# Visual Attention in Omnidirectional Video for Virtual Reality Applications

Cagri Ozcinar, Aljosa Smolic
V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

***Abstract*—Understanding of visual attention is crucial for *omnidirectional video (ODV)* viewed for instance with a head-mounted display (HMD), where only a fraction of an ODV is rendered at a time. Transmission and rendering of ODV can be optimized by understanding how viewers consume a given ODV in virtual reality (VR) applications. In order to predict video regions that might draw the attention of viewers, *saliency maps* can be estimated by using computational visual attention models. As no such model currently exists for ODV, but given the importance for emerging ODV applications, we create a new visual attention user dataset for ODV, investigate behavior of viewers when consuming the content, and analyze the prediction performance of state-of-the-art visual attention models. Our developed test-bed and dataset will be publicly available with this paper, to stimulate and support research on ODV.**

***Index Terms*—Omnidirectional video (ODV); virtual Reality (VR); visual attention; saliency map; visual attention models.**

## I. Introduction

Tremendous interest can be observed in academia and industry these days regarding immersive video technologies and virtual reality (VR) video. Facebooks purchase of kick-starter company Oculus, developer of the Oculus Rift head-mounted displays (HMDs) in a reported US$2bn deal is one major proof of relevance. VR technology has now been used by film producers and streaming service providers, delivering immersive VR video experience using *omnidirectional video* (ODV). This emerging video representation can be captured by omnidirectional multi-camera arrangements and can be rendered through HMDs which allow the viewers to look around a scene from a central point of view in VR. In order to stay compatible with traditional (*i.e.,* rectilinear) video pipelines, ODV is typically produced and stored in a planar representation (*e.g.,* equirectangular projection (ERP)) and then projected back onto a sphere surface at rendering time.

Reaching levels of mass market adoption for ODV in VR applications, however, poses several challenges, of which some can be resolved by understanding the regions of ODV that attract the attention of viewers. The limitations of the ODV technologies are strongly related to the massive volume of video data that needs to be stored, transmitted and rendered compared to traditional video. Since HMDs use only a fraction of an ODV at a time, namely *viewport*, ODV can be optimized by predicting where the viewers' *visual attention* is concentrated at a given point in time. In this context, saliency maps, which predict viewer's eye fixations for given content, can be utilized for ODV. For instance, effective representation [1], cost-efficient resource utilization [2], compression gain [3]–[5], high-quality streaming [6], [7], and foveated rendering [8] would be possible using saliency maps in VR applications. Similarly, understanding of how to guide visual attention in VR, which is still an issue for film producers and story-tellers [9], might be supported by saliency maps. Thus, saliency estimation plays an essential role in the understanding of the visual cues that should be considered in VR applications.

To understand the salient regions of ODV viewed in HMDs, saliency maps can be estimated either by collecting eye fixations during subjective tests *or* by using visual attention models. The former is not always feasible because of the time required and the need to process large amounts of data. On the other hand, there is currently no dedicated visual attention model for ODV, mainly due to the lack of *ground-truth* datasets. Several visual attention models have been introduced over the last decade for *traditional* video, such as [10]–[16]. However, given the interactive look around nature of ODV consumption, these may not produce accurate saliency maps for a given ODV.

Although some recent research works stress the importance of saliency maps for VR [5], [6], [17]–[24], visual attention modeling for ODV with diverse content characteristics has *not* yet been studied sufficiently. Additionally, to our knowledge, no dedicated computational model exists for predicting salient regions of ODV in VR applications. Hence, the problem of predicting visual attention remains an open in the context of ODV. This problem is also emphasized by professional video service providers [18]. Motivated by this gap in the VR research, we analyzed content consumption of ODV viewed in HMDs, and examined the state-of-the-art computational visual attention models using a set of uncompressed ODVs. The contribution of this paper is two-fold. First, we created a dataset which contains a variety of content with different complexity. Our new dataset includes viewport trajectories (VT) and visual attention maps from 17 participants while watching uncompressed ODV. Our dataset and test-bed are available at[1][2]. The developed test-bed can be used to obtain VTs and visual attention maps without the need for eye tracking devices, which is an adequate use-case for many virtual reality applications [7], [25]. Second, we examine state-of-the-art visual attention models for traditional video, that are well cited or very recently published, using our generated dataset. To our knowledge, it is the first time such an evaluation of visual attention models for ODVs has been done with a publicly available test-bed and subjective user data. We expect that our dataset and analysis will be beneficial for future research in compression, streaming, visual quality assessment,

[1]https://v-sense.scss.tcd.ie/?p=1994
[2]https://github.com/cozcinar/omniAttention

978-1-5386-2605-4/18/$31.00 ©2018 IEEE

and computational visual attention modeling techniques for ODV.

