Look Around You: Saliency Maps for Omnidirectional Images in VR Applications

Ana De Abreu, Cagri Ozcinar, Aljosa Smolic
Trinity College Dublin, Dublin, Ireland

***Abstract*—Understanding visual attention has always been a topic of great interest in the graphics, image/video processing, robotics and human-computer interaction communities. By understanding salient image regions, the compression, transmission and rendering algorithms can be optimized. This is particularly important in omnidirectional images (ODIs) viewed with a head-mounted display (HMD), where only a fraction of the captured scene is displayed at a time, namely *viewport*. In order to predict salient image regions, *saliency maps* are estimated either by using an eye tracker to collect eye fixations during subjective tests or by using computational models of visual attention. However, eye tracking developments for ODIs are still in the early stages and although a large list of saliency models are available, no particular attention has been dedicated to ODIs. Therefore, in this paper, we consider the problem of estimating saliency maps for ODIs viewed with HMDs, when the use of an eye tracker device is not possible. We collected *viewport center trajectories* (VCTs) of 32 participants for 21 ODIs and propose a method to transform the gathered data into saliency maps. The obtained saliency maps are compared in terms of image exposition time used to display each ODI in the subjective tests. Then, motivated by the equator bias tendency in ODIs, we propose a post-processing method, namely *fused saliency maps* (FSM), to adapt current saliency models to ODIs requirements. We show that the use of FSM on current models improves their performance by up to 20%. The developed database and testbed are publicly available with this paper.**

***Keywords*—Fixations; head-mounted display (HMD); omnidirectional images (ODIs); saliency maps; viewport; virtual reality (VR).**

I. Introduction

Recent advances in the design of head-mounted displays (HMDs) and in graphics computing power have made feasible the deployment of several virtual reality (VR) applications. One of the most prominent VR applications is omnidirectional images (ODIs). ODIs are spherical captured images that allow users to look around a scene from a central point of view while using an HMD. Compliant with standard 2D image processing, ODIs are stored in a planar representation, *e.g.*, equirectangular, cylindrical or cubic [1], and then projected back into a 3D geometry, *e.g.*, a sphere, at rendering time. Although ODIs application is rapidly increasing in popularity, there are still many barriers that are limiting its progress. Notably, the huge volume of data that needs to be stored, streamed and rendered compared to traditional (rectilinear) images. Some of these limitations could be solved if the viewing direction of the user were known beforehand. For instance, since only a fraction of the ODI is presented on the user HMD, namely the *viewport*, a foveated representation/rendering [2] or a viewport-aware transmission [3] would be possible, which could enhance graphics performance and save power and transmission bandwidth.

QoMEX2017 – Erfurt, Germany; 978-1-5386-4024-1/17/$31.00 ©2017 IEEE

In order to get an understanding of visual attention in a particular image, eye fixation points are usually collected using an eye-tracker. These eye fixations are then processed to get *saliency maps*, which give the visual importance per pixel in the evaluated image. However, in the case of ODIs viewed with HMDs, eye tracking developments are still in the early stages, thus they have an elevated cost and/or a compromised precision. Alternatively, *saliency models* can be used to predict the location of eye fixation points. However, although many models have been proposed in the literature [4], they have been mostly proposed for traditional images and they might not be suitable for the planar representation of the ODIs. For instance, these models do not account for the fact that ODIs in HMDs are presented as continuous spherical images. Thus, different from traditional images that have a strong center bias [5], fixations for ODIs are expected to be distributed through all the image, in particular in the equator surroundings.

In this work, we tackle the problem of estimating saliency maps for ODIs viewed in HMDs from the users' *viewport center trajectories* (VCTs) collected through subjective tests. Unlike traditional images, in ODIs users do not have full visual access to the image content. Thus, visual attention is caused not only by the features of the image, but also by the user curiosity and head movement speed. Therefore, the estimation of the optimal ODI exposition time becomes an important parameter of study. In this work, we conduct and analyze subjective tests where the exposition time is set to $10s$ and $20s$. Test results show that duplicating the exposition time does not necessarily lead to new salient data, concluding, for this limited scenario, that $10s$ is a reasonable ODI exposition time. Moreover, motivated by the lack of saliency models for ODIs and the good performance of current models for traditional images, we propose a simple post-processing method, namely *fused saliency maps* (FSM), to account for the equator bias feature in ODIs. We compare the performance of current saliency models with and without FSM and show that the proposed method improves the performance of current models by up to 20%.

