Saliency in VR: How do people explore virtual environments?

Vincent Sitzmann*, Ana Serrano*, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, Gordon Wetzstein

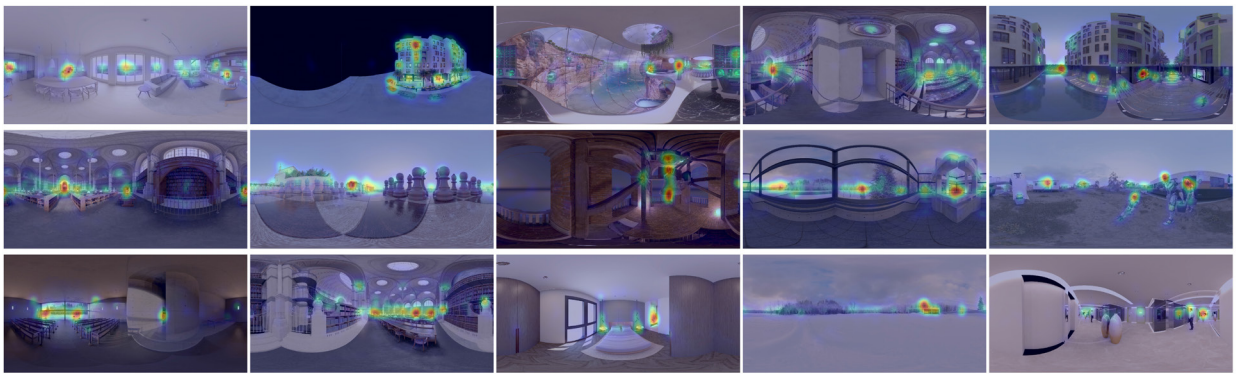

Fig. 1. A representative subset of the 22 panoramas used to analyze how people explore virtual environments from a fixed viewpoint. We recorded almost two thousand scanpaths of users exploring these scenes in different immersive and non-immersive viewing conditions. We then analyzed this data, and provide meaningful insights about viewers' behavior. We apply these insights to VR applications, such as saliency prediction (shown in the image as overlaid heatmaps), VR movie editing, panorama thumbnail generation, panorama video synopsis, and saliency-aware compression of VR content.

Abstract—Understanding how people explore immersive virtual environments is crucial for many applications, such as designing virtual reality (VR) content, developing new compression algorithms, or learning computational models of saliency or visual attention. Whereas a body of recent work has focused on modeling saliency in desktop viewing conditions, VR is very different from these conditions in that viewing behavior is governed by stereoscopic vision and by the complex interaction of head orientation, gaze, and other kinematic constraints. To further our understanding of viewing behavior and saliency in VR, we capture and analyze gaze and head orientation data of 169 users exploring stereoscopic, static omni-directional panoramas, for a total of 1980 head and gaze trajectories for three different viewing conditions. We provide a thorough analysis of our data, which leads to several important insights, such as the existence of a particular fixation bias, which we then use to adapt existing saliency predictors to immersive VR conditions. In addition, we explore other applications of our data and analysis, including automatic alignment of VR video cuts, panorama thumbnails, panorama video synopsis, and saliency-based compression.

Index Terms—Saliency, omnidirectional stereoscopic panoramas

1 Introduction

Virtual reality (VR) systems provide a new medium that has the potential to have a significant impact on our society. The experiences offered by these emerging systems are inherently different from radio, television, or theater, opening new directions in research areas such as cinematic VR capture [1], interaction [53], or content generation and editing [39, 49]. However, the behavior of users who visually explore immersive VR environments is not well understood, nor do statistical models exist to predict this behavior. Yet, with unprecedented capabilities for creating synthetic immersive environments, many important questions arise. How do we design 3D scenes or place cuts in VR videos? How do we drive user attention in virtual environments? Can we predict visual exploration patterns? How can we efficiently compress cinematic VR content?

To address these and other questions from first principles, it is crucial to understand how users explore virtual environments. In this work, we take steps towards this goal. In particular, we are interested in quantifying aspects of user behavior that may be helpful in predicting exploratory user behavior in static and dynamic virtual environments observed from a fixed viewpoint. A detailed understanding of visual attention in VR would not only help answer the above questions, but also inform future designs of user interfaces, eye tracking technology, and other key aspects of VR systems.

A crucial requirement for developing an understanding of viewing behavior in VR is access to behavioral data. To this end, we have performed an extensive study, recording 1980 head and gaze trajectories from 169 people in 22 static virtual environments, which are represented as stereoscopic omni-directional panoramas. Data is recorded using a head-mounted display (HMD) in both standing and seated conditions (*VR* condition and *VR seated* condition), as well as for users observing the same scenes in mono on a desktop monitor for comparison (*desktop* condition).

We analyze the recorded data and discuss important insights, such as the existence of a fixation bias, the mean time until a static stereo panorama can be considered to be fully explored by users, or the existence of two apparent modes in viewer behavior, attention and re-orientation (see Sec. 4 for more details). We then leverage our data to evaluate *existing* saliency predictors, designed for narrow field of view video, in the context of immersive VR, and show how these can be adapted to VR applications. Saliency prediction is a well-explored topic and many existing models are evaluated by the MIT Saliency

- *These authors contributed equally.
- *Correspondence to: sitzmann@cs.stanford.edu and gordon.wetzstein@stanford.edu.*
- *Vincent Sitzmann, Maneesh Agrawala, and Gordon Wetzstein are with Stanford University.*
- *Ana Serrano, Diego Gutierrez, and Belen Masia are with the Universidad de Zaragoza.*
- *Amy Pavel is with the University of California Berkeley.*

Manuscript received 11 Sept. 2017; accepted 8 Jan. 2018.
Date of publication 19 Jan. 2018; date of current version 18 Mar. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2793599

Benchmark [5]. However, these models assume that users sit in front of a screen while observing the images – ground truth data is collected by eye trackers recording precisely this behavior. VR is different from traditional 2D viewing in that users naturally use both significant head movement and gaze to visually explore scenes. We show that this leads to a fixation bias around the equator that is not observed in conventional viewing conditions. Figure 1 shows panoramic views of some of our 22 scenes with superimposed saliency computed from the recorded scan paths in the VR condition. Apart from saliency, we offer several other example applications that are directly derived from our findings. Specifically, our contributions are:

- We record and provide an extensive dataset of visual exploration behavior in stereoscopic, static omni-directional stereo (ODS) panoramas. The dataset contains head orientation and gaze direction, and it captures several different viewing conditions. Scenes, data, and code for analysis (Sec. 3) are available online[1]
- We provide low-level and high-level analysis of the recorded dataset. We derive relevant insights that can be crucial for predicting saliency in VR and other VR applications, such as the existence of an attention bias in VR scenes or differences in head and gaze movement statistics when fixating (Sec. 4)
- We evaluate existing saliency predictors with respect to their performance in VR applications. We show how to tailor these predictors to ODS panoramas and demonstrate that saliency prediction from head movement alone performs on par with state-of-the-art saliency predictors for our scenes (Sec. 5)
- We demonstrate several applications of this saliency prediction, including automatic panorama thumbnails, VR video synopsis, compression, and VR video cuts (Sec. 6)