The remainder of the paper is organized as follows. Section II discusses the related literature on visual attention. Section III describes the technical details of data collection and post-processing methods and Section IV presents the experimental results. Finally, in Section V, conclusions are drawn with directions for future research.

## II. Related Work

To model visual attention, user data must be first gathered through subjective experiments. A head movement dataset, for instance, has been created using several compressed ODVs in [26]. However, estimation and analysis of visual attention was not part of that work. Eye tracking data for omnidirectional images (ODIs) was collected by Rai *et al.* [27] to promote visual attention modeling for ODIs. Additionally, De Abreu *et al.* [19] studied visual attention for ODIs through subjective experiments. A grand challenge at the ICME conference 2017 stimulated further research in this area, by provding a test-set and evaluation framework [20]. Assens *et al.* developed SaltiNet [21] which is based on a deep neural network for scan-path prediction trained on ODIs. Their model estimates the temporal nature of gaze paths when exploring ODIs. Also, SalNet360 was introduced by Monroy *et al.* in [22] to expand a traditional CNN-based saliency estimation algorithm for ODIs. Another participant team of the ICME 2017 challenge, called xd_qsal, extended their previous saliency model [28] to ODIs.

Several computational visual attention models have been introduced over the past decade for *traditional* videos, such as [10]–[16]. For instance, a graph-theoretic solution in [10], named graph-based visual saliency (GBVS) was developed by Harel *et al*. Their method extracts RGB and motion channels in order to predict dynamic saliency maps. A simple solution for visual attention modelling, known as spectral residual (SR), was proposed in [11]. The method considers the rarity of visual features by analyzing the log-spectrum of a given content. Later, Guo *et al.* proposed the phase spectrum of quaternion fourier transform (PQFT) [12]. In addition to SR [11], this model uses the phase spectrum with estimated motion data to predict spatio-temporal dynamic saliency maps. Fang *et al.* in [14] merged the calculated spatial and temporal data into one saliency map using an entropy-based uncertainty weighting approach. Furthermore, Rudoy *et al.* proposed a learning-based visual attention model [13] which considers a sparse set of gaze locations. More recently, dynamic adaptive whitening saliency (AWS-D) was introduced in [15] using the idea that high-order statistical structures carry perceptual relevant information. Also, Wang *et al.* [16] proposed a visual saliency detection algorithm that incorporates information about motion boundaries, edges and color.

Considering that no visual attention model exists for ODV so far, and realizing the importance for VR research, we created a new dataset including diverse ODV contents and evaluated the above mentioned state-of-the-art visual attention models that are currently used for traditional video with our ODV dataset.

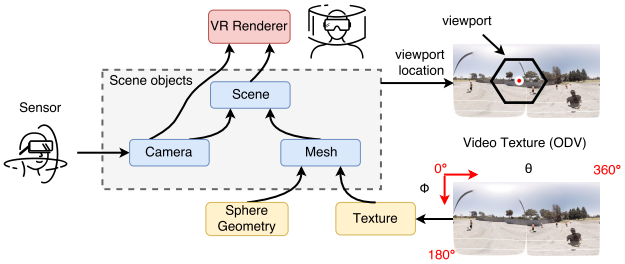



Fig. 1: Schematic diagram of the designed test-bed.

## III. Data Collection and Processing

In this section, we explain our experimental procedure for user data collection, and the processing methods which generate visual attention maps from the collected user data.

### A. Data collection

#### 1) Test-bed design

First, we designed a test-bed to collect the VTs for a given set of ODV from the participants. The testbed was implemented using two APIs, namely, `three.js` [29] and `WebVR` [30]. The former enabled us to create and display GPU-accelerated 3D graphics in a web browser. The latter enabled the creation of fully immersive VR experiences in a browser, allowing us to display a set of ODV without the use of any specific software other than a web browser. In our subjective tests, we used the Oculus Rift consumer version as HMD and Firefox Nightly as web browser.