The contributions of this paper can be summarized as follows. First, we present a database of VCTs for 32 participants and 21 ODIs and a testbed that successively displays ODIs while collecting participants' viewing direction. Both, the testbed and the database are available with this paper. Then, we propose a method for processing the gathered VCTs into saliency maps, considering the pixel distortion due to the 3D to 2D mapping of ODIs. We propose the *fused saliency maps* (FSM) method, an adaptation of current saliency models for ODIs. Finally,

the ODIs exposition time is studied and concluding results are presented in Section VI.

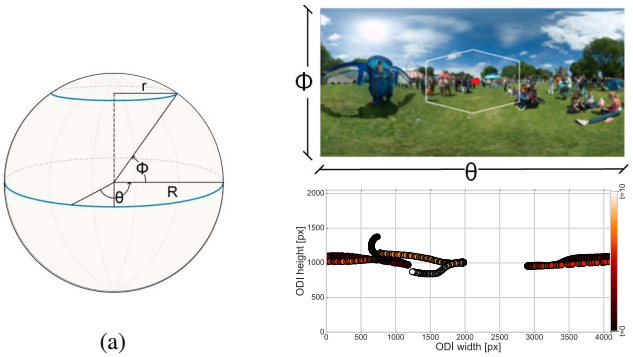
II. Related Work

In an early work [6], an innovative system for coding and transmitting ODIs has been proposed. However, due to low processing capabilities, low resolution displays and poor graphics performance at the time, the VR experience did not match the creative vision. Recently, technology advances have led to the rise of VR applications. Current research works on ODIs focus on the huge bandwidth demand issue of this emerging media type. To maximize coding efficiency, the authors in [2] have proposed novel representations for ODIs and the authors in [7] have adapted classical compression strategies to ODIs needs. Alternatively, in [3], the problem of high bandwidth demands of ODIs has been tackled by optimizing the delivery system using adaptive solutions, where simple strategies to anticipate head position were adopted. However, a rigorous study of visual user attention for ODIs has been missing in these works.

To understand visual attention in ODIs, ground truth data must be gathered through subjective tests. In [8], a testbed for subjective tests for ODIs has been presented. However, the proposed software application is limited to HMDs that use the display of mobile devices, such as Google Cardboard. Moreover, although preliminary results of generated saliency maps have been presented, the authors did not provide any details of the approach followed. In addition, a head movement dataset for omnidirectional videos has been introduced in [9]; however, visual attention studies were overlooked.

Saliency maps can also be predicted using models of visual attention. For the last 20 years many saliency models have been proposed for traditional 2D images [4]. The visual features considered by these models can be classified as low level features (color, intensity, orientation) [10], [11] and high level features (face [5] and object [12] detection, image-center prior [11]). The extracted features can be linearly combined [10] or integrated using learned weights [5] [13]. More recently, models using deep neural networks, trained for object recognition, [14] [15] have shown great performance on predicting visual attention [16]. However, these models may not suit the specificities of ODIs, which are viewed as continuous spherical images. A few works have tackled this problem. In [17], a saliency model for panoramic images (cylindrical ODIs) has been proposed, where conventional saliency models have been used on the planar representation of the ODIs. In [18], a spherical saliency model has been presented, where different visual features are computed on the sphere. However, these works are based on an early model proposed in [10], that only considers low level features. Many saliency models have been recently proposed, showing good performance [16]. Thus, adapting state-of-the-art models to the needs of ODIs would benefit from the developments already made in this area.

Our work targets the prediction of saliency maps for ODIs from data collected through subjective tests. We propose a testbed that collects users VCTs and a processing algorithm that generates saliency maps from the gathered data. Moreover, we introduce a post-processing adaptation of state-of-the-art saliency models for ODIs, namely FSM, and we study the effect of ODIs exposition time in subjective experiments. Our database and testbed are available to promote the study of visual attention for ODIs.



(a) (b)

Fig. 1: (a) Sphere mesh (polar angles θ and ϕ). (b) Top: equirectangular ODI and user viewport (area delimited by the hexagon). ODI: Maasboulevard festival 2006, author: Aldo Hoeben (*Fair*). Bottom: VCT of a participant viewing *Fair* for 10s.

