A Saliency Dataset for 360-Degree Videos

Anh Nguyen
Georgia State University

Zhisheng Yan
Georgia State University

ABSTRACT

Despite the increasing popularity, realizing 360-degree videos in everyday applications is still challenging. Considering the unique viewing behavior in head-mounted display (HMD), understanding the saliency of 360-degree videos becomes the key to various 360-degree video research. Unfortunately, existing saliency datasets are either irrelevant to 360-degree videos or too small to support saliency modeling. In this paper, we introduce a large saliency dataset for 360-degree videos with 50,654 saliency maps from 24 diverse videos. The dataset is created by a new methodology supported by psychology studies in HMD viewing. We describe an open-source software implementing this methodology that can generate saliency maps from any head tracking data. Evaluation of the dataset shows that the generated saliency is highly correlated with the actual user fixation and that the saliency data can provide useful insight on user attention in 360-degree video viewing. The dataset and the program used to extract saliency are both made publicly available to facilitate future research.

CCS CONCEPTS

• **Information systems → Multimedia databases**.

KEYWORDS

Dataset, 360-degree videos, Saliency maps, Head-mounted display, Virtual Reality

ACM Reference Format:
Anh Nguyen and Zhisheng Yan. 2019. A Saliency Dataset for 360-Degree Videos. In *10th ACM Multimedia Systems Conference (MMSys '19), June 18–21, 2019, Amherst, MA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3304109.3325820

1 INTRODUCTION

Virtual Reality (VR) has the potential to become the mainstream of modern life. Its market share is expected to reach $47.7 billion in 2024 [15]. Under the broad umbrella of VR, 360-degree video is an important technology. This emerging video is captured by a 360-degree camera from all directions and then shown as a sphere centered at a user's head. By wearing a head-mounted display (HMD), the user is able to navigate through the panoramic content as she moves her head around. This brings a unique immersive experience that differentiates it from regular videos. Despite the promising experience, viewing 360-degree videos in diverse applications is still challenging. Due to the omnidirectional nature, the desired resolution of 360-degree videos are typically more than 12K [6]. This requires a higher bandwidth than ever to support the video transport. To make it worse, CPU cycles and energy are also significantly consumed on tasks such as streaming, rendering, and decoding.

Considering the unique user interaction pattern when viewing 360-degree videos, understanding users' visual attention, or *saliency*, in HMDs has become a key to 360-degree video research. First, an accurate saliency detection model can improve head movement prediction for 360-degree video viewing [9, 21] and thus optimize viewport adaptive streaming systems [6, 23], where the client only downloads the content likely to be viewed. In addition, including 360-degree saliency as a feature can improve the performance of a wide spectrum of applications, ranging from video compression [12] to salient object segmentation [19], image retargeting [25], and supporting human eye adaptation within HMD [29]. Moreover, a rich saliency dataset could be used to investigate the complex relation between stimuli and user attention. It can help neuroscientists and psychologists to understand the underlying process of human brain and visual cognition.

It is well known that a large and comprehensive dataset is needed to build a strong computational model for saliency prediction[11,21]. Although many saliency models have been developed for regular videos/images thanks to large-scale saliency datasets such as SALICON [14], they cannot be transferred and directly applied to a 360-degree video under a VR headset. Recently, Fan et al. generated saliency maps for 360-degree videos by a model trained on regular images [9]. However, they do not reflect users' visual attention in HMD. Although efforts have been made to extract saliency in HMD, the size of the two existing datasets are still small (60 samples for [24] and 5,700 samples for [8]). The saliency models resulted from small datasets are likely to be biased and overfitted.

In this paper, we introduce a saliency dataset for 360-degree videos with 50,654 samples. Our saliency maps are extracted from viewing sessions of more than 48 users on 24 videos ranging from 60 seconds to 655 seconds. While saliency datasets for regular images/videos can be generated by capturing eye gaze points and fixation using eye-tracking devices, specialized HMDs with accurate eye tracking are not widely available. Considering the unique user interaction in HMDs, we adopt a simple yet proven method as in [1, 26] to represent eye gaze point in HMD by head orientation. This methodology is supported by the fact that the head tends to follow eye movement to preserve the eye-resting position (i.e., eyes looking straight ahead) [17]. We also present a software to to extract saliency maps from head tracking logs. Our software is freely available for the public and can be used to generate saliency maps for any 360-degree videos and head tracking data.