The designed test-bed supports replay of various planar projections of ODV. As the ERP representation is the most widely used ODV format currently, at this stage, we only consider ERP as input format. `three.js` contains all the necessary objects, *i.e.,* scene, geometry, texture, and camera to construct a 3D geometry enabling interactive ODV replay in a browser. For that, a 3D mesh is created defining a sphere with texture. A given video texture (ODV) is mapped onto the sphere geometry such that the ODV can be viewed from the center of the sphere facing outwards *e.g.,* using an HMD. Therefore, the virtual camera is positioned in the center of the geometry. In a case of ERP, each point on the sphere is mapped on the planar surface using coordinates of its longitude ($0° \leq \theta < 360°$) and latitude ($0° \leq \Phi \leq 180°$). Hence, each row and column of the texture (*i.e.,* ODV frame) can be respectively represented by $\theta$ and $\Phi$. The developed test-bed can store $\theta$ and $\Phi$ values of the viewport center location during a viewing session at the HMD's frame refresh rate. Over time a viewport center trajectory (VCT) is recorded. Fig. 1 illustrates the schematic diagram of the designed test-bed for this study.

#### 2) Material

We used the following six *uncompressed* ODVs from the joint video exploration team (JVET) of ITU-T VCEG and ISO/IEC MPEG: $\mathcal{V}$ = {*LRRH*, *Gaslamp360*, *left_Driving360*, *train_le*, *basketball*, *left_Dancing360*} [31]–[33]. We selected only test data with *very high resolution* (*e.g.,* $\geq$ 4K×2K) to provide high quality for a given viewport. For instance, the *LRRH* sequence has 4K×2K resolution, and the other five ODVs have 8K×4K resolution. Each ODV is in ERP and

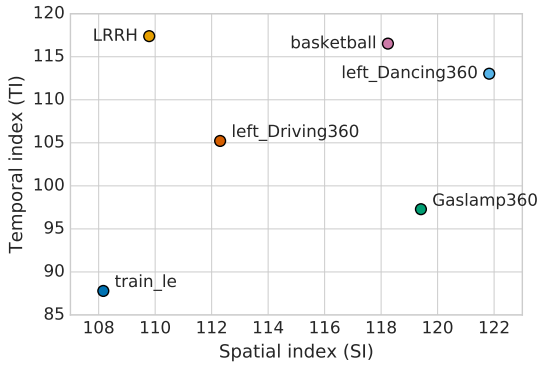



Fig. 2: Video statistics: SI and TI of ODV sequences used in the experiment.

YUV420p format, and of 10 *sec.* length. Also, a set of ODV was chosen to represent a *broad range of content complexities*. Spatial and temporal indices, SI and TI, of each ODV as calculated based on the ITU recommendation P.910 [34], are shown in Fig. 2, exhibiting a variety of different content types.

*3) Participants*

In all, 17 participants, 13 males and four females, took part in our subjective test. Four of the participants were researchers on the VR project, and the others were naïve viewers. All of the observers were screened and reported normal or corrected-to-normal visual acuity.

*4) Procedure*

Subjective tests were performed as *task-free* viewing sessions, *i.e.,* each participant was asked to naturally look at each given ODV. This undirected viewing is the most commonly used procedure for visual attention modeling. Participants were seated in a rotatable chair and allowed to turn freely. Each test session was split into a training and a test session. During the training session, an additional ODV, representative of the content, was shown. Then, during the test session, the six test ODVs were randomly displayed while the individual VCTs were recorded. We also varied the number of repetitions of each single ODV in $l_i \in \mathcal{L} = \{l_2, l_3, l_4\}$. To avoid motion sickness and eye fatigue, between two successive ODVs, we inserted a five *sec.* rest period with a gray screen. Also, before playing each ODV, we reset the HMD sensor to return the initial position ($\theta$=180° and $\Phi$=90°). Similar to [17], [19], we discarded the first fixation recorded for each video, as it added irrelevant information on the viewing direction.

### B. Data processing

We consider a visual attention estimation problem related to the center point of the viewport of the HMD. Because of the early stage of development, existing eye tracking technologies which support HMDs are either expensive or error-prone, and not generally accessible (*e.g.,* the WebVR API does not yet support eye tracking capability [30]). Our aim is to provide a testbed, which is widely usable and applicable without the need for specific hardware. To this end, as the head tends to follow eye movements to preserve the eye resting position (eyes looking straight ahead) [35], we considered the center point of the viewport as the visual target location. As we further consider only fixations as described below, the viewport center and the visual attention are even more likely to be closely related. Further, viewport information is important for many virtual reality applications [25]. Consequently, we recorded VCT user data in all our subjective experiments.