III. Collection of Viewport Center Trajectories

Through subjective tests, viewport information is collected. In particular, in the absence of an eye tracker device, we assume that the center point of the viewport represents the visual target location of the user at any time. This is supported by two facts [19]: (I) visual acuity is at its maximum at the center of the human visual field or fovea, and (II) the head tends to follow eye movements to preserve the eye resting position (eyes looking straight ahead). The collected viewport center locations in time for an ODI and a user is what we call *viewport center trajectory (VCT)*. The bottom image in Fig. 1b shows the VCT for a participant and the ODI *Fair*. In this section, we first describe the implemented testbed that allows the successive display of ODIs while it gathers the participants VCTs. Then, the subjective test is explained in detail.

A. Testbed

The designed testbed is a software application that permits the display of ODIs in an HMD while it collects the VCTs of the participants. The testbed has been implemented using the WebVR [20] and the ThreeJS [21] APIs. The former enables cross-platform compatibility, meaning that our testbed can be used in a variety of VR headsets; *e.g.*, Oculus Rift, HTC Vive, Samsung Gear VR, or Google Cardboard. The latter enables the creation of fully immersive 3D experiences in a browser, allowing us to display the different ODIs without the use of a specific software other than a web browser. In our subjective tests, we used the Oculus Rift DK2 as the HMD and the Firefox Nightly builds as the web browser.

The input of our testbed is the ODIs in their planar representation. Here, we consider the equirectangular planar format as it is the most common output of the 360° cameras. To model a spherical view from an ODI using ThreeJS, a scene, a camera and a few geometries and textures need to be defined. In our scene, a virtual camera is positioned in the center of a 3D sphere mesh (the geometry), which is then wrapped with an ODI in equirectangular format (texture). As we need to display the ODI on an HMD, everything needs to be rendered twice, one view for each eye, with a slight offset to account

Fig. 2: Dataset sample. From left to right. First row: Bellagio Hotel Lobby, author: Bob Dass (*Lobby*); Wedding: Ceremony, author: Aldo Hoeben (*Wed*); Sziget 2008: Solving a maze, author: Aldo Hoeben (*Game*). Second row: Abri de Maguenay, author: Alexandre Duret-Lutz (*Hike*); First, author: Alexandre Duret-Lutz (*Cows*); Cathédrale de Bordeaux, author: Alexandre Duret-Lutz (*Church*)

for the distance between the eyes. In addition, to determine the fraction of the ODI rendered given the head orientation of the user, the application queries the headset for the field-of-view (FOV) of each eye. For instance, Oculus Rift DK2 has a maximum FOV of $100°$ per eye, meaning that around 30% and 55% of the horizontal and vertical domain of the ODI can be shown in the user viewport. The proposed testbed is able to collect the points that delimit the user viewport and its center point (top image in Fig. 1b). This is done at every frame refresh, which depends on the device specifications. For instance, Oculus DK2 has a maximum refresh rate of 75 Hz, meaning 13.33 ms per frame. The collected points are found in the planar format of the ODI, by performing a mapping from the 3D surface to its 2D planar representation. This information, together with the time stamp of each collected point is stored in a CSV file. The built testbed can be easily extended to other ODIs planar representations, as well as, to present omnidirectional videos instead of ODIs. Moreover, the testbed also allows creating customizable instructions to be shown to the participants during the subjective test.

B. Subjective test

We have considered 21 indoor and outdoor ODIs in equirectangular format with the recommended 4096x2048 resolution [22]. These images have been downloaded from the Equirectangular group [23] of the social photography site Flickr. We have only considered ODIs under the Creative Commons (CC) license. Figure 2 shows a sample of the ODIs considered.

We conducted subjective tests under task-free condition, *i.e.,* participants were only asked to naturally look at the ODIs. A total of 32 participants took part of our subjective test and an identifier was associated with each participant, to keep their anonymity. Participants were split into two groups consisting of 16 participants each, and each ODI was presented for $10s$ to one group and $20s$ to the other group. Both experiments were split into a training session and a test session. During the training session, an additional ODI, representative of the content, was presented. Then, during the test session 21 ODIs were randomly displayed while the VCT was collected. Similar to [5], we discarded the first fixation recorded for each ODI, as it adds trivial information on the starting viewing direction. Fixations for ODIs and viewport based data are formally defined in the following section.