MMSys '19, June 18–21, 2019, Amherst, MA, USA

ACM ISBN 978-1-4503-6297-9/19/06…$15.00
https://doi.org/10.1145/3304109.3325820

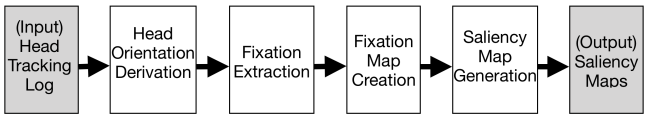


Figure 1: Saliency extraction for 360-degree videos

To extract saliency maps from head movement, we propose a saliency extraction method based on psychology studies on human eye-head movement behavior [10, 11, 17]. After the fixation (where users focus) is extracted by filtering out the saccade (fast movement during which the brain does not process input from the eyes [10]), we generate the saliency maps. We evaluate the consistency between the extracted saliency maps and the actual user fixation using popular metrics such as sAUC, NSS, and CC [3]. The results show that the proposed saliency dataset achieves a similar performance as state-of-the-art saliency datasets [3, 22]. Furthermore, we analyze the saliency dataset to provide insight for 360-degree video viewing. We observe that users tend to focus their attention with less movement when viewing videos captured by fast-moving cameras and videos presenting high contrast and few salient objects.

To the best of our knowledge, the proposed dataset is the largest public saliency dataset for 360-degree videos. The specific contributions are summarized as follows. First, a saliency dataset for 360-degree videos that includes 50,654 saliency maps from 24 videos of various types is generated. Second, a methodology to extract saliency maps is proposed, which can be extended to head movement logs from any 360-degree videos for enlarging the saliency dataset for the research community. Third, an extensive evaluation of the dataset is conducted to validate the consistency between the saliency and the actual user fixation as well as to provide insight on user attention in 360-degree video viewing.

2 RELATED WORKS

Visual attention for regular image/video is an established topic that has been studied for many years. The role of visual attention in the process of objection recognition in human brain was studied in [20]. The concept of saliency map was later mentioned in [18] to address the conspicuity of spatial region in images. Later, several large saliency datasets were introduced to improve the performance of computational saliency models. Borji et al. [2] developed a dataset of 4,000 samples from 120 users viewing regular images from 20 different categories. The fixation data was collected by dedicated eye tracking device in highly controlled experiments. Jiang et al. introduced SALICON [14], a large-scale dataset with 15,000 samples that can be used to train Deep Convolutional Network (DCNN) with minimal overfitting [22]. The dataset utilized Amazon Mechanic Turk to capture mouse clicks. The data collection procedure was set up such that the mouse clicks from users simulate human visual attention in a free viewing context. These datasets have been widely used in regular image and video research. However, they do not reflect the human attention for 360-degree video and thus should not be directly applied to this new technology.

Few works have focused on user attention for 360-degree videos and images. Fan et al. collected head tracking data from 50 users viewing 10 videos [9]. However, the introduced saliency datasets were generated from a model trained on regular images [7]. Thus, the saliency maps do not reflect users' attention in 360-degree videos. Rai et al. created a saliency dataset from 40 users viewing 60 omnidirectional image [24]. Each image was viewed for 25 seconds while the head and eye were tracked by a customized eye tracker installed into the HMD device. Recently, David et al. adopted a similar approach to collect tracking data from 19 videos of five categories [8]. Each video has 20 seconds and was viewed by 57 users. The fixation derived from the tracking logs was then used to generate saliency maps. Unfortunately, the size of these datasets are still limited to support advanced saliency modeling, e.g., using deep neural networks. To bridge this gap, we propose the largest 360-degree video saliency dataset so far and a saliency extraction software to enable further research in saliency prediction and 360-degree video systems.

