corresponding actions with less friction [16, 20]. By recognizing user tasks and attentional states, XR systems can adapt to different states of attention to improve the usability of the system [21]. However, prior works on task recognition typically focused on 2D images and videos [8, 9, 15, 22] and few works have studied immersive virtual reality. Moreover, existing task recognition methods mainly focus on human eye movements [8, 9, 13–16, 22] and paid less attention to human head movements. However, human head movements provide substantial insights into human cognitive behaviours [1–3, 23, 24] and may also have strong correlations with user tasks. Therefore, it is important to investigate the effectiveness of both human eye and head movements in recognizing tasks in immersive virtual reality.

We first perform a user study to collect 30 users' eye and head movements while performing four tasks, i.e. *Free viewing*, *Visual search*, *Saliency*, and *Track*, in 15 360-degree VR videos. Using this dataset, we analyze the characteristics of human eye and head movements, including fixation duration, fixation number, saccade amplitude, head rotation velocity, head rotation acceleration, and eye-head coordination, and observe significant differences across different tasks. Based on our analysis, we then propose *EHTask* – a novel learning-based method to recognize user tasks from eye and head movements in VR. We further conduct extensive experiments to evaluate our model. Results from these experiments show that our model outperforms the state-of-the-art methods derived from 2D viewing conditions by a large margin both on our dataset

- *Zhiming Hu, Sheng Li, Guoping Wang are with School of Computer Science, Peking University, China. E-mail: {jimmyhu | lisheng | wgp}@pku.edu.cn.*
- *Sheng Li and Guoping Wang are with National Biomedical Imaging Center, Peking University, China.*
- *Andreas Bulling is with the University of Stuttgart, Germany. E-mail: andreas.bulling@vis.uni-stuttgart.de.*
- *Sheng Li is the corresponding author.*
- *Project URL: https://cranehzm.github.io/EHTask*

Manuscript received ; revised

(accuracy of 84.4% *vs.* 62.8%) and on a real-world dataset (61.9% *vs.* 44.1%). Our dataset, source code, and pre-trained models are publicly available at https://cranehzm.github.io/EHTask.

The specific contributions of our work are three-fold:

- We provide a new dataset that contains human eye and head movements under four task conditions collected while participants viewed 15 360-degree VR videos.
- We analyze the patterns of human eye and head movements and reveal significant differences across different tasks in terms of fixation duration, saccade amplitude, head rotation velocity, and eye-head coordination.
- We present *EHTask*, a novel learning-based method to recognize user tasks in immersive virtual reality that significantly outperforms the current state of the art.

2 Related Work

Our work is related to previous works on 1) cognitive state estimation, 2) automatic recognition of user tasks, and 3) coordination of eye and head movements.

2.1 Cognitive State Estimation

In the area of cognitive research, cognitive state estimation has become a popular and important research topic in recent years. Lethaus et al. [25] and Sattar et al. [26] both focused on the problem of predicting user intents. Lethaus et al. predicted driver intent based on gaze data while Sattar et al. inferred user search intents from human gaze fixations. Pfleging et al. [27] and Fridman et al. [28] both concentrated on cognitive load estimation. Pfleging et al. presented an approach to estimating cognitive load by measuring pupil diameters under various controlled lighting conditions while Fridman et al. proposed two vision-based methods for cognitive load estimation under real-world driving conditions. Recently, Wang et al. utilized eye movements of a person recalling an image while looking at nothing to estimate mental images [29]. David et al. employed gaze features to predict artificial visual field losses by utilizing hidden Markov models and recurrent neural networks [30]. Ahn et al. decoded a reader's eye movements to estimate their levels of text comprehension and related states [31]. In addition, many researchers have studied the problem of recognizing user tasks and have presented many successful methods [8, 9, 13–15, 22].

In the field of virtual reality, some researchers focused on VR cybersickness prediction [32–34]. For example, Kim et al. developed an electroencephalography driven model to predict VR cybersickness [32] while Anwar et al. proposed a neural network-based method to predict the degree of cybersickness influenced by 360-degree VR videos [33]. Other researchers concentrated on cognitive load estimation in VR [35, 36]. Tremmel et al. utilized electroencephalogram features to estimate cognitive load in an interactive virtual environment [35]. Dell'Agnola et al. simulated cognitive loads in virtual reality and extracted features from different physiological signals to detect the levels of cognitive load [36]. In contrast with previous works, we focus on the problem of recognizing user tasks in immersive virtual reality.

2.2 Recognition of User Tasks

The problem of recognizing user tasks has been explored by many researchers. In his seminal work [12], Yarbus revealed that human eye movement patterns were significantly influenced by the specific tasks assigned to them, suggesting that a user's task may be recognized from his or her eye movements. Since then, many researchers have focused on the link between task and eye movements and have proposed many eye movement-based task recognition methods [8, 9, 13–16, 22]. Coutrot et al. employed hidden Markov models to recognize user tasks from fixations recorded while viewing static natural scene images [8]. Fuhl et al. proposed to use random ferns in combination with saccade angle successions to recognize user tasks [22]. They evaluated this approach on two image-based datasets and showed improvements over other methods. Hild et al. focused on the situation of viewing motion videos and utilized random forests to recognize user tasks from eye movement patterns [15]. However, prior work on task recognition typically focused on 2D viewing conditions, e.g. 2D images and videos, and few works have studied 3D viewing conditions (stereoscopic viewing conditions), e.g. immersive virtual reality. Moreover, existing task recognition methods mainly focus on human eye movements [8, 9, 13–16, 22] and pay less attention to human head movements. However, human head movements provide substantial insights into human cognitive behaviours [1–3, 23, 24] and may also have strong correlations with user tasks. To address the limitations of prior works, in this research, we investigate the effectiveness of both human eye and head movements in recognizing tasks in immersive virtual reality.

2.3 Coordination of Eye and Head Movements

Human eye and head movements have been extensively investigated in the fields of cognitive science and human-centered computing. Some researchers focused on eye-head coordination [2, 3, 24, 37, 38], which refers to the coordinated movements between the eyes and the head. Stahl found that the eyes and the head move in coordination during gaze shifts and that head movement amplitude is proportional to gaze shift amplitude [37]. Fang et al. further discovered that eye-head coordination is involved in gaze fixation and plays a role in visual cognitive processing [24]. Hu et al. focused on eye-head coordination in virtual reality and revealed strong correlations between human gaze positions and head rotation velocities [2, 3]. Sidenmark et al. identified general eye, head, and torso coordination patterns during gaze shifts in virtual reality [38]. Other researchers concentrated on the applications of eye and head movements [1, 23, 39–41]. Gandrud et al. utilized gaze direction and head orientation to predict direction of locomotion in virtual reality [41]. Kytö et al. [39] and Sidenmark et al. [42] leveraged eye and head movements to improve target selection techniques. Kothari et al. employed the magnitudes of eye and head movements to classify gaze events (i.e. fixations, pursuits, and saccades) [23]. Recently, Hu et al. proposed a learning-based method to forecast future eye fixations using past gaze positions and head rotation velocities [1]. In contrast with prior works, in this research we employed eye and head movements to recognize user tasks.

3 Data Collection

3.1 Stimuli

To collect human eye and head movements of performing different tasks in virtual reality, we employed 360-degree VR videos as our stimuli to ensure that the same VR content was presented to a user under different task conditions. Specifically, 15 videos were selected from three publicly available 360-degree video datasets

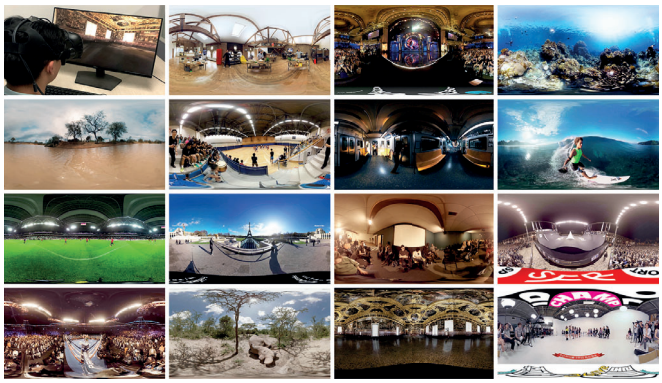


Fig. 1: The experimental setup (top-left) and the 15 360-degree VR videos used in our experiments.