2 Related Work

Modeling human gaze behavior and predicting visual attention has been an active area of vision research. In their seminal work, Koch and Ullman [27] introduced a model for predicting salient regions from a set of image features. Motivated by this work, many models of visual attention have been proposed throughout the last three decades. Most of these models are based on bottom-up, top-down, or hybrid approaches. Bottom-up approaches build on a combination of low-level image features, including color, contrast, or orientation [8, 21, 26, 36] (see Zhao and Koch [62] for a review). Top-down models take higher-level knowledge of the scene into account such as context or specific tasks [16, 22, 25, 35, 55]. Recently, advances in machine learning and particularly convolutional neuronal networks (CNNs) have fostered the convergence of top-down and bottom-up features for saliency prediction, producing more accurate models [20, 34, 42, 59, 63]. Jiang et al. [24] proposed a new methodology to collect attentional data on scales sufficient for these deep learning methods. Volokitin et al. [58] used features learned by CNNs to predict when saliency maps predicted by a model will be accurate and when fixations will be consistent among human observers. Significant prior work explored rigorous benchmarking of saliency models, the impact of the metric on the evaluation result, and shortcomings of state-of-the-art models at the time [4, 6, 45]. Recent work also attempts to extend CNN approaches beyond classical 2D images by computing saliency in more complex scenarios such as stereo images [9, 17] or video [7, 33]. A related line of research is devoted to modeling the gaze scanpath followed by subjects, i.e., the temporal evolution of the viewer's gaze [23, 32]. Marmitt et al. [37] developed a metric to evaluate predicted scanpaths in VR and showed that predictors built for classic viewing conditions perform significantly worse in VR. Building on the rich literature in this area, we explore user behavior and visual attention in immersive virtual environments, which can help build similar models for VR.

What makes VR different from desktop viewing conditions is the fact that head orientation is used as a natural interface to control perspective (and in some cases navigation as well [56]). The interactions of head and eye movements are complex and neurally coupled, for example via the vestibulo-ocular reflex [30]. Koehler et al. [28] showed that saliency maps can differ depending on the instructions given to the viewer. For more information on user behavior in VR, we refer to Ruhland et al. [46], who provide a review of eye gaze behavior, and Freedman [13], who discusses the mechanisms that characterize the coordination between eyes and head during visual orienting movements. With the data recorded in this project, we observe the vestibulo-ocular reflex and other interesting effects. In the paper and supplemental material, we provide an extensive analysis of the user data, and derive statistics describing many low-level aspects of viewing behavior. We hope that this analysis will be useful for basic vision research.

Recent work of Nakashima et al. [38] is closely related to some aspects of our work. They propose a head direction prior to improve accuracy in saliency-based gaze prediction through simple multiplication of the gaze saliency map by a Gaussian head direction bias. Concurrent work by Upenik et al. [57] explores visual attention in VR solely by tracking head orientation. The data collected in this paper and in-depth analyses augment prior work in this field, and may allow for future data-driven models for visual behavior to be learned.

Finally, gaze tracking has found many applications in VR user interfaces [54] and gaze-contingent displays [12, 41, 50]. The ability to predict viewing behavior would be helpful for all of these applications. For example, gaze-contingent techniques may become possible without dedicated gaze trackers, which are currently expensive and not widely available. Moreover, techniques for editing VR content are starting to emerge [39, 49]. The understanding of user behavior we aim to develop in this paper could also influence these and other tools for content creation.

3 Recording Head Orientation and Gaze

In this section, we summarize our efforts towards recording a dataset that contains head orientation and gaze direction for users watching stereoscopic VR panoramas in several different viewing conditions; we provide additional details in the supplemental material. These data form the basis of a statistical analysis of viewing behavior (Sec. 4), as ground truth for saliency prediction (Sec. 5), and also as reference saliency for several higher-level applications (Sec. 6).

3.1 Data capture

Stimuli For the experiments reported in this paper, we used 22 high-resolution omni-directional stereo panoramas (see Figure 1 and supplemental material). We opt for a fixed viewpoint because for the subsequent analyses it is crucial that subjects see the exact same content. Further, in a 3D scenario the variability is likely to be much higher, requiring extremely large numbers of subjects to draw significant conclusions. The scenes include (14) indoor and (8) outdoor scenarios and do not contain landmarks that may be recognized by the users. For each scene we explore different conditions, which limits the number of scenes we can have with the experiment size remaining tractable. With the current stimuli and conditions, we have collected nearly 2,000 trajectories from 169 viewers. All scenes are computer generated by artists; we received permission to use them for this study.

Conditions We recorded users observing the 22 panoramas in three different conditions: in a standing position using a head-mounted display (i.e., the *VR* condition), seated in a non-swivel chair using a head-mounted display (i.e., the *VR seated* condition, making it more difficult to turn around), and seated in front of a desktop monitor (i.e., the *desktop* condition). In the *desktop* condition, the scenes are monoscopic, and users navigate with a mouse. For each scene, we tested four different starting points, spaced at 90° longitude, which results in a total of 264 conditions. These starting points were chosen to cover the entire longitudinal range, while keeping the number of different conditions tractable. We chose not to randomize the starting point over the whole latitude (and rather select randomly from four fixed ones) to limit the number of conditions while being able to analyze the influence of the starting point (Sec. 4.5 and supplement). Complete

[1] https://vsitzmann.github.io/vr-saliency

randomization over the starting point could be of interest for future studies.

Participants For the *VR* condition, we recorded 122 users (92 male, 30 female, age 17-59). The experiments with the *VR seated* condition were performed by 47 users (38 male, 9 female, age 17-39). Users were asked to first perform a stereo vision (Randot) test to quantify their stereo acuity. For *desktop* experiments, we recruited 44 additional participants (27 male, 17 female, age 18-33). All participants reported normal or corrected-to-normal vision.

Procedure All VR scenes were displayed using an Oculus DK2 head-mounted display, equipped with a pupil-labs[2] stereoscopic eye tracker recording at 120 Hz. The DK2 offers a field of view of $95 \times 106°$. The Unity game engine was used to display all scenes and record head orientation while the eye tracker collected gaze data on a separate computer. Users were instructed to freely explore the scene and were provided with a pair of earmuffs to avoid auditory interference. Scenes and starting points were randomized, while ensuring that each user would only see the same scene once from a single random starting point. Each user was shown 8 scenes. Each scene in a certain condition was shown to the user during 30 seconds, while the total time per user that the experiment took, including calibration and explanation, was approximately 10 minutes.

We modeled the *desktop* condition after typical, mouse-controlled desktop panorama viewers on the web (i.e., YouTubeVR or Facebook360). Users sat 0.45 meters away from a 17.3" monitor with a resolution of 1920×1080 px, covering a field of view of $23 \times 13°$. We used a Tobii EyeX eye tracker with an accuracy of 0.6° at a sampling frequency of 55 Hz [15]. The image viewer displayed a rectilinear projection of a $97 \times 65°$ viewport of the panorama. To keep the field of view consistent, no zooming was possible. We instructed the users on how to use the image viewer, before showing the 22 scenes for 30 seconds each. In this condition, we only collected gaze data since users rarely re-orient their head. Instead, we recorded where the users interactively place the virtual camera in the panorama as a proxy for head orientation.

3.2 Data processing

To identify fixations, we transformed the normalized gaze tracker coordinates to latitude and longitude in the 360° panorama. This is necessary to detect users fixating on panorama features while turning their head. We used thresholding based on dispersion and duration of the fixations [47]. For the VR experiments, we set the minimum duration to 150 ms [47] and the maximum dispersion to 1° [2]. For the desktop condition, we found the Tobii EyeX eyetracker to be more noisy than the PupilLabs eyetracker. Thus, we first smoothed this data with a running average of 2 samples, and detected fixations with a dispersion of 2°. We counted the number of fixations at each pixel location in the panorama. Similar to Judd et al. [25], we only consider measurements from the moment where user's gaze left the initial starting point to avoid adding trivial information. We convolved these fixation maps with a Gaussian with a standard deviation of 1° of visual angle, the area of the field of view seen sharply on the fovea of the user, to yield continuous saliency maps [31].