IV. Processing of Viewport Center Trajectories

We are interested in obtaining saliency maps of ODIs in their planar representation. In particular, in the equirectangular planar format. In this Section, we first describe how VCTs are processed into saliency maps. Then, since VCTs are extracted from the 3D geometry and projected on the planar ODIs, we model the pixel deformation in the equirectangular format. This is of great importance when processing the VCTs into saliency maps.

A. From VCTs to saliency maps

Let us first consider the classical two types of eye movements: (I) *saccades*, when the eye rapidly moves from one visual target location to another, and (II) *fixations*, when the eye fixes on the point of interest for a certain period of time, typically around $200ms$ [19]. During a saccade, the eye produces the first movement towards the visual target location and then the head follows [19]. Thus, in this work we assume that the *viewport's center translation* from one visual target location to another is motivated by a saccade. In addition, we generalize the term fixation to refer to the maintaining of the viewport's center on a "single" location on the ODI; otherwise, the term eye fixation is used.

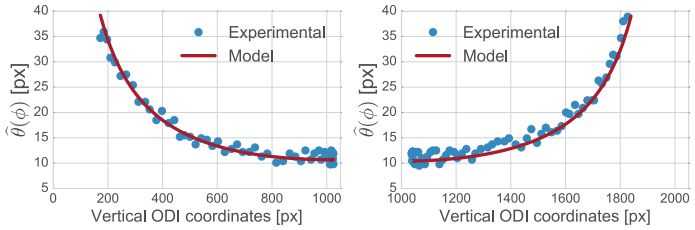
Due to the higher quality of visual information in fixation periods, we disregard viewport's center translations and sequentially process the VCTs to get the fixations. To avoid neglecting minor involuntary head movements during a fixation, the VCT of each participant and ODI is clustered according to a predefined threshold τ. In this work, the DBSCAN clustering algorithm [24] is used and the τ parameter is fixed to $2°$ of any head rotation. Each cluster is a fixation if it has a number of points equivalent to at least $200ms$. The final fixation map of an ODI for N participants is given by: $F_i = 1/N \sum_{n=1}^{N} f_{n,i}$, where $f_{n,i}$ is the fixation map for participant n and ODI i.

To obtain a saliency map for an ODI from a fixation map, we need to consider two facts [19]: (I) gaze shifts smaller than $10°$ can occur without the corresponding head movement, and (II) there is a gradually decreasing acuity from the foveal vision (center point of focus) towards the peripheral vision. Thus, we apply a Gaussian filter G_σ to the fixation map F_i of a particular ODI i, resulting in the final saliency map: $S_i = F_i * G_\sigma$, where σ stands for the standard deviation of the Gaussian filter. Note that, a Gaussian filter follows the $68-95-99.7$ rule, where values located between +/- σ, 2σ and 3σ account for the 68%, 95% and 99.7% data of the set, respectively. Thus in this work, σ is set to $5°$, to account for the $10°$ of small gaze shifts (in 2σ) and the decreased visual acuity from the fovea (in 3σ).

Note that both the τ and σ parameters are given in degrees. Thus, we need to find their projected values on the planar ODI.

B. Distortion in an equirectangular projection

In an equirectangular projection, polar angles from the sphere, $[-\pi \leq \theta \leq \pi]$ and $[-\pi/2 \leq \phi \leq \pi/2]$, are directly used as the horizontal and vertical coordinates of the ODI planar representation (Fig. 1). In this projection, there is a constant vertical sampling density, as the longitudes of the sphere, defined by θ, have the same length from the south to the north pole of the sphere. This is not true for the horizontal sampling, where the circles of latitudes, defined by ϕ, become smaller as one distances oneself from the equator. This leads to a horizontal stretching of the pixels as one moves from the equator, where there is no distortion, to the poles, where distortion tends to infinity.



(a) (b)

Fig. 3: Spherical to planar projection $\widehat{\theta}(\phi)$ of $\theta = 1°$ when the HMD moves: (a) from the south-pole to the equator (from 0 to 1024 px in the planar ODI) and (b) from the equator to the north-pole (from 1024 px to 2048 px in the planar ODI). ODI size: 4096×2048.