[43–45] to provide a wide variety of content, which include indoor scenes, cities, outdoor scenarios, sports, movies, and shows (Figure 1). Eleven videos were captured by a stationary camera and four videos were recorded using a moving camera. These videos are monoscopic and not interactive. They were projected onto the inner surface of a sphere and viewers could observe the videos from the inside of the sphere using a VR headset. Each selected video has a resolution of 3840×2160 pixels and a frame rate of 30 fps. The original videos have different lengths, and to ensure that the task duration is the same in different videos, we extracted a 150-second segment from each video for data collection.

3.2 Participants and Apparatus

We recruited 30 participants (18 male, 12 female, age $\mu = 24.5, \sigma = 5.0$) to take part in our experiments. Each participant reported normal or corrected-to-normal vision. The eye tracker was calibrated to each user before starting the experiment. We conducted the data collection experiments on a computer with an Intel(R) Core(TM) i7-8700 @ 3.20GHz CPU and an NVIDIA GeForce RTX 2060 SUPER GPU. The 360-degree VR videos were displayed using an HTC Vive head mounted device (HMD), equipped with a 7invensun VR eye tracker running at 100 Hz and providing an accuracy of 0.5°. We used the Unity3D game engine to render the VR videos and employed our own Unity scripts to record human eye movements (accessed from the eye tracker) and head movements (accessed from the HMD) at a sampling rate of 100 Hz. The snapshot of our experimental setup is demonstrated on the top left of Figure 1.

3.3 Procedure

In our experiments, each participant was asked to explore three 360-degree VR videos that were randomly chosen from the 15 videos. Each video was played four times for a user, in which the user was required to complete the following four tasks (one task at a time) in random order:

- *Free viewing:* Freely explore the 360-degree VR video;
- *Visual search:* Locate and count as many objects with geometrical shapes, e.g. triangles, circles, and rectangles, as you can find in the scene;
- *Saliency:* Estimate which half of the scene (top or bottom) is more salient;
- *Track:* Keep in view the nearest moving object in your field of view and track it with your eyes.

These tasks are not only typically used in existing task datasets [15, 23, 46, 47] but also have crucial importance for VR applications [1–4, 7, 48]. Studying these tasks contributes to not only understanding the mechanisms of human visual attention in VR [1, 7] but also deriving models recognizing user tasks from visual attention in immersive virtual reality [8–11]. Each task lasted for 150 seconds, i.e. the same length as a video, and the videos were set to silent to avoid auditory disturbance.

During the experiments, we collected the class of the task, human eye movements, and human head movements for further analysis. Specifically, we recorded human gaze positions on the screen of the HMD ($(e_x, e_y), e_x, e_y \in [0, 1]$) and human head orientation in the 360-degree virtual world ($(h_x, h_y), h_x \in [-180°, 180°], h_y \in [-90°, 90°]$). Using the head orientation information, we further converted on-screen gaze positions to the gaze positions in the 360-degree virtual world ($(g_x, g_y), g_x \in [-180°, 180°], g_y \in [-90°, 90°]$). For clarity, we utilized eye-in-head (EiH) data to denote on-screen gaze positions and employed gaze-in-world (GiW) data to represent gaze positions in the virtual world. The EiH data reflects human eye movements with respect to the head while the GiW data shows the combined influence of human eye and head movements [23].

Our dataset (**EHTask-dataset**) contains exploration data of 30 participants, totalling 360 recordings (30 participants $\times$ 3 videos $\times$ 4 tasks). Each recording is annotated with the task, EiH data (100 Hz), GiW data (100 Hz), and head orientation data (100 Hz) in a 150-second 360-degree video. Each of the 15 videos was observed by six users. Table 1 provides a comparison of our dataset with other related datasets. We can see that our dataset is the first VR dataset that contains both human eye and head movements in different tasks.

4 Eye Movements, Head Movements, and Task

Human eye movements and head movements in immersive VR may be severely influenced by the specific tasks assigned to them. To investigate which features or movements are discriminative for which task, in this section we conducted a comprehensive analysis of human eye and head movements in different VR tasks based on our dataset. Specifically, we analyzed the characteristics of human eye movements, the characteristics of human head movements, and the characteristics of eye-head coordination.

4.1 Eye Movements and Task

The patterns of human eye movements can be classified into fixations (pauses over regions of interest) and saccades (rapid eye movements between fixations). To analyze the characteristics of human eye movements in different tasks, we employed a thresholding method based on dispersion and duration to detect fixations and saccades from EiH data [52]. The maximum dispersion of fixations was set to 1° and the minimum duration of fixations was set to 150 ms [4].

We computed the statistical characteristics of the detected fixations and saccades. Specifically, we first calculated the mean fixation duration, fixation number per second, mean saccade duration, saccade number per second, and mean saccade amplitude for each recording. Then we computed the means and standard deviations

TABLE 1: A comparison between our dataset and other related datasets. Our dataset is the first task dataset for immersive virtual reality that contains the information on both human eye and head movements.

Datasets	Stimuli	Viewers	Eye	Head	Task Duration	Tasks
Greene et al. [49]	64 images	16	√	×	10 s	Memory ◇ Decade ◇ People ◇ Wealth
Borji et al. [50]	15 images	21	√	×	30 s	Yarbus' original 7 tasks [12]
Koehler et al. [46]	800 images	19	√	×	2 s	Free viewing ◇ Object search ◇ Saliency
Sugano et al. [47]	480 image pairs	14	√	×	10 s	Free viewing ◇ Preference
Kübler et al. [51]	2 paintings	20	√	×	30-120 s	Free viewing ◇ Age estimating
Hild et al. [15]	1 video	30	√	×	4 min	Explore ◇ Observe ◇ Search ◇ Track
Bulling et al. [17]	Real World	8	√	×	5 min	Copy ◇ Read ◇ Write ◇ Video ◇ Browse ◇ Null
GW dataset [23]	Real world	19	√	√	3 min	Navigation ◇ Ball catching ◇ Visual search ◇Tea making
Ours	15 VR videos	30	√	√	150 s	Free viewing ◇ Visual search ◇ Saliency ◇ Track

(SDs) of the above features for the recordings belonging to the four tasks respectively. The results are indicated in Table 2. To investigate whether the differences between the statistics of the four tasks are significant, we first ran a one-way repeated measures analysis of variance (ANOVA) test to evaluate the differences between the four tasks. If the differences were statistically significant, we further ran a post-hoc Tukey's honest significant difference test (Tukey's HSD test) to perform pairwise comparisons among the four tasks. We find that the differences between the four tasks are statistically significant in terms of mean fixation duration ($F(3,180) = 291.1, p = 8.57E-69 < 0.01$), fixation number per second ($F(3,180) = 399.7, p = 2.61E-79 < 0.01$), mean saccade duration ($F(3,180) = 612.8, p = 3.45E-94 < 0.01$), saccade number per second ($F(3,180) = 59.8, p = 7.10E-27 < 0.01$), and mean saccade amplitude ($F(3,180) = 428.8, p = 1.04E-81 < 0.01$) and the differences between every two tasks also have statistical significance (Tukey's HSD test, $p < 0.01$). The above results correspond with previous findings that human eye movement patterns are different across different tasks [7,12,50,53]. An exception to this is that there is no significant difference between *Free viewing* task and *Track* task (Tukey's HSD test, $p = 0.878$) in terms of saccade number per second. Generally, we expect *Track* task to have fewer saccades than *Free viewing* task because observers are required to fixate on the nearest moving object in the *Track* task. However, the nearest moving object in our VR videos usually moves very fast, which may increase observers' saccades and make the difference between *Free viewing* task and *Track* task not significant.

