PREDICTING HEAD TRAJECTORIES IN 360° VIRTUAL REALITY VIDEOS

A. Deniz Aladagli[1], Erhan Ekmekcioglu[1], Dmitri Jarnikov[2, 3], Ahmet Kondoz[1]

[1]Loughborough University London, Institute of Digital Technologies,
3 Lesney Avenue, E15 2GZ London, UK
[2]Eindhoven University of Technology, Department of Mathematics and Computer Science,
De Lampendriessen, 5612 AZ Eindhoven, Netherlands
[3]Irdeto, High Tech Campus 84, 5656 Eindhoven, Netherlands
a.d.aladagli@lboro.ac.uk, e.ekmekcioglu@lboro.ac.uk, djarnikov@irdeto.com, a.kondoz@lboro.ac.uk

ABSTRACT

In this paper a fixation prediction based saliency algorithm is used in order to predict the head movements of viewers watching virtual reality (VR) videos, by modelling the relationship between fixation predictions and recorded head movements. The saliency algorithm is applied to viewings faithfully recreated from recorded head movements. Spherical cross-correlation analysis is performed between predicted attention centres and actual viewing centres in order to try and identify prevalent lengths of predictable attention and how early they can be predicted. The results show that fixation prediction based saliency analysis correlates with head movements only for limited durations. Therefore, further classification of durations where saliency analysis is predictive is required.

Index Terms— 360, VR, video, omnidirectional, head, prediction, saliency.

1. INTRODUCTION

There has been resurgence of increased interest in virtual reality (VR) in the recent years, particularly with previously expensive head mounted displays (HMD) now being more widely available to general consumers in different forms. A subgenre in this field is VR video (or 360° video) where an immersive viewing experience can be created by providing the users with omnidirectional videos and allowing them to select their viewing directions interactively with the tracking capabilities of HMDs. According to [18], the ideal VR experience with stereoscopic 3D vision requires a video with 60 frames per second frame rate at 6K resolution per eye utilized for the whole spherical view to make full use of the currently available HMDs such as the Gear VR. Depending on the quality, such a video needs to be encoded with 20 to 40 megabits per second (Mbps) [18]. This makes VR videos problematic to provide through on-line streaming services, due to both bandwidth limitations and decrypting capabilities of end-user devices.

An idea explored in many previous works, is to deliver only the portion of the video which the user requires for viewing. One of the earlier works dealing with the partial delivery of omnidirectional video for viewing with HMDs proposes that the source video is sub-divided into tiles and only enough tiles to cover the user's field of vision is transmitted [9]. In this type of system the same region of the video is cached and when users change their viewing angle, they start seeing missing areas in their view due to cache failure and latency in the network. This problem is approached in several ways. In [16], the user is supplied with two videos, one low bandwith omnidirectional video to prevent users from seeing missing areas and one high quality partial video to switch into in order to provide a better experience once the user receives it at each head direction change. In [1], bandwith adaptive streaming techniques are proposed by making tiles of different size and quality available and optimizing the selection of streams depending on the network parameters. Several bandwidth usage optimization algorithms are proposed in [8] where multiple users are sharing a wireless network in a region of interest (RoI) video streaming scenario. Based on utility functions considering network conditions and overlapping user RoIs, the algorithms decide on which tiles are to be streamed and whether to use unicast or multicast. Utilizing these algorithms are shown to increase the number of users supported by wireless networks for RoI video streaming. In these kind of systems the delay for the quality increase or the ratio of high fidelity area that the user is experiencing can be used for evaluating the proposed methods.