Table 1: Statistics of Head Tracking Logs

Logs Name	No. of Users	No. of Videos	Log type
Corbillon [5]	59	5	Quaternion
Wu [27]	48	9	Quaternion
Lo [20]	50	10	Euler angles

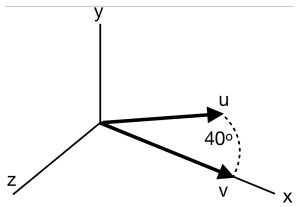


Figure 2: Deriving head orientation vector

3 DATASET GENERATION

In this section, we introduce the steps to extract saliency for 360-degree videos. The procedure is summarized in Figure 1.

3.1 Head Tracking Input

To generate saliency maps, the proposed saliency extraction framework receives head tracking logs as input. The head tracking logs are obtained from three public datasets [5, 20, 27]. Table 1 shows the number of users, the number of videos, and the head orientation representation of each head tracking dataset. We exclude 9 videos from Wu's head tracking logs since the data was collected while users were performing assigned tasks such as tracking/counting objects. Therefore, they were not captured in a free-viewing condition. The average duration of the head tracking logs is 164 seconds, with a minimum of 60 seconds and a maximum of 655 seconds.

3.2 Head Orientation Derivation

We first derive the head orientation vector and treat it as a consistent head orientation format across different head tracking logs. Existing logs record users' head movement by using the rotation between a reference unit vector and the current head orientation. Depending on the platform, the rotation could be represented as 4-tuple quaternion [5, 27] or yaw, pitch, roll [20]. Quaternion is a 4-tuple representation that is equivalent to the rotation matrix in 3D. Euler angles represent the rotations along individual axes. Hence, by applying a rotation operation to the unit reference vector, the head orientation vector can be derived. This process can be illustrated by Figure 2. Given a head orientation represented as a

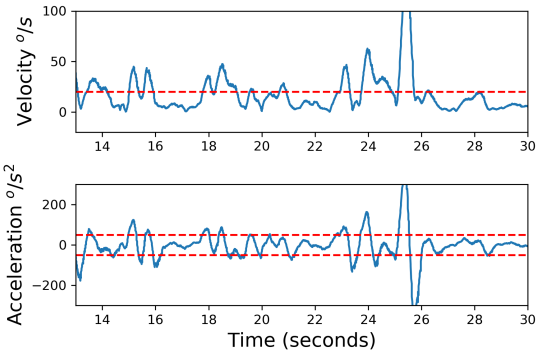


Figure 3: Saccade filtering using thresholds (dotted line)

Euler angle (yaw=40, pitch=0, roll=0) or a quaternion (0.940, 0.0, 0.342, 0.0), the head orientation vector u can be derived by applying a 40^o counter-clockwise rotation along the y axis on the reference unit vector v. Coupled with the timestamps, we are able to derive where the user is looking at on the 3D sphere for any given moment.

3.3 Fixation Extraction

In this step, the fixation is extracted from head orientation vectors. Fixation happens when users' head orientation focuses at a specific area for a short period of time. Before extracting the fixation, saccades must be filtered out. Saccades are very fast movement during which the brain does not process visual input. Thus, they do not reflect user attention. To remove saccades, head turning velocity and acceleration are first derived from head orientations. Then, based on study in [10], head movement with velocity over $20^o/s$ and acceleration magnitude greater than $50^o/s^2$ is considered saccade and filtered out.

Figure 3 illustrates an example of filtering out saccade of the video "Conan1". The red dotted lines are the thresholds of velocity (top figure) and acceleration (bottom figure). Those data whose velocity and acceleration exceed the thresholds are cut off.

We then associate the filtered head orientation vectors with the video frame under viewing. The head orientation vectors are converted into pixel coordinates to create fixation maps (equirectangular frame format) using the following formulas:

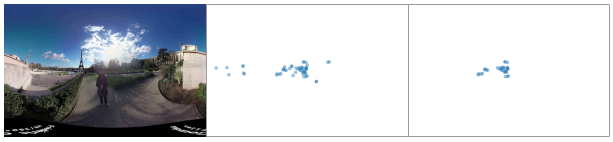
$$a = \frac{\phi}{360} * \mathcal{W} \tag{1}$$

$$b = (\frac{1 - \sin(\theta)}{2}) * \mathcal{H} \tag{2}$$

where a and b are the longitude and latitude positions in the equirectangular frame, ϕ and θ are the vertical and horizontal angles of head orientation vector u in 3D space, and $\mathcal{W}$ and $\mathcal{H}$ are the width and height of the target equirectangular frame.