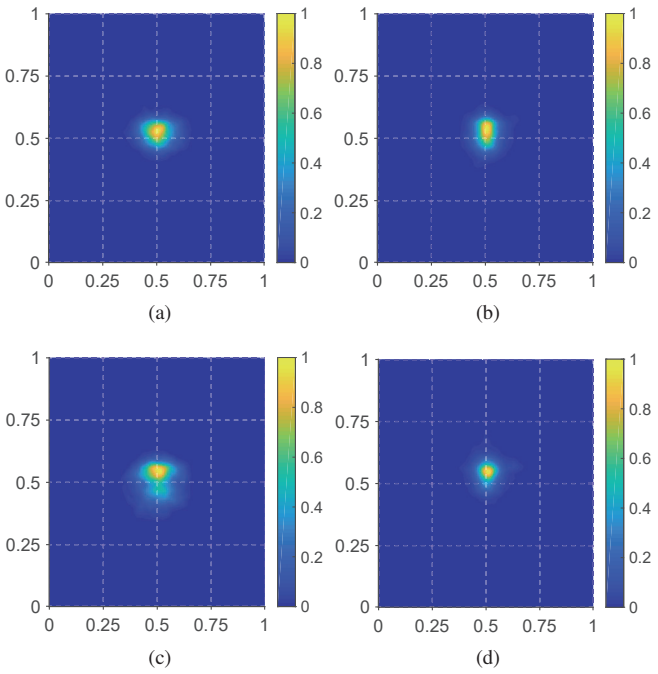


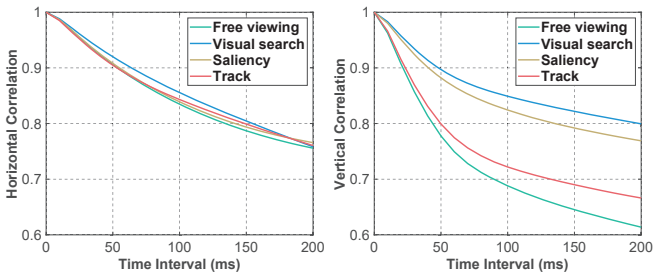
Fig. 2: The distributions of human fixation positions on the HMD's screen in (a) *Free viewing* task, (b) *Visual search* task, (c) *Saliency* task, and (d) *Track* task.

To gain a sound understanding of eye fixations in the four tasks, we analyzed the distributions of fixation positions. Figure 2 illustrates the distributions of fixation positions on the HMD's screen, which are smoothed using a Gaussian filter with sigma equal to one degree of visual angle [54]. We find that, in each task, most of the fixation positions lie in the central region of the screen, which corresponds with previous findings [1–4]. This is because when observers move their heads little or not at all, their gaze-shift sizes are usually limited to a small range of about $\pm 18^\circ$ [24]. As a consequence, observers' eye-in-head fixation positions are usually limited to the central region of the HMD's screen regardless of the task being performed. We further analyzed the dispersions of fixation position distributions. Specifically, we utilized the determinant of the co-variance matrix between horizontal and vertical fixation coordinates as a measure for dispersion [7] and indicated the results in Table 2. We find that there exists a significant difference between the four tasks ($F(3,180) = 194.0, p = 3.70E-56 < 0.01$) and *Saliency* task has significant difference with the other three tasks (Tukey's HSD test, $p < 0.01$). However, there is no significant difference between *Free viewing* task and *Visual search* task (Tukey's HSD test, $p = 0.996$), no significant difference between *Free viewing* task and *Track* task (Tukey's HSD test, $p = 0.352$), and no significant difference between *Visual search* task and *Track* task (Tukey's HSD test, $p = 0.527$). This is because *Saliency* task requires the observers to frequently compare the top and bottom half of the scene, which makes the fixation positions more dispersed than the other three tasks.

To analyze the temporal characteristics of human eye movements in the four tasks, we calculated the auto-correlations of the horizontal and vertical eye coordinates of EiH data respectively [55]. The auto-correlation of a time series is defined as the Pearson correlation between the time series and a delayed copy of itself. We first calculated the horizontal and vertical auto-correlations for each recording and then computed the means for the recordings belonging to the four tasks respectively. Figure 3 illustrates the horizontal and vertical auto-correlations at different time intervals.

TABLE 2: Statistical characteristics of human eye movements in the four tasks. For each item, the difference in the fonts of two tasks indicates that there exists a significant difference between them (Tukey's HSD test, $p < 0.01$). The same font indicates no statistical significance.

		Free viewing	Visual search	Saliency	Track
Mean Fixation Duration	Mean	**263.4** ms	*339.5* ms	241.2 ms	431.7 ms
	SD	25.6 ms	49.0 ms	24.3 ms	106.7 ms
Fixation Number Per Second	Mean	**1.41**	*1.97*	1.22	1.77
	SD	0.38	0.17	0.43	0.19
Mean Saccade Duration	Mean	**633.2** ms	*269.3* ms	776.0 ms	241.1 ms
	SD	218.0 ms	69.2 ms	260.1 ms	56.2 ms
Saccade Number Per Second	Mean	**1.03**	*1.20*	0.95	**1.01**
	SD	0.17	0.18	0.19	0.24
Mean Saccade Amplitude	Mean	**6.51**°	*4.73*°	8.56°	5.40°
	SD	1.24°	1.05°	1.49°	1.58°
Fixation Distribution Dispersion	Mean	**2.21E-6**	**2.25E-6**	7.08E-6	**2.50E-6**
	SD	1.01E-6	1.18E-6	3.50E-6	1.57E-6



(a) (b)

Fig. 3: The auto-correlations of the (a) horizontal eye coordinates and (b) vertical eye coordinates at different time intervals.

In the horizontal direction, we find that the auto-correlations of the four tasks are very close and there is no significant difference between the four tasks ($F(3,180) = 0.848, p = 0.469$) at the time interval of 200 *ms*. However, in terms of vertical auto-correlations, we find that the difference between the four tasks is statistically significant ($F(3,180) = 89.7, p = 1.52E-35 < 0.01$) at 200 *ms* and the differences between every two tasks are statistically significant (Tukey's HSD test, $p < 0.01$). This may be because tasks have a higher influence on observers' vertical gaze behaviours than on horizontal gaze behaviours [1, 3, 55].

To sum up, we conducted a comprehensive analysis of human eye movements in different tasks and observed significant differences in the aspects of mean fixation duration, fixation number per second, mean saccade duration, saccade number per second, mean saccade amplitude, fixation distribution dispersion, and vertical auto-correlation. Our results suggest that the characteristics of human eye movements can serve as clues to recognize tasks in VR.

4.2 Head Movements and Task

To analyze the characteristics of human head movements in the four tasks, we calculated the mean absolute horizontal velocity, mean absolute vertical velocity, mean absolute horizontal acceleration, and mean absolute vertical acceleration for different tasks. We also calculated the dispersions of head velocity distributions by utilizing the determinant of the co-variance matrix between horizontal and vertical head velocities as a measure for dispersion [7]. Specifically, we first calculated the statistics for each recording and then computed the means and SDs for the recordings belonging to the four tasks respectively. The results are indicated in Table 3. To analyze the differences between different tasks, we first ran a one-way repeated measures ANOVA test to evaluate the differences between the four tasks and if the results were significant, we further ran a post-hoc Tukey's HSD test to perform pairwise comparisons among the four tasks. We find that the differences between the four tasks are statistically significant in the aspects of mean absolute horizontal velocity ($F(3,180) = 1328.5, p = 1.68E-122 < 0.01$), mean absolute vertical velocity ($F(3,180) = 1494.8, p = 6.38E-127 < 0.01$), mean absolute horizontal acceleration ($F(3,180) = 296.9, p = 1.96E-69 < 0.01$), mean absolute vertical acceleration ($F(3,180) = 195.6, p = 2.14E-56 < 0.01$), and velocity distribution dispersion ($F(3,180) = 613.7, p = 3.03E-94 < 0.01$) and the differences between every two tasks also have statistical significance (Tukey's HSD test, $p < 0.01$). The above results reveal that the patterns of human head movements are different across different tasks, indicating that the characteristics of human head movements can be employed to recognize user tasks. An exception to this is that there is no significant difference between *Free viewing* task and *Visual search* task in the aspect of mean absolute vertical velocity (Tukey's HSD test, $p = 0.278$). Generally, we expect *Free viewing* task to have larger vertical velocity than *Visual search* task because observers have more freedom to move their heads in *Free viewing* task. However, we find that, agreeing with prior work [6], observers preferred to explore the 360° VR videos in the horizontal direction (mean absolute horizontal velocity: $22.7°/s$) than in the vertical direction (mean absolute vertical velocity: $2.9°/s$) in *Free viewing* task, possibly because the horizontal view (360°) of the 360° VR videos is much larger than the vertical view (180°). As a consequence, the mean absolute vertical velocity in *Free viewing* task is smaller than expected and the difference between *Free viewing* task and *Visual search* task is not significant. Another exception is that there is no significant difference between *Visual search* task and *Track* task (Tukey's HSD test, $p = 0.030 > 0.01$) in the aspect of mean absolute vertical acceleration. Generally, we expect *Track* task to have lower vertical acceleration than *Visual search* task because observers are required to fixate on the nearest moving object in the *Track* task. However, the nearest moving object in our VR videos usually moves very fast, which may increase observers' vertical acceleration and make the difference between *Visual search* task and *Track* task not significant.