4 Understanding Viewing Behavior in VR

With the recorded data, we can gather insights and investigate a number of questions about the behavior of users exploring virtual environments. In the following, we analyze both low-level characteristics, such as duration of the fixations and speed of gaze, and higher-level characteristics, such as the influence of the content or characteristics of the scene.

4.1 Is viewing behavior similar between users?

We first want to assess whether viewing behavior between users is similar; this is also indicative of how robust our data is, and thus how much we can rely on it to draw conclusions. To answer this, we analyze the agreement between users. Specifically, we compute the *inter-observer congruency* metric by means of a *receiver operating characteristic* curve (ROC) [31,55]. This metric calculates the ability of the i^{th} user to predict a *ground truth saliency map*, which is computed from the fixations of all the other users averaged. A single point in the ROC curve is computed by finding the top $n\%$ most salient regions of the ground truth saliency map (leaving out the i^{th} user), and then calculating the percentage of fixations of the i^{th} user that fall into these regions. We show the average ROC for all the 22 scenes in Figure 2 (left), compared with chance (the individual ROCs for each scene are depicted in light gray). The fast convergence of these curves to the maximum rate of 1 indicates a strong agreement, and thus similar behavior, between users for each of the scenes tested. 70% of all fixations fall within the 20% most salient regions. These values are comparable to previous studies viewing regular images on a display [31].

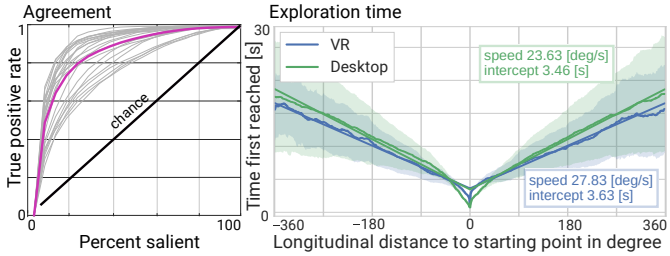

Fig. 2. *Left:* ROC curve of human performance averaged across users (magenta) and individual ROCs for each scene (light gray). The fast convergence to the maximum is indicative of a strong agreement between users. *Right:* Exploration time computed as the average time until a specific longitudinal offset from the starting point is reached.

4.2 How different is viewing behavior for the 3 conditions?

An important question to ask is whether viewing behavior changes when exploring a scene under different conditions. Visual inspection of our three conditions (*VR*, *VR seated*, and *desktop*) shows a high similarity between the saliency maps (see supplement). For a quantitative evaluation of the similarity of saliency maps (here, and in the rest of the paper), we use the Pearson correlation (CC) score, which is a widely used metric in saliency map prediction [6]. It ranges from -1 (perfectly inversely correlated) to 1 (perfectly correlated). The high similarity is confirmed by a median CC score of 0.80 when comparing the *VR* and the *VR seated* conditions, and 0.76 when comparing the *VR* and the *desktop* conditions. The latter is a significant insight: since desktop experiments are much easier to control, it may be possible to use these for collecting adequate training sets for data-driven saliency prediction in future VR systems. Given this similarity, we report only the results of the *VR* (standing) condition throughout the remainder of the paper, unless a significant difference is found, and refer the reader to the supplemental for the *VR seated* and *desktop* conditions.

4.3 Is there a fixation bias in VR?

Several researchers have reported a strong bias for human fixations to be near the center, when viewing regular images [25, 40]. A natural question to ask is whether a similar bias exists in VR. Similar to Judd et al. [25], we calculate the average of all 22 saliency maps, and filter out fixations within the close vicinity (20° longitude) of the starting point. The resulting data indicates that users tend to fixate around the equator of the panoramas, with very few fixations in latitudes far from it. To quantify this *equator bias*, we marginalize out the longitudinal component of the saliency map, and fit a Laplace distribution—with location parameter μ and diversity β—to the latitudinal component (this particular distribution yielded the best match among several tested distributions). Figure 3 depicts the average saliency map, as well as our Laplacian fit to the latitudinal distribution and its parameters, for both the *VR* and the *desktop* conditions. While the mean is almost

[2]https://pupil-labs.com

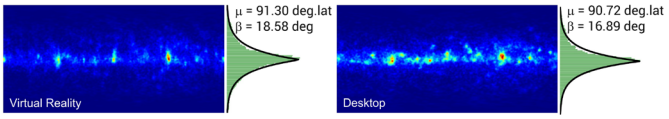

Fig. 3. Average saliency maps computed with all the scenes for both the *VR* (left) and the *desktop* (right) conditions. These average maps demonstrate an "equator bias" that is well-described by a Laplacian fit modeling the probability of a user fixating on an object at a specific latitude.

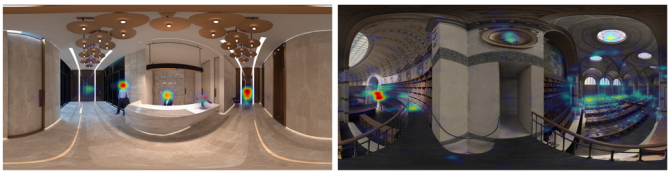

Fig. 4. Saliency maps presenting the lowest (*left*) and highest (*right*) entropy in our dataset. Saliency maps with low entropy have very defined salient regions while in maps with high entropy fixations are scattered all over the scene.

identical, the equator bias for the desktop condition has a lower diversity. As discussed in Section 5, this Laplacian equator bias is crucial for predicting saliency in VR.

Note that most of the scenes in our study have a clear horizon line, which may have influenced the observed equator bias along with viewing preferences, kinematic constraints, as well as the static nature of the scenes. However, most virtual environments share this type of scene layout, so we believe our findings generalize to a significant fraction of this type of content. Further, even for scenes with content scattered along different latitudes (see, e.g., *scene 16* in Fig. 12 of the supplement, displaying very few salient areas near the poles), we observed an equator bias. Nevertheless, different tasks or scenarios, such as gaming, may influence this bias.

4.4 Does scene content affect viewing behavior?

A fundamental issue when analyzing viewing behavior is the potential influence of scene content. This is of particular relevance for content creators; since in a VR setup the viewer has control over the camera, this analysis can help address the key challenge of predicting user attention.

To characterize scene content in a manner that enables insightful analysis, we rely on the distribution of salient regions in the scene, in particular on the *entropy* of the saliency maps. A high entropy results from a large number of similarly salient objects distributed throughout the scene, causing users' fixations to be scattered all over the scene; a low entropy results from a few salient objects that capture all the viewer's attention. Figure 4 shows the saliency maps of the scenes with lowest and highest entropy in our dataset.

Our entropy is computed as the Shannon entropy of the *ground truth saliency map*, computed, per scene, from the average of all users [25]. The entropy is given by: $-\sum_{i=1}^{N} s_i^2 log(s_i^2)$, with s being the ground truth saliency, and N the number of pixels. We consider two entropy levels, low and high, which we term $\{E_0, E_1\}$, respectively. Since a clear threshold for classifying each scene according to its entropy does not exist, we take a conservative approach and analyze only the four scenes with highest and the four with lowest entropy, for a total of eight scenes.