When viewing ODV, users are free to rotate their head to explore the content. For modeling visual attention we are only interested in their fixations. We consider a fixation to be given, if the VCT remains almost stable in a certain location for at least 200*ms*, which is a commonly used lower threshold. This requires clustering the VCT in order to remove influence from minor irrelevant movements and to reduce sensitivity to noise. In this work, we used the DBSCAN clustering algorithm [36], as it allows for setting the cluster neighborhood size $\tau$, which we set to 5° of any head rotation, and detects noise. Thus a fixation is recorded if the clustered VCT remains stable over 200*ms*.

Afterwards, the detected fixations for each participant were fused together to obtain the final fixation map for the $t$-$th$ sec. of a given ODV. More precisely, let $v$ be an ODV in the set of videos $\mathcal{V}$ viewed by participant $n \in \mathcal{N}$, where $\mathcal{N}$ is set of participants. The final fixation map, $F$, for the $t$-$th$ sec. of a given $v$ was calculated using a set of participants as follows:

$$F_v^t = \frac{1}{\eta} \sum_{n=1}^{\eta} f_{n,v}^t, \tag{1}$$

where $\eta$ is total number of participants in $\mathcal{N}$ and $f_{n,v}^t$ is the $t$-$th$ fixation map for the $n$-$th$ participant viewed a given $v$ content.

Finally, a dynamic visual attention map for each ODV was generated by applying Gaussian filtering to its estimated fixation sequence $F_v$. To this end, as there is a gradually decreasing acuity from the foveal vision towards the peripheral vision. Each $t$-$th$ fixation map was filtered using a Gaussian filter with a certain standard deviation corresponding to the high acuity vision area. For this, $\sigma$ was set to 5° in order to account for the 10° related to gaze shifts and the decreased visual acuity from the fovea.

## IV. Results

In the following, we investigate the behavior of viewers when consuming ODV with various content properties. Then, we discuss the prediction performance of state-of-the-art computational visual attention models.

### A. Analysis on our database

In this section, we study the behavior of participants in our VR viewing experiments. In the following, we investigate the effect of exposition time on the amount of fixations, influence of the content complexity, and distribution of fixations.

*1) How does the exposition time affect fixation?*

To study the effect of the exposition time for each ODV, we computed the average entropy from the corresponding dynamic visual attention maps with varying repetitions $\mathcal{L}$. We observe that increasing the number of loops leads to visual

attention maps with higher entropy, meaning higher variation. Fig. 3 (a) shows the average entropy for each ODV with various repetitions.

To further examine the relationship between the exposition time and fixations, we calculated the median value of the ERP longitude distance traveled by participants, a measure of how much users look around. Fig. 3 (b) illustrates these results depending on the number of loops (*i.e.,* $l_2$, $l_3$, and $l_4$) with the median absolute deviation shown on top of each bar. As evident from the figure, except for *LRRH* and *Gaslamp360*, increasing the number of loops does not increase the longitude distance traveled much. More clearly, repeating the content does not necessarily lead to more unique fixation points for most ODVs.

Further, we visualize visual attention maps for each ODV in Fig. 4, which also includes a sample thumbnail for each ODV. We observe that most fixations are densely concentrated at particular moving objects.

*2) Is the number of fixations similar between different viewers?*

In this section, we analyze how the number of fixations varies over participants. Fig. 3 (c) depicts the the average number of fixations (normalized values) for each ODV using the box-and-whisker diagram. With this diagram, we can illustrate the 75*th*, median, and 25*th* percentiles which are the upper, middle, and the lower edges of each box, respectively. The highest and the lowest average number of fixations are also given with the min-max line. Here, we observe that there is a high variation among participants in terms of the average number of fixations per ODV. This observation can be exemplified by the *Gaslamp360* sequence, which has high SI and low-value of TI. As it can be seen in Fig. 3 (c), there is a significant difference between maximum and minimum number of (normalized) average fixations. To support this observation, the differences between participants were statistically measured by the one-way analysis of variance (ANOVA). After the successful normality test, we found that the average number of fixations for each participant is *significantly* different ($p < 0.05$) from each other participant for each ODV group.