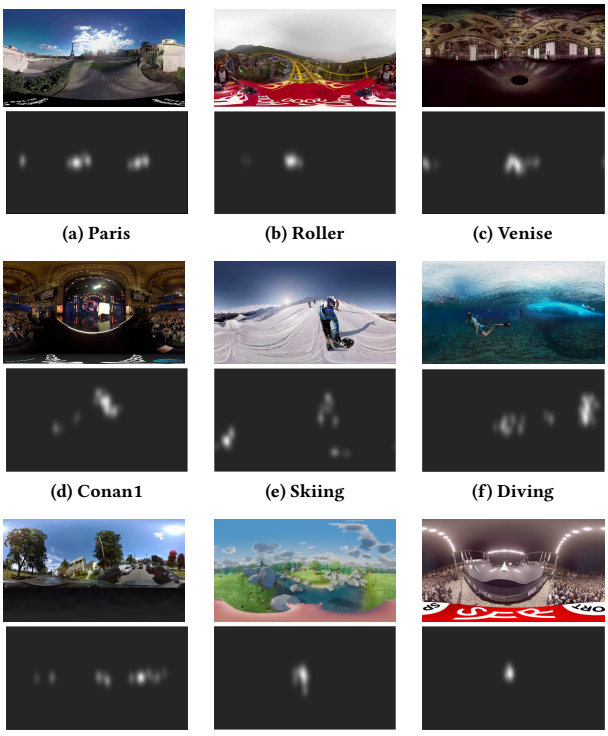
3.4 Fixation Map Creation

Fixation map is the aggregation of fixation points from all users viewing a video at a given timestamp. While fixation from different users usually can mark the region of interest, isolated fixation from few users could be the results of random behavior. Thus, it is necessary to filter out these noisy data points. We choose the Density-Based Spatial Clustering (DBSCAN) algorithm to filter out noise fixation points. This is because DBSCAN, unlike K-mean, returns high-density fixation samples (core samples) without introducing additional data. Based on the density of fixations points



(a) Video frame (b) Raw fixation (c) Filtered fixation

Figure 4: Fixation map filtering by DBSCAN



(a) Paris (b) Roller (c) Venise

(d) Conan1 (e) Skiing (f) Diving

(g) Driving (h) Game (i) Sport

Figure 5: Sample saliency maps in the dataset

in our pre-filtering fixation maps, the DBSCAN is configured to remove most noise in clusters with high density and still be able to retain some core fixations points in cluster with less density. Similar approaches have been previously applied to 360-degree images [1].

Figure 4 illustrates the effect of filtering noisy fixation. While the majority of user fixation focus on the man's face and the feet of the Eiffel tower, some users randomly look around. In Figure 4c, the DBSCAN filters out the most irrelevant fixation points. The core fixation samples now reflect the two most salient areas in the frame.

3.5 Saliency Map Generation

While fixation maps can manifest users' attention at some specific points, they ignore the area in between those points. In fact, it is important to identify continuous regions of interest [4, 26]. This problem could be addressed by applying a Gaussian filter on the fixation maps. Specifically, we assign saliency level to an area based on the density of the fixation around it. Such a classic method has previously been adopted in [1, 16] to generate saliency maps.

Figure 5 shows several illustrative examples of the created saliency maps. These saliency maps imply the areas where the majority of users pay attention to. In these examples, there is a tendency to

focus on small and conspicuous objects such as human (5a, 5d, and 5i). This is similar to the observed behavior in regular image/videos. However, we notice that there are also some distinct phenomena. For example, users highly focus on the target of fast-moving cameras in videos such as Roller (5b), Driving (5g), Game (5h) and Skiing (5e). In addition, users also tend to ignore large and close objects and prefer far-away and small objects that attract their interests in some videos such as Skiing (5e), and Diving (5f).