Let us consider a head rotation in θ with $\phi = 0$, *i.e.*, the user viewport is centered on the equator of the sphere. In this case, the projection of θ on the equirectangular ODI is given by: $\widehat{\theta}(\phi = 0) = (W \times \theta)/2\pi$ pixels, where W stands for the width of the planar ODI. In the case of $\phi \neq 0$, the horizontal pixel stretching is estimated by using the circumferences of the circles of latitudes, defined by $C(\phi) = 2\pi R cos(\phi)$, where R is the sphere radius. Then, the pixel's stretching $P(\phi)$ is defined by the ratio between the circumference at the equator, where $\phi = 0$, and the circumference at a particular latitude ϕ:

$$P(\phi) = \frac{2\pi R}{2\pi R cos(\phi)} = \frac{1}{cos(\phi)} \tag{1}$$

Then, the projection of θ on the planar ODI is given by:

$$\widehat{\theta}(\phi) = \widehat{\theta}(\phi = 0) \times P(\phi) \tag{2}$$

The projection of $\theta = 1°$ from the sphere to an equirectangular ODI of resolution 4096×2048 is illustrated in Fig. 3. Figures 3a and 3b show the pixels' stretching caused by moving the head upwards from the south pole to the equator and from the equator to the north pole, respectively. The scatter plot represents simulated values, while the continuous line is the model approximation from (2).

V. Fused Saliency Maps

Available saliency models for rectilinear images do not address the specific characteristics of ODIs. Unlike traditional images, ODIs are presented as continuous spherical images. Thus, considering a center prior feature, as most saliency models do, is not relevant. Differently, a clear equator bias is seen in ODIs, as viewers are more likely to view content in the equator adjacency rather than in the poles (Fig. 1b). This is an expected user's behavior, as in typical conditions the head is held erect (not tilted backwards or forwards) and photographers tend to frame the object(s) of interest in the equator proximities. This feature may not be easily included in current saliency models, notably in models based on deep neural networks [15] [14] where a large dataset is needed for training.

To deal with the center prior limitation of current saliency models, we propose the *fused saliency maps* (FSM) post-processing method. In FSM, saliency maps are predicted for translated versions of the original planar ODI i using any saliency model designed for traditional images. Then, these saliency maps are translated back into the original ODI setting, denoted as $S_{i,t}$, and linearly combined. Formally, the resulting saliency map is defined as:

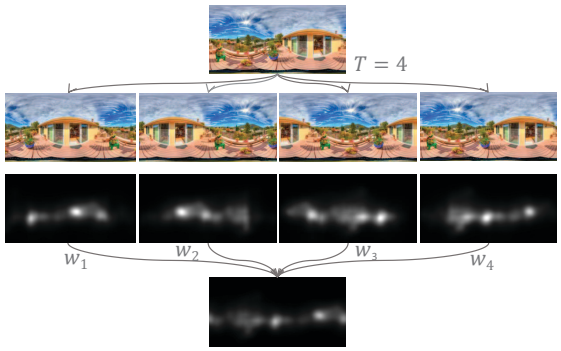


Fig. 4: FSM post-processing method. From top to bottom: (1) Original planar ODI, (2) translated ODIs (T=4), (3) translated back saliency maps $S_{i,t}$ (Salicon model [15]) and (4) FSM output: fused saliency map. ODI: The Porch, author: Nick Hobgood (*Porch*).

$$S_i^{FSM} = \sum_{t=1}^{T} w_t S_{i,t} \tag{3}$$

where T denotes the total number of translations applied to the ODI and w_t assigns the weight to the particular $S_{i,t}$. In this work, we consider an equal weight distribution; however, these weights could be optimized according to the characteristics of the ODI. Figure 4 illustrates our proposed FSM method when the *Salicon* saliency model [15] is used. Note how the saliency maps, $S_{i,t}$, have a strong center bias. This center bias is reshaped into an equator bias in the fused saliency map, *i.e.*, the output of FSM.