*3) Is there any relation between the amount of fixations and content complexity?*

In this section, we investigate the relation between the number of fixations and content complexity (in terms of SI and TI) of a given ODV. We observed that the *train_le* and *Gaslamp360* sequences, which have the lowest TI among all tested ODVs, received most fixations compared to the other ODVs. On the other hand, the *LRRH* sequence, which has the highest TI with low SI, received the lowest number of average fixations. This analysis indicates that a direct relationship exists between motion complexity of ODV and quantity of fixations.

*4) Fixation distributions*

To better understand the salient regions of each ODV, the fixation distributions for the latitude and the longitude of the ERP of each ODV are shown in Figures 5 and 6, respectively. Looking at latitude, we see that the fixation distribution is very dense in regions which are above 100°. This can also be observed in the visual attention maps in Fig. 4. Looking at longitude, we see that moving objects create most fixations for each content. As evident from the visual attention maps, fixations are densely concentrated at particular moving objects, which appear to be most salient.

### *B. Analysis of visual attention models*

In this section, we investigate the performance of state-of-the-art computational visual attention models, which were proposed for traditional video. In this experiment, we consider seven state-of-the-art computational visual attention models for video, namely, GVBS [10], SR [11], PQFT [12], Wang *et al.* [16], Fang *et al.* [14], Rudoy *et al.* [13] and AWS-D [15]. In addition, we included two computational visual attention models, namely SalNet360 [22] and xd_qsal [28], which are designed for ODI and were participants of the ICME 2017 challenge [20].

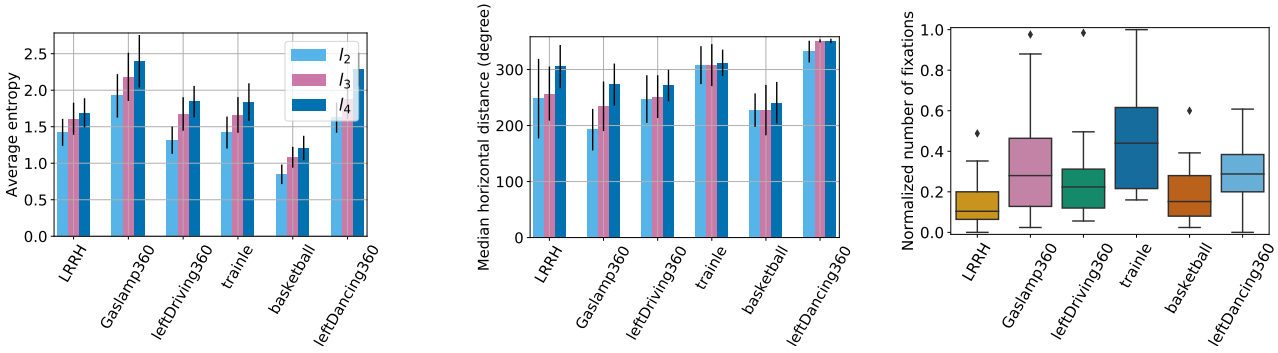
In this evaluation, we applied the area under receiver operating characteristic curve (AUC)-Borji and the normalized scan path saliency (NSS) metrics to measure the accuracy of estimates. The AUC metric considers a classification problem and uses the receiver operator characteristic curve to calculate the accuracy of the predicted visual attention maps in predicting the ground-truth fixations. NSS estimates the average normalized saliency score by measuring the correspondences between saliency maps and ground-truth, computed as the average normalized saliency at fixated locations. Unlike in AUC, NSS is sensitive to false positives [37]. The selected metrics are widely-accepted and standard for evaluating saliency models.

Table I reports the mean values of AUC and NSS with their standard deviations. Note that the model with a higher value of AUC and NSS can better predict the viewer fixations. We can see from the table that the method by Wang *et al.* [16], which emphasizes the salient object regions in dynamic scenes, performs very well (*e.g.,* second best and best in terms of AUC and NSS) compared to other models. To this end, we can conclude that the motion of salient objects plays an essential role in the context of visual attention in ODV, and we propose to take those regions into account for modeling of ODV visual attention. Furthermore, we evaluate two visual attention models which focus on ODIs. As a result, these two tested models perform substantially better compared to the visual attention models designed for standard video. To this end, incorporating information on salient object motion into the dedicated ODIs models can be expected to provide improved visual attention models for ODV.