4.4.1 Viewing behavior metrics

Measuring viewing behavior in an objective manner is not a simple task. First, we define *salient regions* as the 5% most salient pixels of a scene. Figure 5 shows a saliency map and the resulting salient regions computed with this criterion. We then rely on three metrics recently proposed by Serrano et al. [49] in the context of gaze analysis for VR movie editing (time to reach a salient region (*timeToSR*), percentage

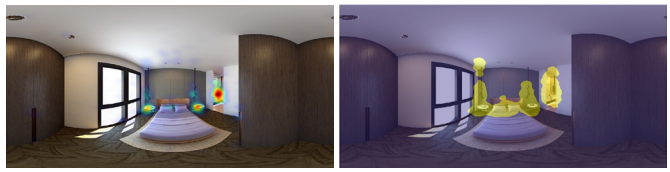

Fig. 5. Salient region computation. *Left:* Ground truth saliency map for a sample scene. *Right:* Corresponding salient regions (yellow) computed by thresholding the 5% most salient pixels of the scene.

of fixations inside the salient regions (*percFixInside*), and number of fixations (*nFix*), which are summarized in the supplemental material), and propose a fourth, novel one, tailored (but not limited) to quantifying the degree of exploration over time in static 360° panoramas:

Convergence time (*convergTime*) For every scene, we obtain the per-user saliency maps at different time steps, and compute the similarity (CC score) with the fully-converged saliency map. We plot the temporal evolution of this CC score, and compute the area under this curve. This metric represents the temporal convergence of saliency maps; it is inversely proportional to how long it takes for the fixation map during exploration to converge to the ground truth saliency map.

4.4.2 Analysis

We first test for independence of observations performing a Wald's test (please refer to the supplement). Based on its results, we employ ANOVA when analyzing *percFixInside*, since the samples are considered to be independent, and report significance values obtained from multilevel modeling for the other three metrics.

We find that the *entropy* of the scene has a significant effect on *nFix* ($p < 0.001$), *timeToSR* ($p < 0.001$), *percFixInside* ($p = 0.022$), and *convergTime* ($p < 0.001$). Specifically, on scenes with low entropy (E_0), the time to reach a salient region (*timeToSR*) is lower. This may be counter-intuitive, since high entropy scenes contain a larger number of salient regions and thus it would be easier to reach one. Interestingly, our results indicate that the viewer explores the scene faster in cases of low entropy, quickly discarding non-salient regions, and that their attention gets directed towards the few salient regions faster. This hypothesis is further supported by the behavior of the *convergTime* metric, which shows that scenes with low entropy do converge faster, and is consistent with the number of fixations, and fixations inside the salient region (*nFix* and *percFixInside*): both are higher for low entropy scenes, indicating that users pay more attention to salient regions when such regions are less, and more concentrated.

4.5 Does the starting point affect viewing behavior?

We also evaluate whether the starting viewport conditions the final saliency map for a given scene: For each scene, we compute the similarity between the final saliency map of the i^{th} viewport and the other three, using again the CC score. We obtain a median CC score of 0.79, which indicates that the final saliency maps after 30 seconds, starting from different viewports, converge and are very similar. Additional analysis on the influence of the viewport, including also a state sequence analysis [14, 49], can be found in the supplement.

4.6 How are head and gaze statistics related?

Many additional insights can be learned from our data, which may be useful for further vision and cognition research, or in applications that require predicting gaze or saliency in VR (see also Section 5). First, we evaluate the speed with which users explore a given scene. Figure 2 (right) shows this *exploration time*, which is the average time that users took to move their eyes to a certain longitude relative to their starting point. On average, users fully explored each scene after about 19 seconds. Indeed, after this time, all saliency maps in our dataset have converged to a CC score of at least 0.8 as compared to their final state. These results suggests that an experimental time of 20 seconds is sufficient to capture fixations in static stereo panoramas.

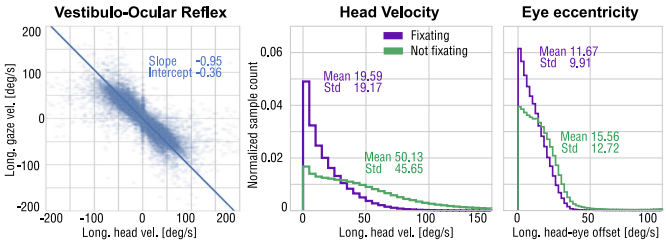

Fig. 6. *Left:* the vestibulo-ocular reflex demonstrated by an inverse linear relationship of gaze and head velocities. *Middle and right:* distributions of longitudinal head velocity and longitudinal eye eccentricity, respectively, while fixating and while not fixating.

	Equirectangular	Cube Map	Patch Based
Without Equator Bias	μ =0.48	μ = 0.37	μ =0.43
With Equator Bias	μ =**0.50**	μ =0.44	μ =0.49

Table 1. Quantitative evaluation of three different projection methods with and without equator bias. We list the mean CC score for all 22 VR scenes used in this study. Applying the equator bias significantly improves the quality of all approaches. Distortions of the equirectangular projection near the poles do not affect saliency prediction as much as the shortcomings of other types of projection after the equator bias is applied.

In our experiments, the mean number of fixations across scenes is 55.35 $\pm$ 12.85 for the VR condition and 49.68 $\pm$ 15.04 for the desktop condition. The mean duration of fixations in the *VR* condition is 266 ms $\pm$ 132, and 253 ms $\pm$ 124 in the *desktop* condition. This is in the range reported for traditional screen viewing conditions [47]. The mean gaze direction relative to the head orientation across scenes is 13.85° $\pm$ 11.73, which is consistent with the analysis performed by Kollenberg et al. [29].

We have also identified the *vestibulo-ocular reflex* [30] in our data. This reflexive mechanism moves the eyes contrary to the head movement, in order to stabilize the line of sight and thus improve vision quality. Figure 6 (left) shows the expected inverse linear relationship between head velocity and relative gaze velocity when fixating. Given this observation, we further analyze the interaction between eye and head movements when shifting to a new target. We offset in time head and gaze acceleration measurements relative to each other, and compute the cross-correlation for different temporal shifts. Our data reveals that head follows gaze with an average delay of 58 ms, where the largest cross-correlation is observed, which is consistent with previous work [11, 13].

It is well-known that gaze velocities differ when users fixate and when they do not [47]. We look at whether this is also the case for head velocities, since they could then act as a rough proxy for fixation classification. Figure 6 (middle) shows that users move their head at longitudinal velocities significantly below the average head speed when they are fixating, and above average when they are not. Further, Figure 6 (right) shows that the longitudinal rotation angle of the eyes relative to the head orientation (eye eccentricity) is significantly smaller when users are fixating. According to this data, users appear to behave in two different modes: *attention* and *re-orientation*. Eye fixations happen in the attention mode, when users have "locked in" on a salient part of the scene, while movements to new salient regions happen in the re-orientation mode. Being able to identify such modes in real time, from either head or gaze movement, can be very useful for interactive applications. Further results for the different conditions, and for the latitudinal direction, can be found in the supplement. Finally, this data and findings can be leveraged for *time-dependent* and *head-based* saliency prediction, as we will show in Sections 5.2 and 5.3.

5 Predicting Saliency in VR

In this section, we show how existing saliency prediction models can be adapted to VR using insights of our data analysis, such as the equator bias. Then, we ask whether the problem of time-dependent saliency prediction is a well-defined one that can be answered with sufficient confidence. Finally, we analyze how well head movement alone, for example captured with inertial sensors, can predict saliency without knowing the exact gaze direction.