VI. Experimental Results

A. ODI exposition time

In ODIs applications, users do not have visual access to the full image. Thus, it is not clear for how long they should be shown in subjective tests to collect participant's fixations. In order to study the effect of the exposition time, we showed each ODI for $10s$ (Test10) or $20s$ (Test20) to 32 participants, with 16 participants for each test.

The median value of the horizontal distance traveled by the participants for Test10 and Test20 is illustrated in Fig. 5, when the maximum distance is 4096 pixels (ODIs width). This figure shows that the increase of the duration of the experiment does not mean an equal increase on the horizontal distance traveled. Meaning that a duration of $20s$ does not necessarily lead to new fixations. Participants tend to travel slower through the ODI and/or go back to previously visited locations when they are given $20s$ instead of $10s$ to look around. Figure 8 shows some examples of saliency maps obtained in the two tests.

The Test10 and Test20 were also compared in terms of the entropy of resulting saliency maps (Fig. 6). A similar trend can be seen between the two curves, meaning that regardless of the exposition time, for some ODIs the fixations tend to be concentrated in the same locations (low entropy), while for others, the fixations tend to be spread out (high entropy).

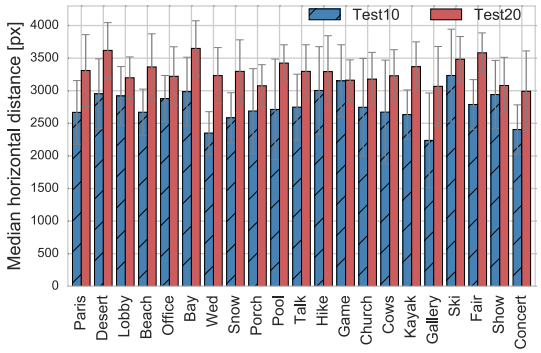


Fig. 5: Median horizontal distance traveled by participants in Test10 and Test20. Median absolute deviation showed on top of each bar.

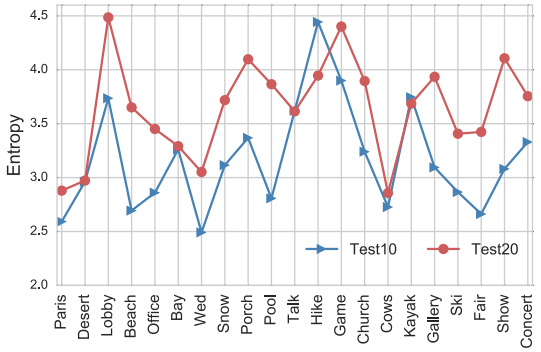
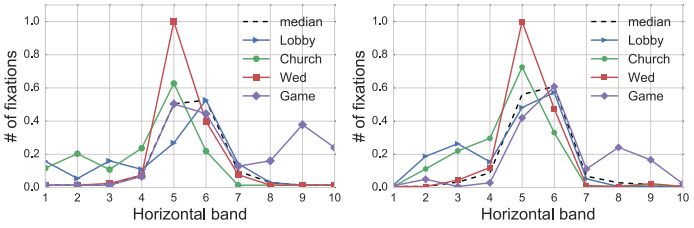


Fig. 6: Saliency maps entropy for Test10 and Test20.

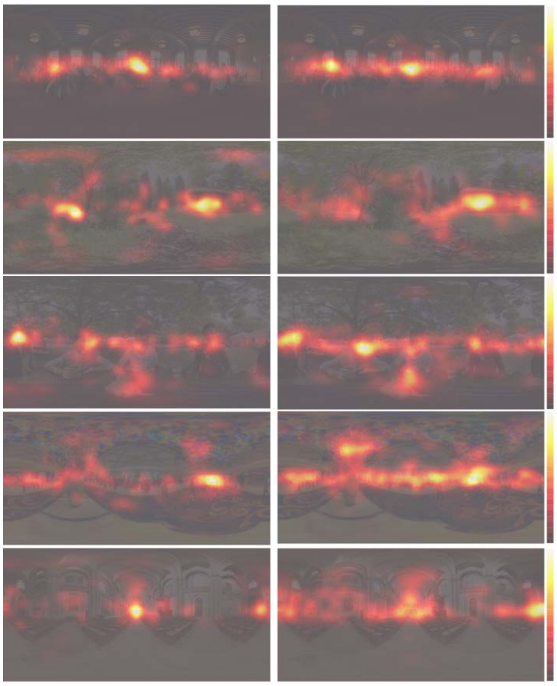
In Fig. 8, the saliency maps for the ODIs *Wed* and *Hike* are shown, which correspond to the saliency maps with the lowest and the highest entropy in Test10, respectively. In general, the entropy is higher in Test20, which demonstrate that the similarity between fixations of different participants is higher at the beginning of the ODI exposition, producing a saliency map with lower entropy. This similarity decreases over time, resulting in a saliency map with higher entropy.