Provision of quality of service (QoS) in such partial delivery systems can benefit from knowing where the user will look at beforehand, which can help decrease or eliminate delay for quality increase or reduce bandwidth consumption by delivering less contingency data overall. Similar challenges were addressed in [15] for RoI streaming of videos, where the RoI is chosen with mouse interactions. Authors of [15] proposed multiple predictors such as auto-regressive moving average (ARMA) of previous RoI centres, centering and/or

978-1-5386-4655-7/17/$31.00 ©2017 IEEE

stabilizing the position of tracked visual features in the centre of the RoI and maintaining the pixels in the centre of the RoI with the use of encoded motion vectors. Their results indicate that a median of multiple predictors provide the most robust prediction. A domain specific predictor developed for the RoI streaming of football matches in [14] is shown to perform marginally better then the ARMA and motion vector based predictors. The domain specific predictor works with several rules designed to keep the tracked ball and players near the RoI centre based on their distances to RoI centre and intensities in the residual from the previous frame. In the previously mentioned [8], the authors tried modifying the utility functions to use the access statistics of the tiles from previous viewings to help the tile selection process, however the wrong predictions based on the probabilistic access patterns were judged to cause too many invalid tile assignments. In [2] and [19] models for view centre prediction are proposed which utilize linear regression of a short window of previous view centres to estimate future view centres. In [2] and [3], a neural network is trained over a dataset of short windows of consecutive view centres and it is shown to perform marginally better than the linear regression approach in [2].

From the previous works, it seems that approaches based on analysis of videos are largely unexplored in the prediction of head movements during VR video viewings. While the visual attention modelling field has been extensively studied and is continuing to be studied, these are mainly developed for tasks such as visual saliency or fixation prediction and object segmentation. For overviews and benchmarks of such models and different visual saliency algorithms, readers are referred to [5] and [4]. The adaptation of saliency algorithms to VR still images is discussed in [21]. The prediction of sequential scanpaths in still images using saliency maps are discussed in both [13] and [11].

With most of the previous saliency studies focusing on still images and not at all on VR video viewings, the purpose of the work presented here is to investigate the utilization of saliency maps for the prediction of head movements of VR video viewers which can later be used for the interactive streaming of VR videos. Focussing on this area, a correlation analysis between real recorded head movements from VR video viewings and sequential fixations predicted from a saliency model is proposed in order to gauge and measure the predictive ability of that saliency model. The details of the fixation prediction method used and the analysis methodology is presented in the following sections followed by the acquired results.

2. FIXATION PREDICTION AND ANALYSIS

As mentioned before, predictive techniques used in streaming VR videos suffer from erroneous predictions. Therefore, in order to design a prediction algorithm utilizing saliency maps, the nature of the relationship between the saliency maps and the head rotations of the user needs to be investigated. The VR video viewing dataset presented in [2] is used in order to perform this investigation. This dataset has been collected using 16 VR videos in total, including documentary style videos, sports and dance performances, extreme sports videos and roller-coaster rides. The head orientations of the users are recorded in degrees in terms of yaw, pitch and roll with a period of approximately 100ms for a duration of 30 seconds. A minimum of 47 and an average of 61 users participated in each video.

According to [21], gazes in the scene are succeeded by head movements towards the same region. Therefore, if such gazes in the scene can be predicted, they might also be used to predict head movements. In order to estimate the accuracy of such a prediction, the method proposed in this work involves performing cross-correlation analysis between sequential predicted fixation locations and past head movements. Below, the details about the fixation prediction method based on the saliency algorithm are presented, followed by the approach for measuring the relationship between the predicted fixation sequences and the recorded head movements.

2.1. Saliency Calculation

GBVS [12] is chosen as the saliency prediction algorithm to perform correlation analysis. While there are more recent algorithms proposed in the literature specialized for videos such as the one presented in [20], GBVS is well cited, readily available, can include a motion feature channel to include temporal cues for use with videos and it runs relatively fast compared to more computationally complex algorithms.

Fixation prediction based saliency algorithms such as GBVS are designed for images presented to subjects on traditional 2D displays, therefore, some additional processing is required in order to create saliency maps for omnidirectional scenes. Since it is not possible for a subject to be influenced by bottom-up features that they cannot see, the fixation prediction algorithm is run on the captured viewport images of each subject for each viewing session, instead of the whole omnidirectional scene. This is achieved by replicating the viewing conditions and using recorded viewing directions where spherical linear interpolation is employed to calculate viewing directions for frames in between recorded samples. The dataset used in this study was presented previously in [2], where it is stated that the data was recorded while the subjects were using an Oculus DK2 headset. While the field of view (FoV) subjects experience may differ with eye distance to lens and eye seperation distance, they will usually be similar [6]. Therefore the default values provided by the Oculus SDK [17] are used while recreating the viewports of the users, where viewports have vertical FoV of 106.19° and binocular horizontal FoV of 95.06°. Given that the vertical resolution (h) of the used headset covers the vertical FoV (fov_v) of the subject, the horizontal resolution (w) required to cover the binocular