5.1 Predicting saliency maps

Instead of learning VR saliency models from scratch, we ask whether existing models could be adopted to immersive applications. This would be ideal, because many saliency predictors for desktop viewing conditions already exist, and advances in that domain could be directly transferred to VR conditions. The fact that gaze statistics are closely related in VR and in traditional viewing (Section 4.6) is indicative of the fact that existing saliency models may be adequate, at least to some extend, to VR. In this context, two primary challenges arise: (i) mapping a 360° panorama to a 2D image (the required input for existing models) distorts the content due to the projective mapping from sphere to plane; and (ii) head-gaze interaction may require special attention for saliency prediction in VR. We address both of these issues in the following.

Which projection is best?

Before running a conventional saliency predictor on a spherical panorama or parts of it, the image has to be projected into a plane. Different projections would naturally result in different types of distortions that may affect the saliency predictor. For an equirectangular projection, for example, we expect large distortions near the poles. A cube map projection may result in discontinuities between some of the cube's faces. Alternatively, smaller patches can be extracted from the panorama, saliency prediction applied to each of them projected onto a plane, and the result stitched together and blended into a saliency panorama. The latter, patch-based approach would result in the least amount of geometrical distortions, but it is also the most computationally expensive approach and it gives up global context for the saliency prediction.

In Figure 8 and Table 1 we compare saliency prediction using all three projection methods qualitatively and quantitatively. For each projection, we compute a saliency map using the state-of-the-art ML-Net saliency predictor [10], and then optionally multiply it by the latitudinal equator bias we derived in Section 4.3. We incorporate the equator bias in a multiplicative manner. This only increases the weight of areas that the saliency predictor has found to be potentially salient, while an additive bias would increase saliency of all points around the equator. Alternatively, it could also be incorporated by addition and re-normalization. Figure 8 shows an example saliency map predicted on the three different sphere projections after applying the equator bias. We also compare the average CC score for all three projection methods and all 22 scenes in Table 1. Quantitatively, saliency computed directly on the equirectangular projection with the equator bias applied not only performs best but it is also the fastest of the three approaches. The benefit of applying the equator bias is smaller for the equirectangular projection than for the other two projections. This may be because the distortions at the poles introduced by the projection may naturally lead to less saliency predicted at the poles than in the cube map and patch-based approaches. While this seems to make this prediction method competitive even without the equator bias, it may lead to inferior generalization as compared to an explicit modeling. Since the equirectangular and patch-based methods using the equator bias perform almost on par, in the following, we use the patch-based method when processing time is not critical, since it is not susceptible to projective distortions.

Which predictor is best?

The fact that existing saliency predictors seem to apply to VR scenarios is important, because rapid progress is being made for saliency prediction with images and videos. Advances in those domains could directly improve saliency prediction in VR. Here, we further evaluate several different existing predictors both quantitatively and qualitatively.

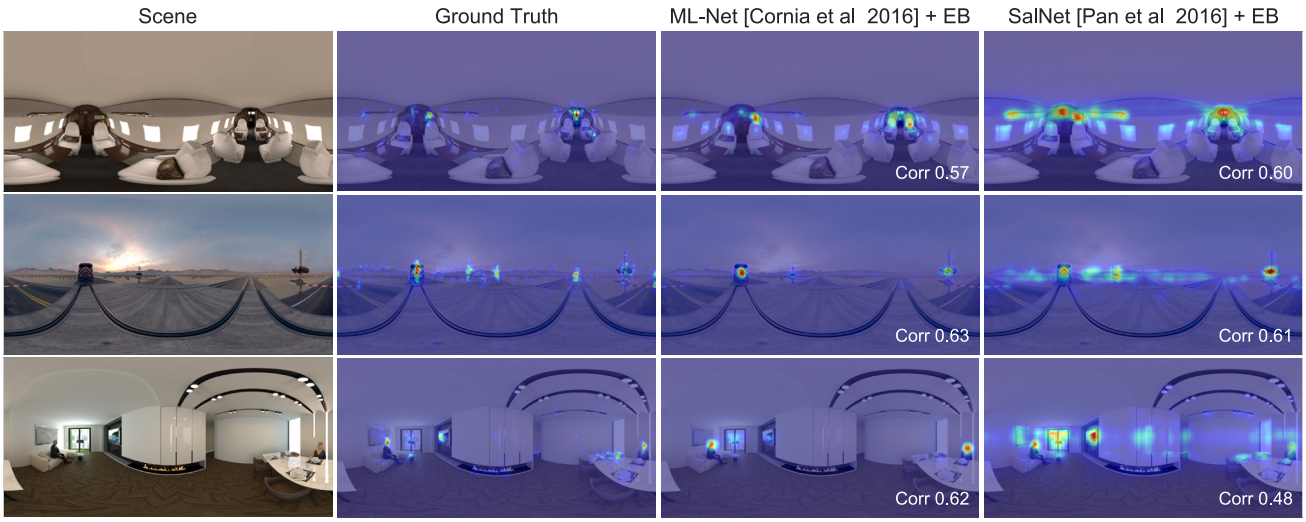

Fig. 7. Saliency prediction for omni-directional stereo panoramas. Existing saliency predictors can be applied to spherical panoramas after they are projected onto a plane, here performed with the patch-based method described in the text. These methods tend to over-predict saliency near the poles. By multiplying the predicted saliency map by the longitudinal equator bias (EB) derived in the previous section, we achieve a good match between ground truth (center left) and predicted saliency (right). Note that this procedure could be applied to any saliency predictor; we chose two top-scoring predictors as an example.

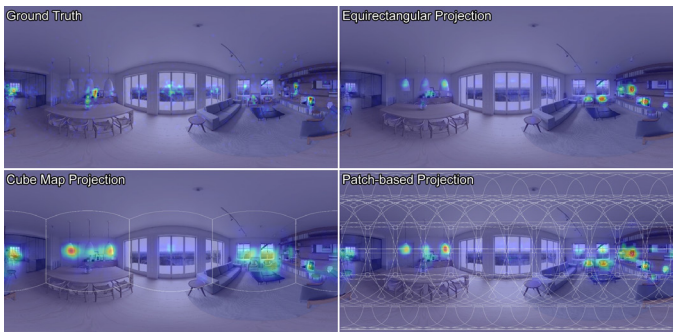

Fig. 8. Comparison of saliency prediction using different projections from sphere to plane. After applying the equator bias, all three projection methods result in comparable saliency maps for this example.

	EB	ML-Net + EB	SalNet + EB
VR	$\mu = 0.34 \pm 0.13$	$\mu = \mathbf{0.49} \pm 0.11$	$\mu = 0.47 \pm 0.13$
Desktop	$\mu = 0.37 \pm 0.11$	$\mu = \mathbf{0.57} \pm 0.11$	$\mu = 0.52 \pm 0.12$

Table 2. Quantitative comparison of predicted saliency maps using a simple equator bias (EB), and two state-of-the-art models together with the EB. Numbers show average mean and standard deviation of CC scores, for each scene, between prediction and ground truth recorded from users exploring 22 scenes in the VR and desktop conditions. The proposed patch-based method was used to predict the saliency maps for both predictors.

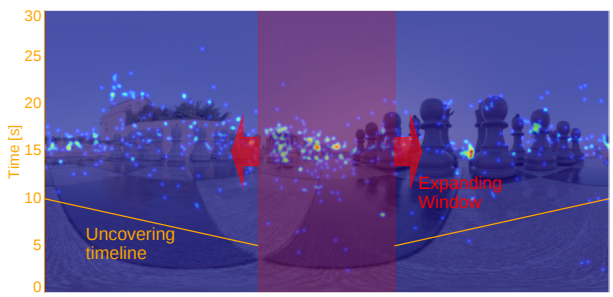

Fig. 9. Time-dependent saliency prediction by uncovering the converged saliency map with the average exploration speed determined in Section 4.