To summarize, we conducted a comprehensive analysis of human head movements in different tasks and observed significant

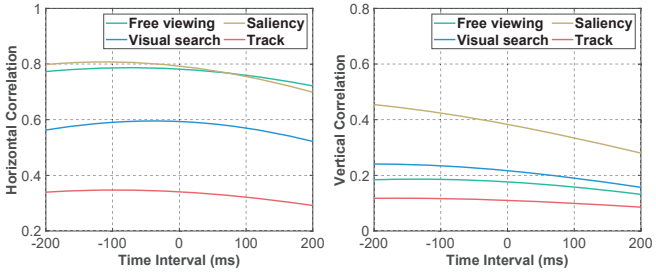
TABLE 3: Statistical characteristics of human head movements in the four tasks. For each item, the same font of two tasks means that the difference between the two tasks is not statistically significant (Tukey's HSD test, $p > 0.01$) while different fonts indicate statistical significance.

		Free viewing	Visual search	Saliency	Track
Mean Absolute Horizontal Velocity	Mean	**22.7**$^\circ/s$	*9.1*$^\circ/s$	26.8$^\circ/s$	6.4$^\circ/s$
	SD	4.3$^\circ/s$	2.3$^\circ/s$	4.4$^\circ/s$	2.4$^\circ/s$
Mean Absolute Vertical Velocity	Mean	**2.9**$^\circ/s$	**2.7**$^\circ/s$	7.5$^\circ/s$	1.9$^\circ/s$
	SD	0.6$^\circ/s$	0.5$^\circ/s$	1.4$^\circ/s$	0.4$^\circ/s$
Mean Absolute Horizontal Acceleration	Mean	**182.6**$^\circ/s^2$	*140.4*$^\circ/s^2$	203.5$^\circ/s^2$	129.8$^\circ/s^2$
	SD	29.4$^\circ/s^2$	14.1$^\circ/s^2$	23.9$^\circ/s^2$	19.4$^\circ/s^2$
Mean Absolute Vertical Acceleration	Mean	**125.0**$^\circ/s^2$	*114.2*$^\circ/s^2$	145.4$^\circ/s^2$	*109.4*$^\circ/s^2$
	SD	15.0$^\circ/s^2$	11.1$^\circ/s^2$	12.0$^\circ/s^2$	11.6$^\circ/s^2$
Velocity Distribution Dispersion	Mean	**2.64E+4**	*6.95E+3*	2.39E+5	3.12E+3
	SD	2.13E+4	7.98E+3	1.27E+5	4.35E+3

differences in the aspects of mean absolute horizontal velocity, mean absolute vertical velocity, mean absolute horizontal acceleration, mean absolute vertical acceleration, and velocity distribution dispersion. Our results reveal that human head movements are severely affected by the specific tasks assigned to them, suggesting that the characteristics of human head movements can be applied to recognize user tasks.

4.3 Eye-Head Coordination and Task

Eye-head coordination refers to the coordinated movements between the eyes and the head. Some researchers found that head movement amplitude is proportional to gaze shift amplitude in real-world situations [24, 37] while other researchers revealed that human on-screen gaze positions are correlated with their head rotation velocities in virtual reality [1–3]. To analyze the eye-head coordinations in the four VR tasks, we calculated the correlations between human on-screen gaze positions and their head rotation velocities in the horizontal and vertical directions respectively using Spearman's rank correlation coefficient [1, 3], which measures the monotonic relationship between two variables. Specifically, we first calculated the horizontal and vertical correlations for each recording and then computed the means for the recordings belonging to the four tasks respectively. Figure 4 illustrates the eye-head correlations in the horizontal and vertical directions. We performed a one-way repeated measures ANOVA test on the correlations of the four tasks at the time interval of 0 *ms* and if the differences between the four tasks were significant, we further ran a post-hoc Tukey's HSD test to perform pairwise comparisons among the four tasks. We find that the differences between the four tasks are statistically significant in terms of horizontal eye-head correlation ($F(3,180) = 548.7, p = 2.80E-90 < 0.01$) and vertical eye-head correlation ($F(3,180) = 308.0, p = 1.27E-70 < 0.01$) and the differences between every two tasks also have statistical significance (Tukey's HSD test, $p < 0.01$). The above results reveal that the patterns of eye-head coordination are different across different tasks. This is because different tasks can induce different visual cognitive processings [1, 7] and thus induce different eye-head coordinations because eye-head coordination is influenced by visual cognitive processing [24]. An exception is that there is no significant difference between *Free viewing* task and *Saliency* task (Tukey's HSD test, $p = 0.086 > 0.01$) in the aspect of horizontal eye-head correlation at the time interval of 0 *ms*. This reflects that the visual cognitive processings of *Free viewing* task and *Saliency* task have similarities in the horizontal direction.



(a) (b)

Fig. 4: The Spearman's correlations between on-screen gaze positions and head rotation velocities in the (a) horizontal direction and (b) vertical direction at different time intervals.

To summarize, we analyzed the characteristics of eye-head coordination in different tasks and observed significant differences in the aspects of horizontal eye-head correlation and vertical eye-head correlation. Our results indicate that different tasks have different influences on eye-head coordination, suggesting that the inner connection between eye movements and head movements can provide meaningful information for task recognition.

5 EHTask Model

Based on the analysis in Section 4, we propose a learning-based model called *EHTask* that combines the eye and head movements to recognize user tasks (Figure 5). *EHTask* consists of four modules: an EiH module that extracts features from eye-in-head data, a GiW module that extracts features from gaze-in-world data, a head module that extracts features from head rotation velocities, and a task recognition module that recognizes user tasks from the extracted features.

The EiH module aims at extracting features from the eye-in-head time-series data ($E_i \in R^2$). Previous work on gaze prediction reveals that 1D convolutional neural network (CNN) has good performance for extracting features from gaze and head time-series data [1, 3] while bidirectional gated recurrent unit (BiGRU) has also been proven to be powerful for processing sequence data [56]. An intuitive idea is to combine the above two architectures to produce better results. Therefore, the EiH module first employs 1D CNN layers to extract features for each time step of the eye-in-head data and then applies BiGRU to extract temporal features from the output of the 1D CNN layers. Specifically, three 1D CNN layers, each with 16 channels and a kernel size of three,

are employed for feature extraction. Each CNN layer is followed by a batch normalization layer, a ReLU activation function, and a max-pooling layer with a kernel size of two. After the CNN layers, a BiGRU layer with hidden size (the number of features in the hidden state) of 64 is applied to extract temporal features. The BiGRU layer outputs the hidden states of the first and last time steps respectively for task recognition.