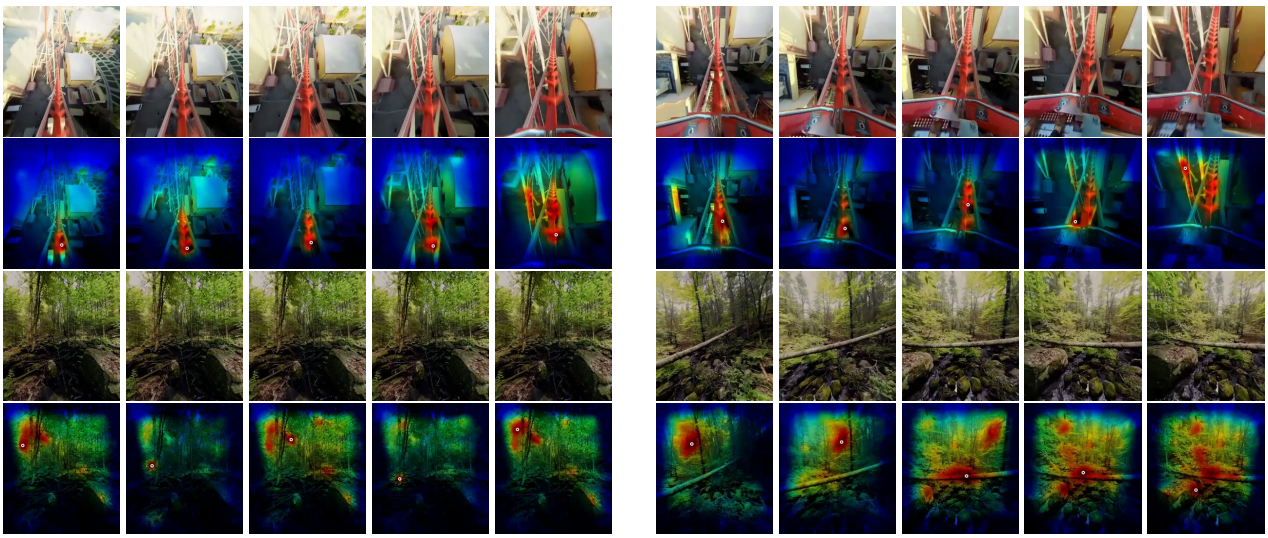


Fig. 1. Actual recreated viewports and overlaid saliency maps for 2 viewers (left and right) from roller coaster and forest documentary videos (up and down). The predicted fixation points are indicated as white circles in the overlaid saliency maps.

horizontal FoV (fov_h) that would be experienced with the headset can be calculated as below.

$$w = h * \frac{\tan \frac{fov_h}{2}}{\tan \frac{fov_v}{2}} \tag{1}$$

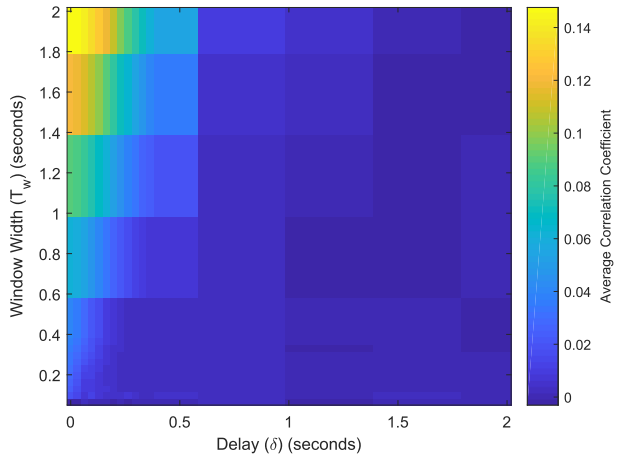
Using VR video projection, the viewports of the subjects are recreated for the durations of their viewings and a saliency map video is created by using GBVS on each viewport frame. With this process 985 saliency map videos in total are created matching the head movements recorded. Examples of extracted squences of saliency maps can be seen in Fig. 1. The techniques presented in [10] are used to identify a singular centre of attention on each frame of the saliency map video, where the pixel with the maximum saliency value in each frame is chosen as the location of attention in a winner-take-all manner.