3.6 Program Structure

The generated saliency maps are filtered one more time to remove maps with negligible saliency, i.e., users' head orientations are randomly scattered due to the lack of region of interest. We eventually create a dataset of 50,654 saliency maps from 24 videos. The saliency maps for each video are stored together in one file. The data in each file is organized into records. Each record has three fields: *timestamp*, *fixation*, and *saliency map*. The first field is the relative video time in seconds for the saliency maps. The second field is a list of fixation points. Each fixation point is a unit vector representing the head orientation in the three-dimensional space. The third field is the saliency map, where each pixel is a float number representing the saliency level in the original video frame.

Our software can receive head tracking logs for any 360-degree videos and return saliency maps stored in pickle formats. The scripts (written in Python) reside in the root folder. The *data* folder contains saliency map files and the URLs to the original 360-degree videos hosted in Youtube. The naming convention for saliency map files is *saliency_ds<ds>_topic<vid>* where *ds* is the index of the data source (Corbillon, Wu, and Lo) and *vid* is the video name. The saliency maps and the software are made publicly available at Github. [1]

4 DATASET EVALUATION AND ANALYTICS

In this section, we evaluate the consistency and explore some of the characteristics of the 360-degree saliency maps. The Intel AI DevCloud framework is used to calculate the evaluation metrics, analyze and visualize the saliency data.

4.1 Dataset Evaluation

First, we evaluate if the generated saliency maps are consistent with human attention by using several popular corresponding measures [3] such as shuffle Area Under Curve (sAUC), Normalized Scanpath Saliency (NSS), and Pearson's Correlation Coefficient (CC). These measures indicate the correlation between the saliency maps and the actual user fixation. We also compare the proposed saliency dataset with two baseline saliency generation methods. The Equator Bar is a model which linearly increases saliency level from the pole to the equator of the sphere. This results in a nonlinear increasing of saliency in the projected equirectangular frame. Similarly, in the Circle at Center baseline, a circle on the surface of the sphere is expanded from a given point. The saliency level decreases as it moves further away from that point. The projected equirectangular shows high saliency around the center of the frame.

Table 2 shows the evaluation results. The proposed saliency dataset achieves a higher score than both baselines in all metrics. The scores of these metrics are also consistent with the results of other state-of-the-art datasets [3, 22] indicating that the saliency dataset is reasonable and captures user attention. Moreover, the low scores of the baselines imply that user attention cannot be captured by simple heuristics that attempt to simulate biases. Finally, the scores of Circle at Center are better than those of Equator Bar. This is because the camera placement in many 360-degree videos tend to capture important targets at the center of the frame.

Table 2: Saliency Dataset Evaluation

Models	sAUC	NSS	CC
Our Saliency Dataset	0.862	4.873	0.916
Equator Bar	0.409	1.110	0.302
Circle at Center	0.589	2.472	0.502

4.2 Dataset Analytics

Next, we analyze the proposed saliency dataset to provide insight on user behavior in 360-degree videos.

4.2.1 Accumulated Saliency. We first examine accumulated saliency to investigate user attention on 360-degree videos in the spatial domain. Specifically, we sum together the saliency maps of a video across the time domain to create the accumulated saliency map. The accumulated saliency maps identify the most salient regions attracting the highest attention. We also randomly sample 600 fixation maps for each video to indicate which spatial regions have been actually explored by users. Note that some fixation points may not be shown on the accumulated saliency map since they may be viewed by only a few users.

The accumulated saliency and fixation maps are shown in Figure 6. We can observe that users' exploring pattern across the spatial domain is highly distinct for different videos. For very fast moving videos such as Roller, Drive, Game and Landscape, the salient regions are small and the fixation points are relatively clustered. This is because the fast camera-moving speed strongly restricts users' movement and therefore users focus on the moving direction of the camera. Similarly, the small size of highly salient region and the clusters of fixation points can also be observed in videos with few salient objects such as Cooking, Conan1, and Sport. This is because the small number of salient objects limits the users' navigating options.