The GiW module extracts features from the gaze-in-world time-series data ($G_i \in R^2$) while the head module is utilized to extract features from the time series of the head rotation velocities ($H_i \in R^2$). The same network structure as EiH module is employed for the GiW module and head module respectively for feature extraction.

The task recognition module combines the outputs of the EiH module, GiW module, and head module to recognize user tasks. Specifically, this module employs two fully connected (FC) layers, each with 64 neurons, to integrate the extracted features. Each FC layer is followed by a batch normalization layer, a ReLU activation function, and a dropout layer with dropout rate 0.5 to improve the network's generalization ability. A Softmax layer is applied after the second FC layer to generate the probability of each task.

To train our model, we first down-sampled the original recordings in the dataset to a frequency of 25 Hz for simplicity [8]. Then we segmented the recordings into small windows and employed these windows to train our model. The window size was set to 10 seconds ($\Delta t_1 = \Delta t_2 = \Delta t_3 = 10$ s) because the duration of 10 seconds has been proven to be effective enough for recognizing user tasks [8,9,22,47]. The interval between two adjacent windows was set to one second. We employed cross entropy loss as the loss function and utilized Adam optimizer with weight decay $1.0e^{-4}$ to minimize the loss. We set the initial learning rate to $1.0e^{-2}$ and employed an exponential decay strategy that decayed the learning rate by γ every epoch. We set γ to 0.75 and employed a batch size of 256 to train our model for totally 30 epochs. Our model was implemented using the PyTorch framework.

6 Experiments and Results

In this section, we conducted extensive experiments to evaluate our model's task recognition performance. Specifically, we first compared our model with the state-of-the-art methods derived from 2D viewing conditions on our dataset using a cross-user evaluation and a cross-scene evaluation respectively. We further evaluated our model's performance on a newly released real-world task dataset [23] to test our model's generalization capability for different situations. We also performed an ablation study to validate the effectiveness of each component in our model.

6.1 Evaluation Metric and Comparison

As commonly used in prior works [8, 9, 13, 22], we employed classification accuracy as the metric to evaluate the performances of task recognition methods. We compared the performance of our model with the following state-of-the-art methods derived from 2D viewing conditions:

- *Linear Discriminant Analysis (LDA):* Linear discriminant analysis has been proven to be effective for task recognition in prior works [8, 15]. We utilized the implementation provided in Coutrot et al.'s Matlab toolbox [8] and trained the model from scratch using its default settings. This Matlab toolbox extracts features from raw eye movements using hidden Markov models (HMM) and then utilizes the HMM features to train task recognition methods. We respectively utilized the raw eye movements and the HMM features of the raw eye movements to train LDA and got *LDA_r* (LDA using raw eye movements) and *LDA_h* (LDA using HMM features) for comparison.
- *Support Vector Machine (SVM):* As shown in prior works [11,16–18,51,53], support vector machines can be applied to recognize user tasks. We employed the implementation provided in Coutrot et al.'s Matlab toolbox [8] and used the default settings to train it from scratch. Raw eye movements and the HMM features of the raw eye movements were trained respectively to produce two models, i.e. *SVM_r* and *SVM_h*.
- *Boosting Classifier (BC):* Boosting classifiers have been successfully used for task recognition in previous works [8,50]. We used the implementation of AdaBoost provided in Coutrot et al.'s Matlab toolbox [8] and trained it from scratch using the default settings. *BC_r* and *BC_h* were trained for comparison using the raw eye movements and the HMM features of the raw eye movements respectively.
- *Random Forests (RFo):* Random forests were frequently used to recognize user tasks [9, 13, 15, 47]. We used the implementation provided in Coutrot et al.'s Matlab toolbox [8] and trained *RFo_r* and *RFo_h* from raw eye movements and the HMM features of the raw eye movements respectively using the default settings.
- *Random Ferns (RFe):* Random ferns were recently applied to the problem of task recognition [22]. We employed the implementation provided by Fuhl et al. [22], which recognizes user tasks from raw eye movements. We trained *RFe* from scratch using the default parameters for comparison.

6.2 Recognition Performance

6.2.1 Cross-User Evaluation

We first performed a cross-user evaluation to evaluate our model's generalization capability for different users. Specifically, we first segmented the original recordings into windows of 10 seconds (Section 5) and then evenly divided all the windows into five folds according to different users. We trained the methods on four folds, and tested on the remaining one fold. Each method was trained and tested for five times in total in which each fold was tested once. The recognition results in each test were collected for further analysis. We calculated the mean classification accuracy of the five tests for each method and indicated the results in Table 4 (Cross-User, Window). We can see that our model outperforms the state-of-the-art methods by a large margin (accuracy of 84.4% *vs.* 62.8%). We further performed a paired Wilcoxon signed-rank test to compare the recognition results of our model with the second-best method and validated that the difference between our model and the second-best method is statistically significant ($p < 0.01$). The above results validate that our model has a high accuracy for recognizing user tasks and has a strong generalization capability for different users.

Figure 6 (a) shows the confusion matrix of our model's cross-user recognition results. Each diagonal element of the confusion matrix represents the recognition accuracy for each class while the off-diagonal elements indicate the probabilities of mislabeling one class as another. The higher the diagonal values of the confusion matrix, the better the recognition performance. We can see that

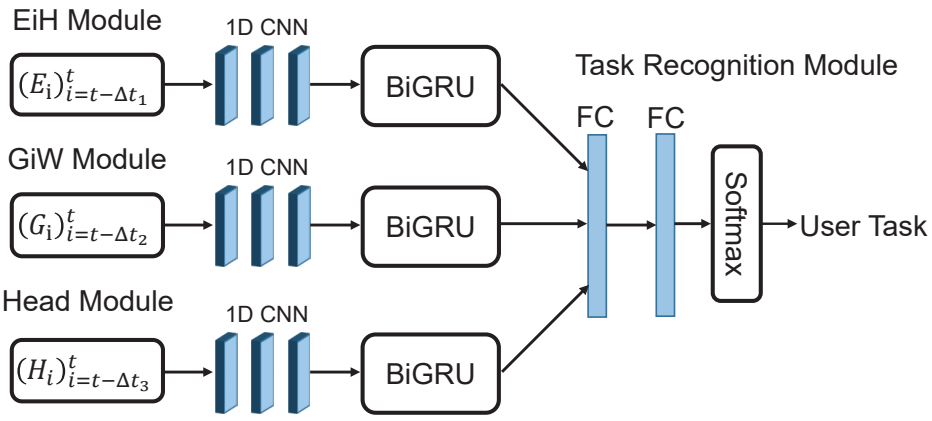


Fig. 5: Architecture of the proposed model *EHTask*.

our model maintains a high recognition accuracy for each class, which validates the effectiveness of our model. Furthermore, we find that the largest confusion takes place between *Visual search* task and *Track* task (11.8% and 12.9%). By examining our analysis in Section 4, we find that *Visual search* task and *Track* task have many similarities in terms of influences on human eye movements and head movements. For example, there is no significant difference between *Visual search* task and *Track* task in the aspects of fixation distribution dispersion (Table 2) and mean absolute vertical acceleration of head movements (Table 3). The similar influences on human eye and head movements may degrade our model's performance because our model relies on the features extracted from human eye and head movements to discriminate different tasks. In addition, we also find that our model has the highest accuracy for recognizing *Saliency* task. This may be because *Saliency* task has some distinct influences on human eye and head movements compared with the other three tasks. For example, in terms of fixation distribution dispersion (Table 2), *Saliency* task has the largest dispersion and the differences between *Saliency* task and the other three tasks are statistically significant while there exists no statistical significance between the other three tasks.

In the above evaluation, we segmented a whole recording into windows of 10 seconds and only evaluated our model's performance on the windows. To further evaluate our model's performance on the whole recordings, we employed a majority voting (MV) strategy that utilized the majority voting result of all the windows belonging to one recording to recognize the task of this recording. The majority voting recognition performances of different methods are indicated in Table 4 (Cross-User, MV). We can see that our model outperforms the state-of-the-art methods and the difference between our model and the second-best method is statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$). Furthermore, we find that our model achieves a large improvement using majority voting over that of using windows (97.8% *vs.* 84.4%), which validates the effectiveness of our majority voting strategy.