2.2. Cross-Correlation Analysis

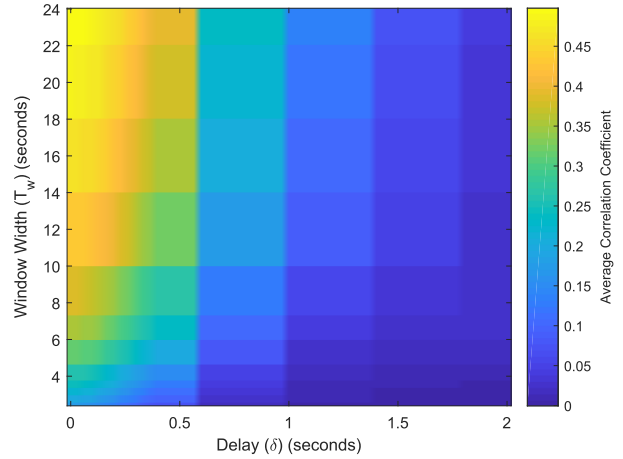
In order to understand the possible association between the predicted attention centres and head movements, the work here aims to identify 2 separate aspects of such relationship. The first is the size of the predictive lookahead margin (δ) that this relationship has and the second is the duration of strong association (T_w) required to identify a predictive association. Therefore, the qualitative test proposed here measures the cross-correlation between the series of predicted attention centres and actual recorded view centres across many different window widths of samples and delays. Higher correlation coefficients found with higher delay values could point to the predictive ability of the fixation prediction method.

The identified sequential attention centres are calculated in different 2D viewport spaces and therefore cannot be analysed together to represent attention shifts in 3D. In order to perform a time-series cross correlation analysis, all the attention centres need to be transformed into the same coordinate system. Firstly, they are projected to the unit sphere in 3D cartesian coordinates using the reverse of the VR video projection process. Second, their axis are aligned compensating for the head orientation at the moment of calculation with the following. Given the subject's head orientation at moment t is defined by yaw, pitch and roll values, projected attention centres on the unit sphere can aligned using a single yaw-pitch-roll rotation matrix.

With this, the attention centres for the whole viewing session are aligned with each other and are limited to being on the unit sphere's surface, same as the subject's actual view centres. Since the points being analysed can move all around on the sphere, performing correlation analysis on any single dimension might not reveal the nature of the relationship between predicted attention centres and head movements. Therefore, the spherical correlation coefficient described in [7] is deemed appropriate for this type of analysis. The spherical correlation coefficient measures the correlation between two series of spherical data. Positive and negative correlations are measured as having values of 1 and -1 respectively while a value of 0 is interpreted as having no correlation, similar to Pearson's correlation coefficient. In the context of spherical data, this coefficient measures how well the two series can be matched with an orthogonal transformation, where a positive correlation corresponds to the transformation being a rotation. In other words, this coefficient measures the similarity of the