To study the distribution of fixations in the considered ODIs, we equally split each planar ODI along its vertical axis into horizontal bands. For illustration, we consider 10 bands, each with size 4096×205 pixels. Figure 7, shows the distribution of the number of fixations (normalized) for each horizontal band, starting from the top of each ODI. With a dashed line, we show the median value of the number of fixations for the 21 ODIs, which illustrates their strong tendency of being located in the vicinity of the equator. We also show the distribution of fixations for some particular ODIs. The ODI *Wed*, shows a strong tendency of the fixations to be around the equator. On the other hand, ODIs *Lobby*, *Church* and *Game* have appealing features at the top (*Lobby* and *Church*) and at the bottom of the ODI (*Game*). However, even if fixations are more distributed for these ODIs, the highest number of fixations are still concentrated in the equator vicinity. Moreover, a similar trend on the distribution of the number of fixations in the different ODIs can be seen for Test10 and Test20. Thus, even if the participants had more time to look around, the proportion of time spent on each horizontal band is comparable. This is another evidence that longer exposition time does not lead to relevant new fixations. Therefore, in the remaining of this paper, we only consider the subjective results for Test10.



(a) Test10 (b) Test20

Fig. 7: Fixations distribution for each ODI horizontal region.



(a) Test10 (b) Test20

Fig. 8: Saliency maps for Test10 and Test20. ODIs from top to bottom: *Wed*, *Hike*, *Game*, *Lobby*, *Church*

B. *Saliency models performance*

Motivated by the strong tendency of fixations to be located in the vicinities of the equator (Fig. 7), in this section we evaluate the performance of the proposed FSM method (Section V) that targets the capture of this ODI feature. In particular, the predicted saliency maps from subjective test data (Test10) are used as *ground truth*. Then, the performance of current saliency models, with and without our FSM method, is evaluated by measuring their accuracy on predicting the ground truth. The performance is quantified using the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) [25]. We use the AUC-Borji [16] variant of AUC. We consider the models *Salicon* [15], *Deep Gaze2* [14] and *Judd* [5] due to their proven good prediction accuracy [16]. In the case of *Judd* model, we replace the center prior by an equator prior, to indicate the distance of each pixel to the equator of the ODI. We denote this modified version of the model as *Judd-E*. Our FSM method is applied with $T = 4$.

From the ROC curve in Fig. 10, the following remarks can be drawn. (I) The saliency models with FSM always outperform the original models. In particular, *Salicon & FSM* model had

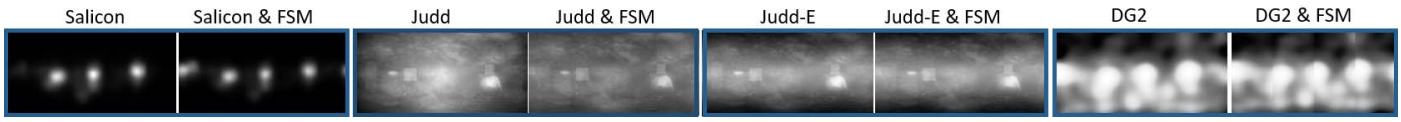


Fig. 9: Saliency maps estimated by *Salicon*, *Deep Gaze2*, *Judd* and *Judd-E* models with and without the FSM method. ODI: (*Game*).