6.2.2 Cross-Scene Evaluation

Our dataset consists of recordings from 15 scenes. To evaluate our model's generalization capability for different scenes, we segmented the original recordings into 10-second windows, evenly divided all the windows into five folds according to different scenes, and performed a five-fold cross-scene evaluation to test our model and other methods. The recognition performances of different methods are indicated in Table 4 (Cross-Scene, Window). We can see that our model achieves a large improvement over the state-of-the-art methods (82.1% *vs.* 62.6%). We further performed a paired Wilcoxon signed-rank test to compare our model with the second-best method and the result validates that the difference between our model and the second-best method is statistically significant ($p < 0.01$). The above results validate that our model has a high accuracy for recognizing user tasks and a strong generalization capability for different scenes.

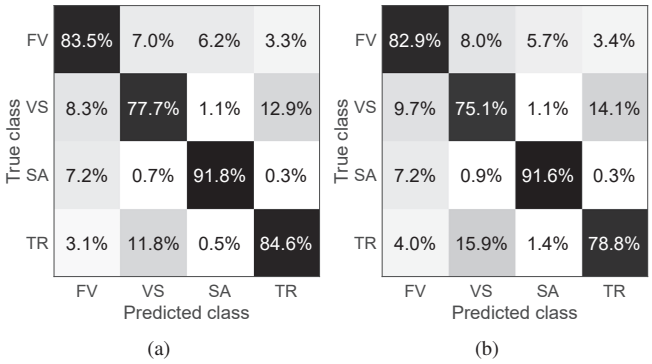


Fig. 6: Confusion matrices of our model's (a) cross-user recognition results and (b) cross-scene recognition results, normalised across ground truth rows. FV: *Free viewing*; VS: *Visual search*; SA: *Saliency*; TR: *Track*.

The confusion matrix of our model's cross-scene recognition results is illustrated in Figure 6 (b). We can see that, similar to the situation of cross-user evaluation (Figure 6 (a)), our model has a high accuracy for recognizing each class and the largest confusion takes place between *Visual search* task and *Track* task (14.1% and 15.9%). In addition, the confusion matrix also indicates that our model has the highest accuracy for recognizing *Saliency* task.

We further evaluated the majority voting recognition performances of our model and other methods. Specifically, we employed the majority voting result of all the windows belonging to one recording to recognize user tasks in this recording. The results are indicated in Table 4 (Cross-Scene, MV). We can see that our model outperforms other methods, achieving a high accuracy of

TABLE 4: Task recognition performances of different methods on our dataset. In each row, the best method is emphasized using bold font and the second-best method is stressed using a underline. Our model outperforms other methods in both cross-user and cross-scene settings.

		Ours	*LDA_r*	*LDA_h*	*SVM_r*	*SVM_h*	*BC_r*	*BC_h*	*RFo_r*	*RFo_h*	*RFe*
Cross-User	Window	**84.4%**	37.2%	54.0%	29.5%	54.3%	41.5%	49.3%	<u>62.8%</u>	58.0%	48.7%
	MV	**97.8%**	42.8%	76.1%	34.2%	75.3%	47.5%	65.3%	83.1%	<u>88.9%</u>	68.3%
Cross-Scene	Window	**82.1%**	37.2%	53.8%	26.3%	54.1%	41.2%	49.0%	<u>62.6%</u>	57.9%	48.3%
	MV	**96.4%**	41.9%	74.2%	26.7%	75.3%	47.5%	64.4%	83.6%	<u>87.2%</u>	72.2%

96.4%. The result of a paired Wilcoxon signed-rank test reveals that there exists a significant difference between our model and the second-best method ($p < 0.01$). In addition, we find that our model achieves a large improvement using majority voting over that of using windows (96.4% *vs.* 82.1%). This validates that the majority voting strategy is effective in cross-scene settings.

6.3 Performance in Real World

Our model recognizes task using only the eye and head movements of a user. This ensures that our model can be easily applied to other situations besides immersive virtual reality as long as human eye and head movements in the corresponding situations are available. To test our model's generalization capability for different situations, we evaluated our model on a newly released real-world dataset, i.e. GW dataset [23]. GW dataset contains the eye and head movements of 19 participants performing four everyday tasks, i.e. indoor navigation, ball catching, visual search, and tea making, in an indoor environment. The duration of each task was approximately three minutes. Task recognition in real-world situations may be more challenging than that in immersive VR. This is because human physical movements are usually limited to a small region in immersive VR, making different users share similar eye and head movements in the same task. However, in real-world situations, human physical movements may have more freedom and greater randomness, which leads to different eye and head movements even in the same task.

We first segmented the original recordings in GW dataset into windows of 10 seconds and then evenly divided all the windows into five folds according to different users. We further performed a five-fold cross-user evaluation to test our model and other methods. The recognition performances of different methods are indicated in Table 5 (Window). We can see that our model achieves a large improvement over the state-of-the-art methods (61.9% *vs.* 44.1%) and the difference between our model and the second-best method is statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$).

We further evaluated the majority voting recognition performances of our model and other methods on GW dataset. Specifically, the majority voting result of all the windows belonging to one recording was employed to recognize user tasks in this recording. The results are indicated in Table 5 (MV). We find that our model achieves a high accuracy of 87.7%, outperforming the state-of-the-art methods. We further performed a paired Wilcoxon signed-rank test and the result indicates that the difference between our model and the second-best method is statistically significant ($p < 0.01$). Furthermore, we find that our model achieves higher accuracy using majority voting than that of using windows (87.7% *vs.* 61.9%), which validates that the majority voting strategy helps improve recognition accuracy in real-world settings. By comparing all the methods' performances on our dataset (Table 4) with that on the real-world dataset (Table 5), we find that all the methods achieve a higher accuracy in immersive VR than in real-world situations. This validates that task recognition in real-world situations is more challenging than that in immersive VR.

6.4 Ablation Study

We performed an ablation study to evaluate the effectiveness of each component in our model. Specifically, we retrained our model on our dataset using only EiH data, using only Head data, using only GiW data, using EiH and Head data, using EiH and GiW data, using Head and GiW data, using only the CNN modules, and using only the BiGRU modules, respectively. We segmented the recordings in our dataset into 10-second windows and evaluated the ablated models using a five-fold cross-user evaluation and a five-fold cross-scene evaluation. Table 6 indicates the recognition performances of our model and the ablated models. We find that our model achieves higher accuracy than all the ablated models in terms of both cross-user evaluation and cross-scene evaluation. We further employed paired Wilcoxon signed-rank tests to perform pairwise comparisons between our model and each ablated model and validated that the differences between our model and the ablated models are statistically significant ($p < 0.01$). The above results indicate that each component in our model helps improve our model's task recognition accuracy.

To further validate the effectiveness of our model's architecture, we also evaluated other architectures on our dataset: (1). We replaced our CNN+BiGRU architecture with a bidirectional long short-term memory (BiLSTM) layer with hidden size of 64 to extract features for task recognition (2). We replaced our CNN+BiGRU architecture with a CNN+BiLSTM architecture for feature extraction. The same CNN architecture as our model was employed and a BiLSTM layer with hidden size of 64 was applied after the CNN. (3). We employed the eye-head statistics indicated in Table 2 and Table 3 as hand-crafted features to train the state-of-the-art methods, i.e. *LDA*, *SVM*, *RFo*, and *BC*. The cross-user and cross-scene recognition performances of these architectures are indicated in Table 7. We can see that our model outperforms these architectures and the results are statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$).

6.5 Runtime Performance

Our model was implemented on an NVIDIA TITAN Xp GPU platform with an Inter(R) Xeon(R) E5-2620 v4 2.10 GHz CPU. The average run time for recognizing task from a 10-second window was 0.10 *ms* on the GPU and 1.19 *ms* on the CPU. These results show that our model is light-weight enough and ready for real-time usage.