## V. Conclusion

While visual attention maps for ODV can be estimated through existing prediction models for traditional video, these models may not produce accurate results for a given ODV due to the interactive look around nature of ODV consumption. For development of dedicated visual attention models for ODV, understanding of the viewing behavior and evaluation of state-of-the-art visual attention models are beneficial.

In this research, a new test-bed and a new visual attention user dataset for ODV were introduced, which can be used for development of new algorithms for processing of ODV. The developed test-bed can be used to obtain VTs and visual




(a) Entropy of dynamic saliency map for each ODV at $l_2$, $l_3$, and $l_4$.

(b) Median ERP longitude distance traveled by participants in $l_2$, $l_3$, and $l_4$.

(c) Box-and-whisker diagram for number of fixations.

Fig. 3: Qualitative analysis of the behavior of participants when consuming ODVs using HMD.

Fig. 4: A sample thumbnail frame with its visual attention for each ODV [31]–[33]. A frame for each ODV from left to right: *LRRH*, *Gaslamp360*, *left_Driving360*, *train_le*, *basketball*, and *left_Dancing360*.

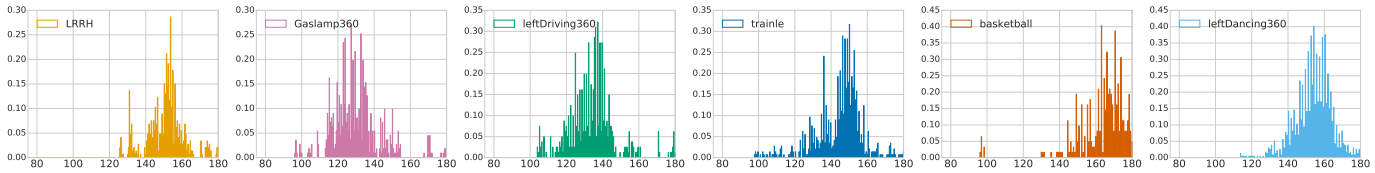



Fig. 5: Fixation distribution for the latitude of each ODV. In each histogram, the latitude value of the ERP and its number of fixations (normalized) are respectively represented by the x- and y- axes.

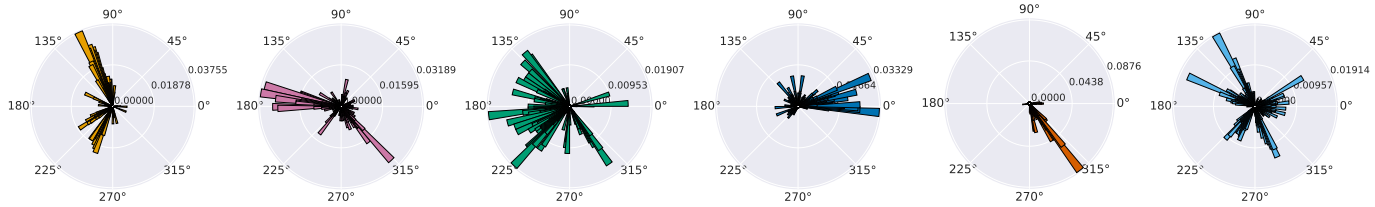


Fig. 6: Fixation distribution for the longitude of each ODV. In each polar sub-figure, the longitude value of the ERP and its number of fixations (normalized) are respectively represented by the angle and the radius of the polar plot.

attention maps without the need for eye tracking devices, which is an adequate use-case for many VR applications. The dataset includes VTs of seventeen participants and visual attention maps for a variety of content with different complexity.

Viewing behavior of participants when consuming ODV and prediction performance of state-of-the-art visual attention models were statistically analyzed using the gathered fixations in the conducted subjective studies. This is, to the best of our knowledge, the first comprehensive analysis for ODV viewing and evaluation of prediction models with a publicly available test-bed and subjective user data.

Our results show that repeating the content does not produce unique fixation points, and the quantity of fixations depends on the motion complexity of ODV. Also, we observed that the average number of fixations significantly differs among participants. Further, we learned that the evaluated visual attention models for standard video do not produce accurate visual attention maps for ODV.

In the future, we plan to further analyze the problem by investigating motion aspects, and formalize visual attention of ODV, by developing a new computational model.