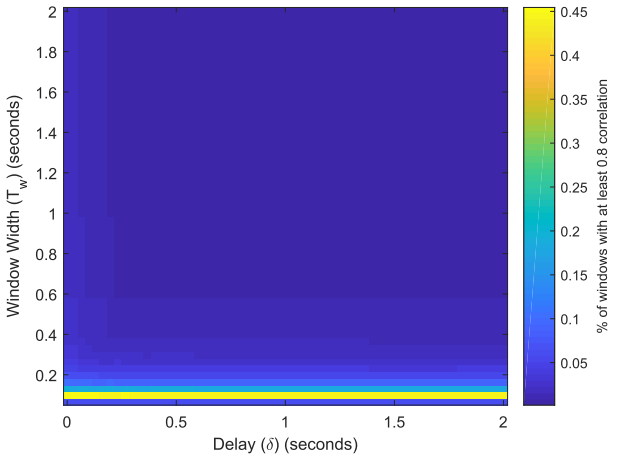


(a) $T_w \in [0.06s - 2s]$

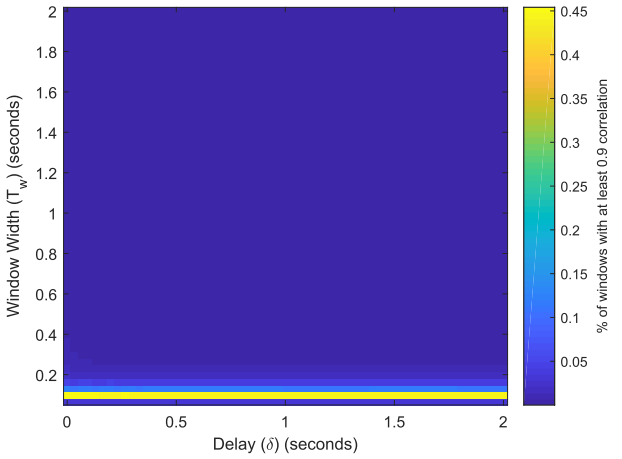


(b) $T_w \in [2s - 24s]$

Fig. 2. Averages of the measured correlations for different window sizes, $\delta \in [0s - 2s]$.



(a) Percantage of correlations higher than 0.8



(b) Percantage of correlations higher than 0.9

Fig. 3. Percantage of high correlations over different (δ, T_w) value pairs.

shapes of the paths drawn by the two sequences on the sphere.

In order to identify the look-ahead duration (δ) where the prediction method is most effective at and the lengths of attention windows (T_w) that the prediction method is accurate at, cross correlation is run with different values for both variables where the window of actual attention is slid over the whole viewing session for all viewings. Coefficients are evaluated for window widths (T_w) ranging from $0.06s$ to $2.0s$ and delay values (δ) ranging from $0.0s$ to $2.0s$, with a granularity of $0.03s$ until $0.4s$ and a granularity of $0.4s$ until $2.0s$ for both of the variables. As the identified attention centres are always within an approximately $50°$ radius of the actual viewing centre, there's a trend being forced between the two sequences. In order to get a baseline measure of this trend, window widths between $2.0s$ and $24.0s$ are also evaluated.

3. RESULTS AND DISCUSSION

The results shown in Fig. 2 (a) are not very encouraging as the highest average measured correlation is approximately 0.14 which is very low and since the (δ, T_w) pair that yields this result is measured for no delay ($0s$) and the widest window tested ($2.0s$). It is suspected that the increase in average measured correlation as the window width increases is due to the expected trend which was discussed in the previous section. This is seen more clearly in Fig. 2 (b), where the increase in the number of samples in the analysis window causes the smaller dissimilarities in the two sequences to be less apparent in the larger context. A subjective comparison of the distributions of measured correlations for different (δ, T_w) value pairs show that even though there are highly correlated sequences measured, they are not significant among any other measured sequences. With wider T_w values, the distributions

are more dense towards the higher mean, which is expected again due to the viewport trend.

From the results acquired here, it seems that a frame based winner-take-all approach to predict sequential fixation applied on saliency maps from GBVS is not robust enough. This is supported by the visual inspection of the predicted fixation points on extracted viewport videos, where the predicted points move around too much and too fast due to the flickering in the extracted saliency map. Compared to the smooth head movements of real viewers, this seems to be at least one of the obstacles for utilizing the predicted fixation spots for the prediction of head movements. There are indeed a certain percentage of sequences where the correlation is measured to be high as shown in Fig. 3, although the percentage not changing significantly with the changing delay is cause for caution since it indicates that the amount of delay might be insignificant. Therefore, the analysis performed here is deemed to be insufficient to identify a clear method to model the head movements after the predicted fixations.