Table 2 lists mean and standard deviation of the CC score for all 22 scenes in the VR condition, and for users exploring the same scenes in the desktop condition. These numbers allow us to analyze how good and how consistent across scenes a particular predictor is. We test the equator bias by itself as a baseline, as well as two of the highest-ranked models in the MIT benchmark where source code is available: ML-Net [10] and SalNet [42], together with the equator bias. We see that the two advanced models perform very similar and do much better than the equator bias alone. We also see that both of these models predict viewing behavior in the desktop condition better than for the VR condition. This makes sense, because the desktop condition is what these models were trained for originally. In Figure 7 we also compare qualitatively the saliency maps of three scenes recorded under the VR condition (all scenes in the supplement).

5.2 Can time-dependent saliency be predicted with sufficient confidence?

Virtual environments impose viewing conditions much different from those of conventional saliency prediction. Specifically, the question of temporal evolution arises: for users starting to explore the scene at a given starting point, is it possible to predict the probability that they fixate at specific coordinates at a time instant t? This problem is also closely related to scanpath prediction. We use data from Section 4 to build a simple baseline model for this problem: Figure 2 (right) shows an estimate for when users reach a certain longitude on average. We can thus model the time-dependent saliency map of a scene with an initially small window that grows larger over time to progressively uncover more of a converged (predicted or ground truth) saliency map. The part of the saliency map within this window is the currently active part, while the parts outside this window are set to zero. The left and right boundaries of the window are widened with the speed predicted in Figure 2 (right).

Figure 9 visualizes this approach. We generate the time-dependent saliency maps from the converged ground truth maps for all 22 scenes and compare them with the actual ground truth at each timestep. We use the fully-converged saliency map as a baseline. The time-dependent, constructed saliency maps model the recorded data better than the converged saliency map within the first 6 seconds. Subsequently, they perform slightly worse until the converged map is fully uncovered after

about 10 seconds, and the model is thus identical to the baseline. Our simple time-dependent model achieves an average CC score of 0.57 over all scenes, viewports, and the first 10 seconds (uncovering the ground truth saliency map), while using the converged saliency map as a predictor yields a CC of just 0.47.

Although this is useful as a first-order approximation for time-dependent saliency, there is still work ahead to adequately model time-dependent saliency over prolonged periods. In fact, due to the high inter-user variance of recorded scanpaths[3], the problem of predicting time-dependent saliency maps may not be a well-defined one. Perhaps a real-time approach that would use head orientation measured by an inertial measurement unit (IMU) to predict where a specific user will look next could be more useful than trying to predict time-dependent saliency without any knowledge of a specific user.

5.3 Can head orientation be used for saliency prediction?

The analysis in Section 4 indicates a strong correlation between head movement and gaze behavior in VR. In particular, Figure 6 (middle) shows that fixations usually occur with low head velocities (except for the vestibulo-ocular reflex). This insight suggests that an approximation of a saliency map may be obtained from the longitudinal head velocity alone, e.g. measured by an IMU, without the need for gaze tracking.

We validate this hypothesis by counting the number of measurements at pixel locations where the head speed falls below a threshold of $19.6\,^{\circ}/s$ for all experiments in the VR condition. We then blur this information with a Gaussian kernel of size 11.7° of visual angle, to take into account the mean eye offset while fixating (Figure 6, right). Qualitative results are shown in the supplemental material. For a quantitative analysis, we compute the CC score between these *head saliency maps* and the ground truth and compared it with the results obtained from the predictors examined in Table 2. Our CC score of 0.50 places our approximation on par with the performance of both saliency predictors tested; this is a positive and interesting result, given the fact that no gaze information is used at all. Head saliency maps could therefore become a valuable tool to analyze the approximate regions that users attend to from IMU data alone, without the need for additional eye-tracking hardware.

6 Applications

In this section, we outline several applications for VR saliency prediction. Rather than evaluating each of the applications in detail and comparing extensively to potentially related techniques, the goal of this section is to highlight the importance and utility of saliency prediction in VR for a range of applications with the purpose of stimulating future work in this domain.

6.1 Automatic alignment of cuts in VR video

How to place cuts in VR video is a question that was recently addressed by Serrano et al. [49]. In a number of situations, alignment of the objects of interest before and after the cut is a safe assumption, since it facilitates the viewer to "lock in" on the action immediately after the cut. The proposed saliency prediction facilitates automatic alignment of such cuts. We show in Figure 10 and in the supplemental video that predicted saliency maps can be used to align VR video before and after a cut by shifting the cuts in the longitudinal direction such that the Pearson CC of the predicted saliency maps is maximized. We use the 72 scenes provided by Serrano et al. [49], which were manually aligned to overlapping regions of interest (ROI) before and after a cut. However, in some there are multiple ROIs, and thus multiple meaningful alignments possible. We predict saliency maps before and after the cut using the predictor described as performing best in Section 5.1 (i.e., ML-Net with equator bias on equirectangular projection), and then shift the saliency map after the cut with respect to the saliency map before the cut such as to maximize the Pearson correlation. For the scenes with one ROI visible before and after the cut, the median error of our method with respect to the manually aligned results is 2.11°, which mildly increases to 9.14° if we include the scenes with two ROIs in the same field of view. Qualitative analysis shows that the alignments are meaningful and succeed to align salient regions, however, performance is strongly dependent on the quality of the saliency predictor used. This indicates that saliency-based automatic alignment of video cuts is a useful way to guide users when editing VR videos, suggesting good initial alternatives, but it may not be able to completely replace user interaction. Full alignment results can be found in the supplemental.

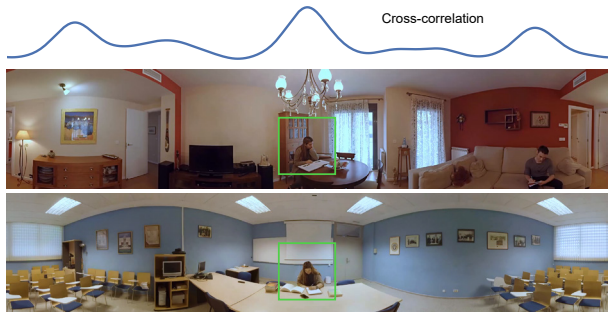

Fig. 10. Automatic alignment of cuts in VR video. To align two video segments, we can maximize the correlation between the saliency maps of the last frame in the first segment and the first frame of the second one. The cross-correlation accounting for all horizontal shifts is shown on top of this example, which has been automatically aligned with the proposed algorithm.

Fig. 11. Automatic panorama thumbnail generation. The most salient regions of a panorama can be extracted to serve as a representative preview of the entire scene.

6.2 Panorama thumbnails

Extracting a small viewport that is representative of a panorama may be helpful as a preview or thumbnail. However, VR panoramas cover the full sphere and most of the content may not be salient at all. To extract a thumbnail that remains representative of a scene in more commonly used image formats and at lower resolutions, we propose to extract the gnomonic, or rectilinear patch of the panorama that maximizes saliency within. To this end, we predict the saliency map of the entire panorama as discussed in Section 5.1. Then, we use an exhaustive search for the subregion with a fixed, user-defined field of view, that maximizes the integrated saliency within its gnomonic projection. A 2D Gaussian weighting function is applied to the predicted saliency values within each patch before integration to favor patches that center the most salient objects. While this is an intuitive approach, it is also an effective one. Results are shown in Figure 11 and, for all 22 scenes, in the supplemental material. Note that this approach to thumbnail generation is also closely related to techniques for gaze-based photo cropping [48].