## ACKNOWLEDGMENT

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. We gratefully acknowledge the support of NVIDIA Corporation for the donation of GPUs used in this work.

| **Model ↓ / ODV ↗** | *LRRH* | | *Gaslamp360* | | *left_Driving360* | | *train_le* | | *basketball* | | *left_Dancing360* | | *Mean* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | NSS | AUC | NSS | AUC | NSS | AUC | NSS | AUC | NSS | AUC | NSS | AUC | NSS |
| GVBS [10] | **0.80**(0.08) | *1.41*(0.45) | 0.52(0.04) | 0.92(0.09) | **0.80**(0.03) | *1.22*(0.15) | 0.47(0.02) | 0.54(0.26) | 0.52(0.08) | 1.03(0.43) | *0.66*(0.06) | 0.88(0.44) | **0.62** | 1.00 |
| SR [11] | 0.50(0.01) | 0.39(0.10) | 0.50(0.01) | 0.67(0.31) | 0.49(0.01) | 0.25(0.15) | 0.51(0.01) | 0.69(0.15) | 0.50(0.01) | 0.99(0.64) | 0.51(0.01) | 0.96(0.41) | 0.50 | 0.66 |
| PQFT [12] | 0.51(0.01) | 0.32(0.17) | 0.51(0.01) | 0.69(0.41) | 0.50(0.01) | 0.34(0.17) | 0.51(0.01) | *1.29*(0.33) | 0.51(0.01) | 1.42(0.89) | 0.50(0.01) | *1.29*(0.57) | 0.51 | 0.89 |
| Wang *et al.* [16] | *0.60*(0.11) | **1.67**(0.82) | *0.59*(0.07) | **2.13**(0.61) | 0.51(0.06) | 0.70(0.56) | **0.55**(0.06) | 0.74(0.20) | *0.61*(0.09) | **2.40**(0.88) | **0.69**(0.05) | **2.36**(0.36) | *0.60* | **1.67** |
| Fang *et al.* [14] | 0.51(0.02) | 1.33(1.37) | 0.51(0.03) | 1.02(0.63) | *0.54*(0.05) | **1.49**(0.51) | 0.51(0.04) | 0.31(0.54) | 0.51(0.03) | 1.17(0.75) | 0.52(0.03) | 1.53(0.59) | 0.52 | 1.14 |
| Rudoy *et al.* [13] | 0.54(0.08) | 1.08(1.31) | 0.51(0.06) | 0.67(0.70) | 0.45(0.10) | 0.65(0.70) | 0.51(0.04) | 0.51(0.51) | 0.48(0.04) | 0.67(0.96) | 0.53(0.07) | 1.56(0.46) | 0.50 | 0.86 |
| AWS-D [15] | 0.52(0.15) | 0.51(1.30) | **0.61**(0.09) | *1.43*(0.85) | 0.48(0.09) | 0.56(0.32) | *0.55*(0.14) | **1.39**(0.67) | **0.75**(0.14) | *2.26*(0.99) | 0.63(0.10) | *2.00*(0.52) | 0.59 | *1.36* |
| SalNet360 [22] | **0.70**(0.13) | 0.71(0.35) | **0.82**(0.06) | **1.67**(0.39) | 0.77(0.10) | **1.13**(0.28) | 0.65(0.16) | 0.96(0.55) | **0.81**(0.08) | **1.44**(0.35) | 0.70(0.05) | 1.08(0.03) | **0.74** | **1.17** |
| xd_qsal [28] | 0.69(0.14) | **1.79**(0.08) | 0.68(0.10) | 0.89(0.22) | **0.78**(0.09) | 1.02(0.26) | **0.77**(0.09) | **1.02**(0.25) | 0.59(0.16) | 0.88(0.75) | **0.72**(0.06) | **1.38**(0.33) | 0.71 | 1.16 |

TABLE I: Mean(standard deviation) values for saliency detection accuracy of state-of-the-art visual attention models over six ODVs (best in **blue**, second best in *red*).

## REFERENCES


[1] 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 *Panoramic Photogrammetry Workshop*, Berlin, Germany, Feb. 2005, pp. 24–25.

[2] C. Ozcinar, A. De Abreu, S. Knorr, and A. Smolic, "Estimation of optimal encoding ladders for tiled 360 VR video in adaptive streaming systems," in *The 19th IEEE International Symposium on Multimedia (ISM 2017)*, Taichung, Taiwan, Nov. 2017.