4. CONCLUSION AND FUTURE WORK

The first contribution in this paper is a novel method to apply saliency algorithms to VR video viewings. Second, the relationship between sequences of fixations predicted from saliency maps and actual head movements of individual users is investigated. An inherent similarity due the viewport limitation is identified between the two sequences. This finding may benefit future works performing similar analysis, in disregarding the trivial correlations. A predictively meaningful relationship could not be identified between the two sequences.

In the future, other saliency algorithms which are more successful with videos such as the one in [20] will be explored together with a more robust sequential fixation prediction approach as proposed in [11]. As there is not an expectation of saliency maps to be accurate all the time for all users, a more detailed and continuous analysis will be performed to identify instances where the saliency map will be useful. The challenge here therefore is the real time evaluation of the accuracy of the saliency approach.

5. ACKNOWLEDGEMENTS

The work presented in this paper was carried out as part of CLOUDSCREENS, a Marie Curie Initial Training Networks action funded by the European Commissions 7th Framework Program under the Grant Number 608028.

6. REFERENCES

[1] P. R. Alface, J.-F. Macq, and N. Verzijp. Interactive omnidirectional video delivery: A bandwidth-effective approach. *Bell Labs Technical Journal*, 16(4):135–147, 2012.

[2] Y. Bao, H. Wu, T. Zhang, A. A. Ramli, and X. Liu. Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1161–1170. IEEE, 2016.

[3] Y. Bao, T. Zhang, A. Pande, H. Wu, and X. Liu. Motion-prediction-based multicast for 360-degree video transmissions. In *Sensing, Communication, and Networking (SECON), 2017 14th Annual IEEE International Conference on*, pages 1–9. IEEE, 2017.

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

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

[6] Field of view and scale. `https://developer.oculus.com/design/latest/concepts/bp_app_fov_scale/`. [Online; accessed 06-October-2017].

[7] N. I. Fisher, T. Lewis, and B. J. Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1987.

[8] R. Guntur and W. T. Ooi. On tile assignment for region-of-interest video streaming in a wireless lan. In *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*, pages 59–64. ACM, 2012.

[9] S. Heymann, A. Smolic, K. Mueller, Y. Guo, J. Rurainsky, P. Eisert, and T. Wiegand. Representation, coding and interactive rendering of high-resolution panoramic images and video using mpeg-4. In *Proc. Panoramic Photogrammetry Workshop (PPW)*, 2005.

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

[11] 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.

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

[13] . D. A. T. Marmitt, G. Modeling visual attention in vr : measuring the accuracy of predicted scanpaths. In *Eurographics 2002, Short Presentations*, page 217226, 01 2002.

[14] A. Mavlankar and B. Girod. Background extraction and long-term memory motion-compensated prediction for spatial-random-access-enabled video coding. In *Picture Coding Symposium, 2009. PCS 2009*, pages 1–4. IEEE, 2009.

[15] A. Mavlankar, D. Varodayan, and B. Girod. Region-of-interest prediction for interactively streaming regions of high resolution video. In *Packet Video 2007*, pages 68–77. IEEE, 2007.

[16] D. Ochi, Y. Kunita, K. Fujii, A. Kojima, S. Iwaki, and J. Hirose. Hmd viewing spherical video streaming system. In *Proceedings of the ACM International Conference on Multimedia*, pages 763–764. ACM, 2014.

[17] Oculus sdk for windows. `https://developer.oculus.com/downloads/package/oculus-sdk-for-windows/`. [Online; accessed 06-October-2017].

[18] Optimizing 360 video for oculus. `https://developers.facebook.com/videos/f8-2016/optimizing-360-video-for-oculus//`, 2016. [Online; accessed 06-October-2017].

[19] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan. Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, pages 1–6. ACM, 2016.

[20] A. Singh, C.-H. H. Chu, and M. A. Pratt. Learning to predict video saliency using temporal superpixels. In *ICPRAM (2)*, pages 201–209, 2015.

[21] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *arXiv preprint arXiv:1612.04335*, 2016.