4.2.2 Local Randomness Saliency. To quantify the extent of user navigation in accumulated saliency maps, we propose the Local Randomness Saliency (LRS) metric. It is calculated by applying an entropy filter [28] to the accumulated saliency map and then taking the mean of the output map. The entropy filter passes a convolution mask on the accumulated saliency map and calculates the Shannon Entropy each time. Therefore, the LRS metric can assign more energy to regions that attract users attention and capture the extent of users' spatial navigation behavior.

Table 3 shows the LRS values for each 360-degree video. There is a strong agreement between the LRS values and the accumulated saliency maps shown in Figure 6. Notably, videos with faster camera movement such as Roller, Coaster, Coaster2, Game, Ride and Drive all result in lower LRS values. This is because there are only a few regions with high saliency in these videos. On the other hand, videos with a static camera and fewer focus points such as Venise, Diving, Timelapse, Diving2, and Panel achieve higher values of LRS. This is attributed to the fact that there is no clear foreground object

[1] https://github.com/phananh1010/PanoSaliency

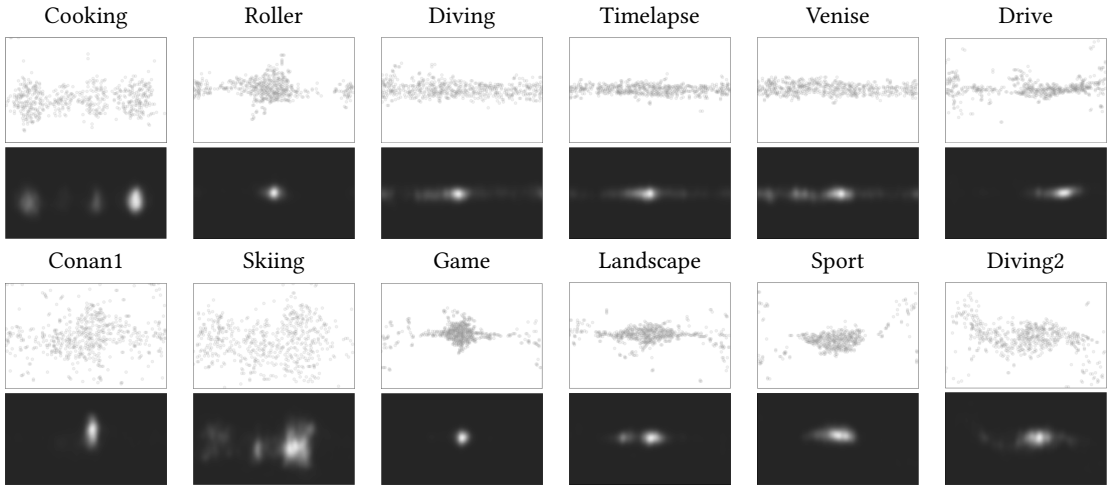


Figure 6: Sample accumulated saliency and fixation maps

Table 3: Local Randomness Saliency of the dataset

Coaster	Coaster2	Diving2	Landscape	Pacman	Panel	Drive	Ride	Game	Sport	Roller	Venise
0.35	0.265	0.723	0.503	0.266	0.61	0.453	0.339	0.252	0.492	0.347	0.952
Conan1	Skiing	Alien	Conan2	Surfing	War	Cooking	Football	Rhinos	Paris	Timelapse	Diving
0.439	1.187	1.548	0.434	1.303	0.801	0.839	0.745	1.378	0.557	0.779	0.892

to focus on when viewing these videos. As a result, users spend most of the time exploring around.

4.2.3 Head Movement Velocity and Saccade Percentage. We now investigate user attention in the time domain by studying some intermediate data that results in the saliency dataset. Specifically, we track the median head movement velocity of all users and the saccade percentage of each video. The saccade percentage is the portion of data removed during the saccade removal process discussed in Section 3.3. A high saccade percentage of a video implies that users move their head frequently to explore new content.

Table 4 shows the user interaction results in the time domain. In all cases, videos with higher head movement speed have more data identified as saccade. More interestingly, videos captured by a fast-moving camera such as Roller, Drive, Game, Landscape, Coaster, Coaster2, and Pacman have a low head movement speed. This verifies the effects we discussed in Section 4.2.1 and 4.2.2. In addition, the head movement velocity in videos such as Conan1, Conan2, and Cooking tend to be much lower. Since these videos have few salient objects, users do not have many options to explore the content.