[3]While converged saliency maps show a high inter-user agreement (Section 4.1), this is not necessarily the case for scanpaths, and thus for time-dependent saliency.

Fig. 12. Automatic panorama video synopsis. Saliency prediction in VR videos can be used to create a short, stop-motion-like animation that summarizes the video. For this application, we predict saliency of each frame, extract a panorama thumbnail from one of the first video frames, and then search every N^{th} frame for the window with highest saliency within a certain neighborhood of the last window.

6.3 Panorama video synopsis

Automatically generating video synopses is an important and active area of research (e.g., [44]). Most recently, Su et al. [51, 52] introduced the problem of automatically extracting paths of a camera with a smaller field-of-view through 360° panorama videos, dubbed *pano2vid*. Good saliency prediction for monoscopic and stereoscopic VR videos can help improve these and many other applications. Figure 12, for example, shows an approach to combining video synopsis and *pano2vid*. Here, we predict the saliency for each frame in a video as discussed in Section 5.1, and extracted the panorama thumbnail from the first frame, as discussed in the previous subsection. In subsequent frames, we search for the window in the panorama with the highest saliency that is close to the center of the last window. Neither the saliency prediction step nor this simple search procedure enforce strict temporal consistency, but the resulting panorama video synopsis works quite well (see supplemental video).

6.4 Saliency-aware VR image compression

Emerging VR image and video formats require substantially more bandwidth than conventional images and videos. Yet, low latency is even more critical in immersive environments than for desktop viewing scenarios. Thus, optimizing the bandwidth for VR video with advanced compression schemes is important and has become an active area of research [61]. Inspired by saliency-aware video compression schemes [19], we test an intuitive approach to saliency-aware compression for omni-directional stereo panoramas. Specifically, we propose to maintain a higher resolution in more salient regions of the panorama.

To evaluate potential benefits of saliency-aware panorama compression, we downsample a cube map representation of the omni-directional stereo panoramas with a bicubic filter by a factor of 6. We then upsample the low-resolution cube map and blend it with the 10% most salient regions of the high-resolution panoramas, using the ground-truth saliency maps. Overall, the compression ratio of the raw pixel count is thus 25%. Figure 13 shows this saliency-aware compression for an example image.

To evaluate the proposed saliency-aware VR image compression, we carried out a pilot study to assess the perceived quality of saliency-aware compression when compared to regular downsampling for a comparable compression ratio. To this end, users were presented with ten randomized pairs of stereo panoramas, and they were asked to pick the one that had better quality in a two-alternative forced choice (2AFC) test. For each pair, we sequentially displayed the two panoramas in randomized order, with a blank frame of 0.75 seconds between the two alternatives [43]. A total of eight users participated in the study, all reported normal or corrected-to-normal vision. The results of the study are shown in Figure 13 (bottom left). Saliency-aware compression was preferred for most scenes, and performed worse in only one scene. These preliminary results encourage future investigations of saliency-aware image and video compression for VR.

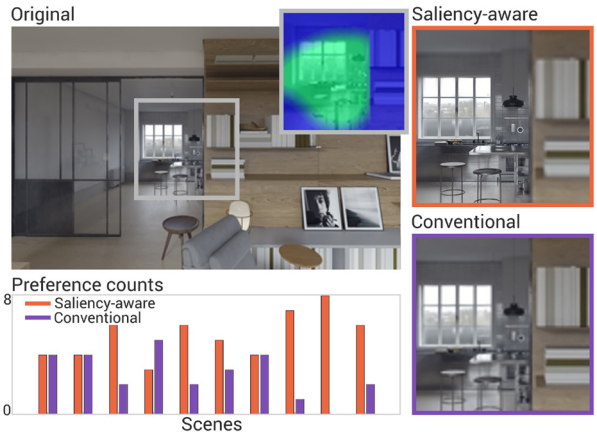

Fig. 13. Saliency-aware panorama compression. *Top left:* original, high-resolution region of the input panorama. Inset shows the compression map based on saliency information, where green indicates more salient regions. *Right:* Close-ups showing the differences between saliency-aware compression and conventional downsampling. Note that salient regions retain a better quality in our compression, while non-salient regions get more degraded. *Bottom left:* Preference counts for the ten scenes displayed during the user study.

7 Discussion

In summary, we collect a dataset that includes gaze and head orientation for users observing omni-directional stereo panoramas in VR, both in a standing and in a seated condition. We also capture users observing the same scenes in a desktop scenario, exploring monoscopic panoramas with mouse-based interaction. The data encompasses 169 users in three different conditions, totaling 1980 head and gaze trajectories.

The primary insights of our data analysis are: (1) gaze statistics and saliency in VR seem to be in good agreement with those of conventional displays; as a consequence, existing saliency predictors can be applied to VR using a few simple modifications described in this paper; (2) head and gaze interaction are coupled in VR viewing conditions – we show that head orientation recorded by inertial sensors may be sufficient to predict saliency with reasonable accuracy without the need for costly eye trackers; (3) we can accurately predict time-dependent viewing behavior only within the first few seconds after being exposed to a new scene but not for longer periods of time due to the high inter-user variance; (4) the distribution of salient regions in the scene has a significant impact on how viewers explore a scene: the fewer salient regions, the faster user attention gets directed towards any of them and the more concentrated their attention is; (5) we observe two distinct viewing modes: attention and re-orientation, potentially distinguishable via head or gaze movement in real time and thus useful for interactive applications.

These insights could have a direct impact on a range of common tasks in VR. We outline a number of applications, such as panorama thumbnail generation, panorama video synopsis, automatically placing cuts in VR video, and saliency-aware compression. These applications show the potential that saliency has for emerging VR systems and we hope to inspire further research in this domain.

Future Work Many potential avenues of future work exist. We did not use a 3D display or mobile device since we wanted to closely resemble the most standard viewing condition (regular monitor or laptop). Alternative viewing devices could be interesting for future work. Nevertheless, one of our goals is to analyze whether viewing behavior using regular desktop screens is similar to using a HMD, and our analysis seems to support this hypothesis. We believe this is an important insight, since it could enable future work to collect large saliency datasets for omni-directional stereo panoramas without the need for HMDs equipped with eye trackers.

Predicting gaze scanpaths of observers when freely exploring a

VR panorama would be very interesting in many fields, including vision, cognition, and of course, any VR-related application. Since the seminal work of Koch and Ulman [27], many researchers have proposed models of human gaze when viewing regular 2D images on conventional displays (e.g., [3, 18, 32, 60]). An important element to derive such models is gaze statistics, and whether those found in our VR setup are comparable to the ones reported for traditional viewing conditions; this would inform to what extent we can use existing gaze predictors in VR applications, or be useful as priors in the development of new predictors. Our data can be of particular interest to build gaze predictors using just head movement as input, since head position is much cheaper to obtain than actual gaze data.

Our data may still be insufficient to train robust data-driven behavioral models; we hope that making our scenes and code available will help gather more data for this purpose. We also hope it will be a basis for people to further explore other scenarios, such as dynamic or interactive scenes, the influence of the task, or the presence of motion parallax, etc. These future studies could leverage our methodology and metrics, and build upon them for the specific particularities of their scenarios. It would be interesting to explore how behavioral models could improve low-cost but imprecise gaze sensors, such as electrooculograms. Future work could also incorporate temporal consistency for saliency prediction in videos, or extend it to multimodal experiences that include audio.