7 Discussion

Our work has made an important step towards understanding human visual attention under different VR tasks and recognizing

TABLE 5: Task recognition performances of different methods on GW dataset. In each row, bold font is used to emphasize the best method and a underline is applied to indicate the second-best method. Our model performs better than other methods in terms of both the results of windows and the results of majority voting.

	Ours	*LDA_r*	*LDA_h*	*SVM_r*	*SVM_h*	*BC_r*	*BC_h*	*RFo_r*	*RFo_h*	*RFe*
Window	**61.9%**	26.4%	39.0%	26.2%	37.9%	36.3%	34.1%	<u>44.1%</u>	42.3%	36.1%
MV	**87.7%**	26.2%	60.0%	32.3%	46.2%	33.8%	40.0%	53.8%	60.0%	<u>64.6%</u>

TABLE 6: The recognition performances of our model and the ablated models. Our model outperforms all the ablated models, validating the effectiveness of each component in our model.

	Ours	EiH	Head	GiW	EiH+Head	EiH+GiW	Head+GiW	CNN	BiGRU
Cross-User	**84.4%**	74.5%	79.9%	75.5%	82.1%	80.0%	83.1%	83.5%	81.3%
Cross-Scene	**82.1%**	73.9%	79.5%	70.3%	81.2%	77.0%	80.1%	81.5%	78.8%

TABLE 7: The cross-user and cross-scene recognition performances of our model and other architectures.

	Ours	BiLSTM	CNN+BiLSTM	*LDA*	*SVM*	*RFo*	*BC*
User	**84.4%**	80.8%	83.8%	74.3%	31.4%	75.8%	67.4%
Scene	**82.1%**	78.1%	81.3%	74.0%	32.4%	75.7%	67.1%

user tasks in immersive virtual reality. Our dataset, the analyses, and the new method advance VR research in several ways.

On our dataset: Existing VR datasets typically only cover a single user task [1–4, 6]. In contrast, our dataset contains human eye and head movements during four common task conditions in immersive virtual reality (Section 3). As such, our dataset paves the way towards a better understanding of visual attention in VR and can be very useful for fostering more research in this field. Furthermore, although in this work we only used our dataset to train a task recognition model, it can also be used to evaluate other data-driven models for immersive virtual reality, such as saliency prediction models [4] or gaze prediction models [1–3, 6]. Our dataset enables researchers to extend existing models that were only trained for one specific task to other task conditions, which will significantly increase the impact of these models as well as their generalization capability to different VR tasks.

On our analyses: We analyzed the patterns of human eye and head movements in immersive virtual reality and revealed significant differences across different tasks (Section 4). Our analyses are significant in that they provide information that are crucial for the development of future VR applications, for example those employing gaze guidance [5] and gaze-contingent rendering [2, 3]. Our analyses also guide future research on the important topic of visual attention analysis in immersive virtual reality.

On our recognition model: Our proposed model achieves a high recognition accuracy in immersive virtual reality and demonstrates strong generalization capabilities for both different users and visual scenes (Table 4). As such, it significantly advances research on the emerging research area of task recognition in VR. The method can also become a crucial component of important VR applications, such as adaptive virtual environment design [7] or low-friction predictive interfaces [16, 20]. In addition, our model also exhibits good recognition performance and strong generalization capability for different users in real-world situations (Table 5). This means our model also has a significant impact on task recognition in real-world situations. Furthermore, our model recognizes user tasks using only the eye and head movements. As such, our model can also be easily extended to other systems like AR and MR systems and can be very useful for fostering new research in these systems.

Limitations: Despite all of these advances, we identified several limitations that we plan to address in future work. First, we only explored the four tasks that are most commonly used in VR applications. However, there exist other VR tasks worth investigating in future work, such as reading or memory tasks. Furthermore, we employed non-interactive 360-degree VR videos instead of interactive 3D virtual environments as our stimuli to ensure that the same VR content was presented to a user under different task conditions. Human visual attention is influenced by both the scene content and the specific tasks assigned to them [1,7]. Since we were interested in the differences between different tasks, employing the same content to collect data avoided the interference from different scene content. However, employing non-interactive VR videos inevitably restricts our analysis to non-interactive VR tasks, neglecting interactive tasks, such as navigation task. The characteristics of human eye and head movements in interactive VR tasks still remain to be explored. Finally, we mainly focused on the differences between different tasks rather than the differences between different stimuli. For our 15 videos, each video was observed by only six users. Our dataset may be insufficient for analyzing the differences between different stimuli.

Future Work: Besides overcoming the above limitations, many potential avenues of future work exist. First, it will be interesting to explore the effectiveness of other factors, such as human body movements and hand movements, in recognizing user tasks. In addition, we can apply our model to other systems besides immersive virtual reality, such as real-world system, augmented reality system, and mixed reality system. Our model only relies on human eye and head movements, ensuring that it can be easily applied to other systems. Furthermore, we are also looking forward to exploring our model's applications in human computer interaction, human-centered computing, and intelligent user interfaces. Finally, recognizing other mental states in immersive VR besides user tasks, such as user cognitive loads [35, 36] and the levels of VR cybersickness [32–34], from human eye and head movements is an interesting avenue of future work.

8 Conclusion

In this work, we focused on understanding human visual attention under different VR tasks and recognizing user tasks in immersive VR. We first presented a dataset of users performing four tasks in immersive VR and showed that the patterns of human eye and head movements are significantly different across different tasks in terms of fixation duration, saccade amplitude, head rotation velocity, and eye-head coordination. Based on these insights, we proposed a novel method to recognize user tasks that outperformed

the state-of-the-art methods both on our dataset and on a real-world dataset by a large margin. As such, our work represents an important advance in understanding human visual attention under different VR tasks and guides future research on task recognition in immersive virtual reality.

ACKNOWLEDGMENTS

This project was supported by the National Key R&D Program of China (No.2017YFB1002700) and the National Natural Science Foundation of China (No.61632003, 61631001, 62172013). A. Bulling was funded by the European Research Council (ERC; grant agreement 801708). We would further like to thank Rakshit Kothari for providing the dataset as well as Wolfgang Fuhl and Antoine Coutrot for providing source code.

REFERENCES

[1] 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*, vol. 27, no. 5, pp. 2681–2690, 2021.

[2] 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*, vol. 25, no. 5, pp. 2002–2010, 2019.

[3] 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*, vol. 26, no. 5, pp. 1902–1911, 2020.

[4] 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*, vol. 24, no. 4, pp. 1633–1642, 2018.

[5] S. Grogorick, M. Stengel, E. Eisemann, and M. Magnor, "Subtle gaze guidance for immersive environments," in *Proceedings of the 2017 ACM Symposium on Applied Perception*, 2017, pp. 1–7.

[6] 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 2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[7] J. Hadnett-Hunter, G. Nicolaou, E. O'Neill, and M. Proulx, "The effect of task on visual attention in interactive virtual environments," *ACM Transactions on Applied Perception*, vol. 16, no. 3, pp. 1–17, 2019.

[8] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior Research Methods*, vol. 50, no. 1, pp. 362–379, 2018.

[9] J. F. Boisvert and N. D. Bruce, "Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features," *Neurocomputing*, vol. 207, pp. 653–668, 2016.

[10] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk, "Predicting cognitive state from eye movements," *PloS One*, vol. 8, no. 5, p. e64937, 2013.

[11] C. Kanan, N. A. Ray, D. N. Bseiso, J. H. Hsiao, and G. W. Cottrell, "Predicting an observer's task using multi-fixation pattern analysis," in *Proceedings of the 2014 ACM Symposium on Eye Tracking Research and Applications*, 2014, pp. 287–290.

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

[13] H. Liao, W. Dong, H. Huang, G. Gartner, and H. Liu, "Inferring user tasks in pedestrian navigation from eye movement data in real-world environments," *International Journal of Geographical Information Science*, vol. 33, no. 4, pp. 739–763, 2019.

[14] M. E. Król and M. Król, "The right look for the job: decoding cognitive processes involved in the task from spatial eye-movement patterns," *Psychological Research*, vol. 84, no. 1, pp. 245–258, 2018.