5 DATASET SAMPLE USAGE

5.1 360-degree Saliency Prediction Model

In 360-degree videos, developing an accurate attention models has become the major challenge due to its role in many applications. While several strong saliency prediction models [7, 22] have been proposed for regular videos/images thanks to the large-scale dataset such as SALICON [14], the development of 360-degree saliency model has been limited due to the lack of similar large-scale 360-degree datasets. With the proposed large dataset, an improved saliency model could be trained to address various problems in 360-degree videos. One of the potential application of 360-degree saliency model is video compression. Similar to previous approaches in regular videos [12], identifying and encoding region of interest in higher quality could allow higher compression ratio and satisfactory user experience.

5.2 Head Movement Prediction

Head movement prediction is the key to bandwidth-efficient viewport-adaptive streaming for 360-degree videos, where only the viewport that users would look at in the near future is streamed. However, head movement prediction that only explores past head orientations was shown to achieve limited accuracy [13, 23]. Since most head movement are the users' reaction to video content, saliency maps could be used to identify areas that attract users' attention. By adding near-future saliency maps as an additional feature, the head movement prediction performance can be significantly improved. In fact, we have used a small subset of the proposed saliency dataset to build a preliminary head movement prediction model successfully [21]. Future work is needed to fully explore the larger saliency dataset and further improve head movement prediction performance.

5.3 360-degree Video Tile Preparation

Tile-based streaming systems cut the 360-degree video into tiles at server side and then stream the tiles covering user viewport. The saliency dataset could be used to improve tile preparation strategies. For example, in videos where users tend to focus on few locations such as Roller, Game, Drive, an aggressive tile preparation approach could be used i.e. only those few tiles covering the salient region would be prepared with higher bitrates while the remaining tiles would be encoded in low bitrates only. This can expedite the tile preparation and encoding at the server and is especially beneficial for live broadcasting systems. This approach is scalable because the server only needs to calculate saliency maps once in an offline fashion before developing a streaming strategy.

Table 4: Median head movement velocity and saccade percentage of the dataset

	Coaster	Coaster2	Diving2	Landscape	Pacman	Panel	Drive	Ride	Game	Sport	Roller	Venise
Velocity	4.609	5.602	12.226	12.701	7.678	15.619	10.505	9.523	7.27	12.508	8.468	11.326
Saccade pct.	0.35	0.265	0.723	0.503	0.266	0.61	0.453	0.339	0.252	0.492	0.369	0.427
	Conan1	Skiing	Alien	Conan2	Surfing	War	Cooking	Football	Rhinos	Paris	Timelapse	Diving
Velocity	6.183	9.2767	6.557	7.207	11.316	7.356	2.218	8.8711	6.297	9.016	15.275	11.959
Saccade pct.	0.323	0.380	0.320	0.348	0.425	0.366	0.217	0.395	0.337	0.388	0.429	0.512

6 CONCLUSION

In this paper, we introduce a saliency dataset for 360-degree videos with more than 50,654 samples and an open-source software to extract saliency maps from 360-degree videos with head tracking data. Motivated by psychology studies on user behavior in HMDs, we propose a methodology to capture fixation maps and then generate saliency maps for 360-degree videos. Evaluation results show that the proposed saliency dataset is highly consistent with the ground truth user fixation. Analytics of the dataset on the spatial and temporal are also presented to provide insight on user interaction pattern in HMDs.

Both the dataset and the source code for saliency maps extraction have been posted on public website for sharing. This work will bring a large dataset of 360-degree video saliency to the research community. It could potentially enable new powerful computational models for saliency detection that were impossible with the existing small datasets. The saliency dataset can also be used in various other areas, such as 360-degree video streaming and compression.

By using the thresholds to remove saccade, we do not consider the cases where users try to track fast-moving objects while their head is still moving. In this case, head orientation might not perfectly address the eye fixation. To address this issue and improve the accuracy of the saliency maps, we plan to incorporate the fixation data from the eye tracking VR headset in our future work.

7 ACKNOWLEDGEMENT

This work is supported by Intel AI DevCloud Usage for Research.