8 Acknowledgements

The authors would like to thank Jaime Ruiz-Borau for support with experiments. This research has been partially funded by an ERC Consolidator Grant (project CHAMELEON), the Spanish Ministry of Economy and Competitiveness (projects TIN2016-78753- P, and TIN2016-79710-P), and the NSF/Intel Partnership on Visual and Experiential Computing (NSF IIS 1539120). Ana Serrano was supported by an FPI grant from the Spanish Ministry of Economy and Competitiveness. Gordon Wetzstein was supported by a Terman Faculty Fellowship and an Okawa Research Grant. We thank the following artists, photographers, and studios who generously contributed their omni-directional stereo panoramas for this study: Dabarti CGI Studio, Attu Studio, Estudio Eter, White Crow Studios, Steelblue, Blackhaus Studio, immortal-arts, Chaos Group, Felix Dodd, Kevin Margo, Aldo Garcia, Bertrand Benoit, Jason Buchheim, Prof. Robert Kooima, Tom Isaksen (Charakter Ink.), Victor Abramovskiy (RSTR.tv).

References

[1] R. Anderson, D. Gallup, J. Barron, et al. Jump: Virtual reality video. *ACM Trans. on Graphics (TOG)*, 35(6):198:1–198:13, Nov. 2016.

[2] P. Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics.*, 71(4):881–895, 2009.

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

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

[5] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT Saliency Benchmark. `http://saliency.mit.edu`, 2017.

[6] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv:1604.03605*, 2016.

[7] S. Chaabouni, J. Benois-Pineau, O. Hadar, and C. Amar. Deep learning for saliency prediction in natural video. *arXiv preprint arXiv:1604.08010*, 2016.

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

[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. A deep multi-level network for saliency prediction. In *Proc. of the International Conference on Pattern Recognition (ICPR)*, 2016.

[11] A. Doshi and M. Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of Vision*, 12(2):9, 2012.

[12] A. Duchowski, N. Cournia, and H. Murphy. Gaze-contingent displays: a review. *CyberPsychology & Behavior*, 7(6):621–34, 2004.

[13] E. Freedman. Coordination of the eyes and head during visual orienting. *Experimental Brain Research*, 190(4):369–387, 2008.

[14] A. Gabadinho, G. Ritschard, N. Müller, and M. Studer. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(1), 2011.

[15] A. Gibaldi, M. Vanegas, P. Bex, and G. Maiello. Evaluation of the tobii eyex eye tracking controller and matlab toolkit for research. *Behavior Research Methods*, 2016.

[16] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 34(10):1915–26, 2012.

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

[18] S. Hacisalihzade, L. Stark, and J. Allen. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Trans. on Systems, Man, and Cybernetics*, 22(3):474–81, 1992.

[19] H. Hadizadeh and I. Bajic. Saliency-aware video compression. *IEEE Trans. on Image Processing*, 23(1):19–33, 2014.

[20] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 262–270, 2015.

[21] L. Itti, C. Koch, E. Niebur, et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–59, 1998.

[22] Y. Jia and M. Han. Category-independent object-level saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1761–68, 2013.

[23] M. Jiang, X. Boix, G. Roig, J. Xu, L. Van Gool, and Q. Zhao. Learning to predict sequences of human visual fixations. *IEEE transactions on neural networks and learning systems*, 27(6):1241–1252, 2016.

[24] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1072–1080, 2015.

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[26] W. Kienzle, F. A. Wichmann, M. O. Franz, and B. Schölkopf. A nonparametric approach to bottom-up visual saliency. In *Proc. of the Conference on Neural Information Processing (NIPS)*, pp. 689–96, 2006.

[27] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence*, pp. 115–41. Springer, 1987.

[28] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14–14, 2014.

[29] T. Kollenberg, A. Neumann, D. Schneider, T. Tews, T. Hermann, H. Ritter, A. Dierker, and H. Koesling. Visual search in the (un) real world: how head-mounted displays affect eye movements, head movements and target detection. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 121–24. ACM, 2010.

[30] V. Laurutis and D. Robinson. The vestibulo-ocular reflex during human saccadic eye movements. *The Journal of Physiology*, 373(1):209–33, 1986.

[31] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–66, 2013.

[32] O. Le Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152–64, 2015.

[33] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar. Learning gaze transitions from depth to improve video saliency estimation. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[34] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5455–63, 2015.

[35] R. Liu, J. Cao, Z. Lin, and S. Shan. Adaptive partial differential equation

learning for visual saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3866–73, 2014.

[36] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 33(2):353–367, 2011.

[37] G. Marmitt and A. T. Duchowski. *Modeling visual attention in VR: Measuring the accuracy of predicted scanpaths*. PhD thesis, Clemson University, 2002.

[38] R. Nakashima, Y. Fang, Y. Hatori, et al. Saliency-based gaze prediction based on head direction. *Vision Research*, 117:59–66, 2015.

[39] C. Nguyen, S. DiVerdi, A. Hertzmann, and F. Liu. Vremiere: In-headset virtual reality video editing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2017.

[40] A. Nuthmann and J. Henderson. Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8):20, 2010.

[41] N. Padmanaban, R. Konrad, T. Stramer, E. Cooper, and G. Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proc. of the National Academy of Sciences (PNAS)*, 14:2183–88, 2017.

[42] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[43] 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. on Graphics (TOG)*, 35(6):179:1–179:12, Nov. 2016.

[44] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 435–41, 2006.

[45] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1153–1160, 2013.

[46] K. Ruhland, C. Peters, S. Andrist, et al. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In *Computer Graphics Forum*, vol. 34, pp. 299–326, 2015.

[47] D. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 71–8. ACM, 2000.

[48] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 771–80, 2006.

[49] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia. Movie editing and cognitive event segmentation in virtual reality video. *ACM Trans. on Graphics (TOG)*, 36(4), 2017.

[50] M. Stengel and M. Magnor. Gaze-contingent computational displays: Boosting perceptual fidelity. *IEEE Signal Processing Magazine*, 33(5):139–48, 2016.

[51] Y. Su, D. Jayaraman, and K. Grauman. Pano2vid: Automatic cinematography for watching 360° videos. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, 2016.

[52] Y.-C. Su and K. Grauman. Making 360° video watchable in 2d: Learning videography for click free viewing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[53] Q. Sun, L. Wei, and A. Kaufman. Mapping virtual and physical reality. *ACM Trans. on Graphics (TOG)*, 35(4):64:1–64:12, July 2016.

[54] V. Tanriverdi and R. Jacob. Interacting with eye movements in virtual environments. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 265–72, 2000.

[55] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766, 2006.

[56] S. Tregillus, M. Al Zayer, and E. Folmer. Handsfree omnidirectional VR navigation using head tilt. In *Proc. of the CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 4063–4068, 2017.

[57] E. Upenik and T. Ebrahimi. A Simple Method to Obtain Visual Attention Data in Head Mounted Virtual Reality. In *IEEE International Conference on Multimedia and Expo 2017*. IEEE, Hong Kong, July 2017.

[58] A. Volokitin, M. Gygli, and X. Boix. Predicting when saliency maps are accurate and eye fixations consistent. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 544–552, 2016.

[59] L. Wang, H. Lu, X. Ruan, and M. Yang. Deep networks for saliency detection via local estimation and global search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3183–92, 2015.

[60] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 441–8, 2011.

[61] M. Yu, H. Lakshman, and B. Girod. A framework to evaluate omnidirectional video coding schemes. In *Proc. of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2015.

[62] Q. Zhao and C. Koch. Learning saliency-based visual attention: A review. *Signal Processing*, 93(6):1401–7, 2013.

[63] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1265–74, 2015.