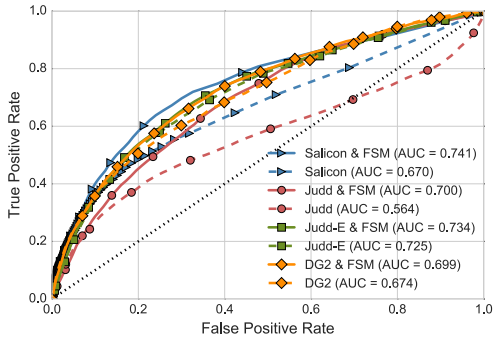


Fig. 10: ROC curves and AUC values averaged over all 21 ODIs.

the best performance overall with an AUC = 0.74, while the largest AUC gain over an original model is observed for the *Judd & FSM* model, with a gain of 20%. (II) The *Judd-E* model (with and without FSM) notably outperforms the original *Judd* model. Note that FSM provides a small improvement on the *Judd-E* model; meaning that both the FSM method and the replacement of the center prior by an equator prior have comparable impacts on the original model. Figure 9, shows the saliency maps for the *Salicon*, *Deep Gaze2*, *Judd* and *Judd-E* models with and without the FSM method for ODI *Game*, which has been shown in Fig. 2.

VII. Conclusion

In this work, we estimated saliency maps for ODIs viewed in HMDs. We developed a testbed and a database of VCTs for 32 participants and 21 ODIs. To estimate saliency maps from the collected data, we proposed a method that considers both the 3D geometry and the planar representation of an ODI. Moreover, we studied the effect of ODI exposition time during subjective tests for saliency estimation, where ODIs were shown for $10s$ or $20s$. Tests results showed that increasing the exposition time from $10s$ to $20s$, did not always lead to new fixations, concluding that between the two options, $10s$ was a reasonable ODI exposition time. The optimization of this parameter is left for future work. In addition, we illustrated the equator bias tendency in ODIs, which motivated the use of our proposed FSM method. We showed that the proposed FSM improved current models performance by up to 20%.

References

[1] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE Int. Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, Sept 2015.

[2] ——, "Content adaptive representations of omnidirectional videos for cinematic virtual reality," in *ACM Multimedia Conf.*, Brisbane, Australia, Oct 2015.

[3] 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*, ser. ATC '16, New York, NY, Oct 2016, pp. 1–6.

[4] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," *ArXiv e-prints*, Oct 2015.

[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE Int. Conf. on Computer Vision*, Sept 2009.

[6] 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.

[7] F. D. Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding," in *Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec 2016.

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

[9] X. Corbillon, F. D. Simone, and G. Simon, "360-degree video head movement dataset," in *ACM Multimedia Systems*, Taipei, Taiwan, June 2017.

[10] L. Itti, C. Koch, E. Niebur *et al.*, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[11] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proceedings of the British Machine Vision Conf.*, Sept 2011, pp. 110.1–110.12.

[12] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Int. Conf. on Computer Vision*, Barcelona, Spain, Nov 2011.

[13] J. Wang, A. Borji, C. C. J. Kuo, and L. Itti, "Learning a combined model of visual saliency for fixation prediction," *IEEE Trans. on Image Processing*, vol. 25, no. 4, pp. 1566–1579, 2016.

[14] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Deepgaze II: Reading fixations from deep features trained on object recognition," *ArXiv e-prints*, Oct 2016.

[15] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE Int. Conf. on Computer Vision*, Santiago, Chile, Dec 2015.

[16] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. [Online]. Available: http://saliency.mit.edu/results_mit300.html

[17] A. Bur, A. Tapus, N. Ouerhani, R. Siegwar, and H. Hiigli, "Robot navigation by panoramic vision and attention guided fetaures," in *Int. Conf. on Pattern Recognition*, Hong Kong, Aug 2006.

[18] I. Bogdanova, A. Bur, and H. Hugli, "Visual attention on the sphere," *IEEE Trans. on Image Processing*, vol. 17, no. 11, pp. 2000–2014, 2008.

[19] O.-J. Grüsser and U. Grüsser-Cornehls, "The sense of sight," in *Human Physiology*. Berlin, Heidelberg: Springer, 1983, pp. 237–276.

[20] WebVR API. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/WebVR_API

[21] ThreeJS API. [Online]. Available: https://threejs.org/

[22] Oculus 360 photos support. [Online]. Available: https://support.oculus.com/help/oculus/866319816819547

[23] Equirectangular flickr group. [Online]. Available: https://www.flickr.com/groups/equirectangular/

[24] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Conf. on Knowledge Discovery and Data Mining*, Portland, OR, Aug 1996.

[25] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *ArXiv e-prints*, Apr 2016.