[15] J. Hild, M. Voit, C. Kühnle, and J. Beyerer, "Predicting observer's task from eye movement patterns during motion image analysis," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications*, 2018, pp. 1–5.

[16] A. Keshava, A. Aumeistere, K. Izdebski, and P. Konig, "Decoding task from oculomotor behavior in virtual reality," in *Proceedings of the 2020 ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[17] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 741–753, 2010.

[18] A. Bulling, C. Weichel, and H. Gellersen, "Eyecontext: recognition of high-level contextual cues from human visual behaviour," in *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems*, 2013, pp. 305–308.

[19] J. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behaviour using topic models," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 75–85.

[20] B. David-John, C. E. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker, "Towards gaze-based prediction of the intent to interact in virtual reality," in *Proceedings of the 2021 ACM Symposium on Eye Tracking Research and Applications*, 2021, pp. 1–7.

[21] L.-M. Vortmann and F. Putze, "Attention-aware brain computer interface to avoid distractions in augmented reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[22] W. Fuhl, N. Castner, T. Kübler, A. Lotz, W. Rosenstiel, and E. Kasneci, "Ferns for area of interest free scanpath classification," in *Proceedings of the 2019 ACM Symposium on Eye Tracking Research and Applications*, 2019, pp. 1–5.

[23] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz, "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities," *Scientific Reports*, vol. 10, no. 1, pp. 1–18, 2020.

[24] Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, and S. Shioiri, "Eye-head coordination for visual cognitive processing," *PloS One*, vol. 10, no. 3, p. e0121035, 2015.

[25] F. Lethaus, M. R. Baumann, F. Köster, and K. Lemmer, "A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data," *Neurocomputing*, vol. 121, pp. 108–130, 2013.

[26] H. Sattar, M. Fritz, and A. Bulling, "Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations," *Neurocomputing*, vol. 387, p. 369–382, 2020.

[27] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5776–5788.

[28] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–9.

[29] X. Wang, A. Ley, S. Koch, D. Lindlbauer, J. Hays, K. Holmqvist, and M. Alexa, "The mental image revealed by gaze tracking," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[30] E. J. David, P. Lebranchu, M. P. Da Silva, and P. Le Callet, "Predicting artificial visual field losses: a gaze-based inference study," *Journal of Vision*, vol. 19, no. 14, pp. 22–22, 2019.

[31] S. Ahn, C. Kelton, A. Balasubramanian, and G. Zelinsky, "Towards predicting reading comprehension from gaze behavior," in *Proceedings of the 2020 ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[32] J. Kim, W. Kim, H. Oh, S. Lee, and S. Lee, "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *Proceedings of the 2019 IEEE International Conference on Computer Vision*, 2019, pp. 10 580–10 589.

[33] M. Shahid Anwar, J. Wang, S. Ahmad, A. Ullah, W. Khan, and Z. Fei, "Evaluating the factors affecting qoe of 360-degree videos and cybersickness levels predictions in virtual reality," *Electronics*, vol. 9, no. 9, p. 1530, 2020.

[34] S. Balasubramanian and R. Soundararajan, "Prediction of discomfort due to egomotion in immersive videos for virtual reality," in *Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality*, 2019, pp. 169–177.

[35] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski, "Estimating cognitive workload in an interactive virtual reality environment using eeg," *Frontiers in Human Neuroscience*, vol. 13, p. 401, 2019.

[36] F. Dell'Agnola, N. Momeni, A. Arza, and D. Atienza, "Cognitive workload monitoring in virtual reality based rescue missions with drones," in *Proceedings of the 2020 International Conference on Human-Computer Interaction*, 2020, pp. 397–409.

[37] J. S. Stahl, "Amplitude of human head movements associated with horizontal saccades," *Experimental Brain Research*, vol. 126, no. 1, pp. 41–54, 1999.

[38] L. Sidenmark and H. Gellersen, "Eye, head and torso coordination during gaze shifts in virtual reality," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 1, pp. 1–40, 2019.
[39] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst, "Pinpointing: Precise head-and eye-based target selection for augmented reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.
[40] L. Sidenmark, D. Mardanbegi, A. R. Gomez, C. Clarke, and H. Gellersen, "Bimodalgaze: Seamlessly refined pointing with gaze and filtered gestural head movement," in *Proceedings of the 2020 ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–9.
[41] J. Gandrud and V. Interrante, "Predicting destination using head orientation and gaze direction during locomotion in vr," in *Proceedings of the 2016 ACM Symposium on Applied Perception*, 2016, pp. 31–38.
[42] L. Sidenmark and H. Gellersen, "Eye&head: Synergetic eye and head movement for gaze pointing and selection," in *Proceedings of the 2019 ACM Symposium on User Interface Software and Technology*, 2019, pp. 1161–1174.
[43] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "360 video viewing dataset in head-mounted virtual reality," in *Proceedings of the 2017 ACM on Multimedia Systems Conference*, 2017, pp. 211–216.
[44] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proceedings of the 2017 ACM on Multimedia Systems Conference*, 2017, pp. 199–204.
[45] C. Wu, Z. Tan, Z. Wang, and S. Yang, "A dataset for exploring user behaviors in vr spherical video streaming," in *Proceedings of the 2017 ACM on Multimedia Systems Conference*, 2017, pp. 193–198.
[46] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *Journal of Vision*, vol. 14, no. 3, pp. 14–14, 2014.
[47] Y. Sugano, Y. Ozaki, H. Kasai, K. Ogaki, and Y. Sato, "Image preference estimation with a data-driven approach: A comparative study between gaze and image features," *Journal of Eye Movement Research*, vol. 7, no. 3, 2014.
[48] L. Sidenmark, C. Clarke, X. Zhang, J. Phu, and H. Gellersen, "Outline pursuits: Gaze-assisted selection of occluded objects in virtual reality," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.
[49] M. R. Greene, T. Liu, and J. M. Wolfe, "Reconsidering yarbus: A failure to predict observers' task from eye movement patterns," *Vision Research*, vol. 62, pp. 1–8, 2012.
[50] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of Vision*, vol. 14, no. 3, pp. 29–29, 2014.
[51] T. C. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci, "Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies," *Behavior Research Methods*, vol. 49, no. 3, pp. 1048–1064, 2017.
[52] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 ACM Symposium on Eye Tracking Research and Applications*, 2000, pp. 71–78.
[53] M. I. Coco and F. Keller, "Classification of visual and linguistic tasks using eye-movement features," *Journal of Vision*, vol. 14, no. 3, pp. 11–11, 2014.
[54] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.
[55] Z. Hu, S. Li, and M. Gai, "Temporal continuity of visual attention for future gaze prediction in immersive virtual reality," *Virtual Reality and Intelligent Hardware*, vol. 2, no. 2, pp. 142–152, 2020.
[56] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, 2014.

Zhiming Hu is a Ph.D. candidate at Graphics & Interactive Technology Lab., School of Computer Science, Peking University. His research interest includes virtual reality, visual attention, human-computer interaction, and eye tracking.

Andreas Bulling is Full Professor of Human-Computer Interaction and Cognitive Systems at the University of Stuttgart, Germany. Before, he was a Feodor Lynen and Marie Curie Research Fellow at the University of Cambridge, UK, Senior Researcher at the Max Planck Institute for Informatics, and an Independent Research Group Leader at Saarland University, Germany. His research interests include computer vision, machine learning, and human-computer interaction.

Sheng Li is currently an Associate Professor of the School of Computer Science, Peking University. He works as a member of the Graphics & Interactive Technology Lab. and has published over 40 refereed papers in prestigious journals and conferences, including ACM TOG, IEEE TVCG, CGF, etc. His research interests include virtual reality, rendering, physical simulation and animation. He is a member of ACM and IEEE.

Guoping Wang is currently a professor in Peking University, where he is also the director of Graphics & Interactive Technology Laboratory. He achieved the National Science Fund for Distinguished Young Scholars in 2009. His research interests include Virtual Reality, Computer Graphics, Human-Computer Interaction, and Multimedia.