[3] C. Grunheit, A. Smolic, and T. Wiegand, "Efficient representation and interactive streaming of high-resolution panoramic views," in *2002 International Conference on Image Processing (ICIP)*, Rochester, NY, USA, USA, Sept. 2002, vol. 3, pp. III–209–III–212 vol.3.

[4] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *Journal of visual communication and image representation*, vol. 22, no. 8, pp. 704–711, 1 Nov. 2011.

[5] G. Luz, J. Ascenso, C. Brites, and F. Pereira, "Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2017, pp. 1–6.

[6] S. Rossi and L. Toni, "Navigation-Aware adaptive streaming strategies for omnidirectional video," in *IEEE International Workshop on Multimedia Signal Processing (MMSP 2017)*, London-Luton, UK, Oct. 2017.

[7] C. Ozcinar, A. De Abreu, and A. Smolic, "Viewport-aware adaptive 360°video streaming using tiles for virtual reality," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017.

[8] 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 transactions on graphics*, vol. 35, no. 6, pp. 179, 11 Nov. 2016.

[9] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "How do people explore virtual environments?," *arXiv:1612.04335v2*, 13 Dec. 2016.

[10] J. Harel, C. Koch, and P. Perona, "Graph-Based visual saliency," *NPS*, pp. 545–552, 2007.

[11] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[12] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[13] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 1147–1154.

[14] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE transactions on image processing*, vol. 23, no. 9, pp. 3910–3921, Sept. 2014.

[15] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 893–907, May 2017.

[16] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-Aware video object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[17] E. Upenik, M. Rerabek, and T. Ebrahimi, "A testbed for subjective evaluation of omnidirectional visual content," in *Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016.

[18] "Enhancing high-resolution 360 streaming with view prediction," https://code.facebook.com/posts/118926451990297/, Apr 2017.

[19] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–6.

[20] Y. Rai, P. Le Callet, and P. Guillotel, "Salient360 - the training dataset for the ICME grand challenge," in *IEEE International Conferance on Multimedia & Expo*, Hong Kong, July 2017.

[21] M. Assens, K. McGuinness, X. Giro-i Nieto, and N. E. O'Connor, "SaltiNet: Scan-path prediction on 360 degree images using saliency volumes," *arXiv*, 11 July 2017.

[22] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *arXiv*, 19 Sept. 2017.

[23] M. Xu, C. Li, Z. Wang, and Z. Chen, "Visual quality assessment of panoramic video," *arXiv:1612.04335v2*, 19 Sept. 2017.

[24] T. Maugey, O. Le Meur, and Z. Liu, "Saliency-based navigation in omnidirectional image," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct 2017, pp. 1–6.

[25] V.R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen, "Tiling in interactive panoramic video: Approaches and evaluation," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1819–1831, Sept 2016.

[26] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, Taipei, Taiwan, June 2017, pp. 199–204, ACM.

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

[28] C. Xia, F. Qi, and G. Shi, "Bottom-Up visual saliency estimation with deep Autoencoder-Based sparse reconstruction," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1227–1240, June 2016.

[29] "JavaScript 3D library. https://threejs.org/," https://github.com/mrdoob/three.js/, Feb 2017.

[30] "WebVR: Bringing virtual reality to the web," https://webvr.info/, Feb 2017.

[31] A. Abbas and B. Adsumilli, "Ahg8: New gopro test sequences for virtual reality video coding," Tech. Rep. JVET-D0026, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Oct 2016.

[32] E. Asbun, H. He, He. Y., and Y. Ye, "Ahg8: Interdigital test sequences for virtual reality video coding," Tech. Rep. JVET-D0039, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Oct 2016.

[33] G. Bang, G. Lafruit, and M. Tanimoto, "Description of 360 3D video application exploration experiments on divergent multiview video," Tech. Rep. MPEG2015/ M16129, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Feb. 2016.

[34] Telephone Transmission Quality, Telephone Installations, and Local Line Networks, "ITU-T p-series recommendations," 1999.

[35] O.-J. Grüsser and U. Grüsser-Cornehls, "The sense of sight," in *Human Physiology*, Robert F. Schmidt and Gerhard Thews, Eds., pp. 237–276. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[36] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 226–231, AAAI Press.

[37] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv:1604.03605*, 12 Apr. 2016.