REFERENCES

[1] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. 2017. Look Around You: Saliency Maps for Omnidirectional Images in VR Applications. In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*.

[2] Ali Borji and Laurent Itti. 2015. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. In *arXiv:1505.03581 [cs.CV]*.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. 2018. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (March 2018), 1–1.

[4] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. 2011. Fusing generic objectness and visual saliency for salient object detection. In *IEEE International Conference on Computer Vision*.

[5] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-Degree Video Head Movement Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*.

[6] Xavier Corbillon, Gwendal Simon, Alisa Devlic, and Jacob Chakareski. 2017. Viewport-adaptive Navigable 360-degree Video Delivery. In *IEEE International Conference on Communications (ICC)*.

[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep Multi-level Network for Saliency Prediction. In *IEEE International Conference on Pattern Recognition (ICPR)*.

[8] Erwan J. David, Jesus Gutierrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360-degree videos. In *roceedings of the 9th ACM on Multimedia Systems Conference (MMSys'18)*.

[9] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation Prediction for 360 Video Streaming in Head-Mounted Virtual Reality. In *ACM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*.

[10] Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Eye-head coordination for visual cognitive processing. In *PLoS ONE 10(3): e0121035. https://doi.org/10.1371/journal.pone.0121035.*

[11] John M Findlay and Iain D Gilchrist. 2008. *Active Vision: The Psychology of Looking and Seeing.* Oxford Scholarship Online.

[12] Hadi Hadizadeh and Ivan V. Bajic. 2013. Saliency-Aware Video Compression. In *IEEE Transactions on Image Processing.*

[13] Jian He, Mubashir Adnan Qureshi, Lili Qiu, Jin Li, Feng Li, and Lei Han. 2018. Rubiks: Practical 360-Degree Streaming for Smartphones. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'18).*

[14] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV.*

[15] Mordor Intelligence. 2018. MS Windows NT Kernel Description. https://www.mordorintelligence.com/industry-reports/virtual-reality-market

[16] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV).*

[17] Wolf Kienzle, Bernhard Scholkopf, Felix Wichmann, and Matthias Franz. 2007. How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements. In *Hamprecht F.A., Schnorr C., Jahne B. (eds) Pattern Recognition. DAGM 2007. Lecture Notes in Computer Science.*

[18] Christof Koch and Shimon Ullman. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. In *Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science).* Springer.

[19] Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu. 2016. Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[20] Wen-Chih Lo, Ching-Ling Fan, and Jean Lee. 2017. 360-degree Video Viewing Dataset in Head-Mounted Virtual Reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17).*

[21] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. 2018. Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction. In *ACM Multimedia Conference for 2018 (ACMMM2018).*

[22] Junting Pan, Elisa Sayrol, Xavier G. Nieto, Kevin McGuinness, and Noel E. O'Connor. 2016. Shallow and Deep Convolutional Networks for Saliency Prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[23] Feng Qian, Bo Han, Lusheng Ji, and Vijay Gopalakrishnan. 2016. Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular Operations, Applications and Challenges - ATC '16.*

[24] Yashas Rai, Jesus Gutierrez, , and Patrick Le Callet. 2017. A Dataset of Head and Eye Movements for 360 Degree Images. In *roceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17).*

[25] Vidya Setlurand, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. 2005. Automatic image retargeting. In *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia (MUM'05).*

[26] Evgeniy Upenik and Touradj Ebrahimi. 2017. A Simple Method to Obtain Visual Attention Data in Head Mounted Virtual Reality. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW).*

[27] Chenglei Wu, Zhihao Tan, and Zhi Wang. 2017. 360-degree Video Viewing Dataset in Head-Mounted Virtual Reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17).*

[28] Chengxin Yan, Nong Sang, Tianxu, and Zhangb. 2003. Local entropy-based transition region extraction and thresholding. In *Pattern Recognition Letters.*

[29] Zhisheng Yan, Chen Song, Feng Lin, and Wenyao Xu. 2018. Exploring Eye Adaptation in Head-Mounted Display for Energy Efficient Smartphone Virtual Reality. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications (HotMobile'18).*