A SIMPLE METHOD TO OBTAIN VISUAL ATTENTION DATA IN HEAD MOUNTED VIRTUAL REALITY

Evgeniy Upenik and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

ABSTRACT

Automatic prediction of salient regions in images is a well developed topic in the field of computer vision. Yet, virtual reality omnidirectional visual content brings new challenges to this topic, due to a different representation of visual information and additional degrees of freedom available to viewers. Having a model for visual attention is important to continue research in this direction. In this paper we develop such a model for head direction trajectories. The method consists of three basic steps: First, a computed head angular speed is used to exclude the parts of a trajectory where motion is too fast to fixate viewer's attention. Second, fixation locations of different subjects are fused together, optionally preceded by a re-sampling step to conform to the equal distribution of points on a sphere. Finally, a Gaussian based filtering is performed to produce continuous fixation maps. The developed model can be used to obtain ground truth experimental data when eye tracking is not available.

Index Terms— visual attention, fixation maps, omnidirectional visual content, virtual reality, 360-degree images and video

1. INTRODUCTION

Omnidirectional visual content or cinematic virtual reality is a technology which provides immersive experience to viewers by displaying still images or video with a full spherical coverage of the field of view. Content of such type is acquired with special devices performing a particular work-flow. Certain omnidirectional acquisition systems, e.g. multi-lens and catadioptric cameras, produce ready-to-display images or video, whilst others, e.g. multi-camera systems, require an additional step of off-line stitching. The latter combines signals from several image sensors into a panoramic planar representation, such as an equirectangular or cubic projection. Omnidirectional images and video are typically consumed using a virtual reality (VR) head-mounted display (HMD). Visual content represented in one of the projections is rendered on a viewport of an HMD where data from acceleration and orientation sensors is used to define which part of the content is to be displayed. This data, if stored, can then be used for analysis of human visual attention in VR imaging.

Computational prediction methods for human visual attention have been studied for decades in conventional flat images. The first theoretical computational model of human visual attention was introduced by Koch and Ullman in [1], and the first practical implementation was presented by Clark and Ferrier in [2]. Detailed descriptions and classifications of state-of-the-art visual attention models can be found in [3–5]. There exist two main approaches for modeling human visual attention, namely, bottom-up and top-down. The former starts by computing different features in images, typically intensity, color and orientation characteristics. These features are then fused together to produce a saliency map. The latter approach takes into account certain high level information about the scene which is used, for example, by incorporating face, object, and text detection. Top-down methods are often combined with bottom-up models.

Visual attention for spherical images has been studied in [6, 7]. Bogdanova et al. propose bottom-up methods to obtain saliency maps from omnidirectional images for static and dynamic cases. Features are computed and fused in a spherical domain. However, these studies do not provide any detailed descriptions about interpretation of experimental visual attention data for omnidirectional images.

Experimental visual attention data, unlike prediction models, does not provide saliency maps. After initial processing, one can obtain fixation locations, i.e. points in the image where observers fixated their attention. This data can be further processed to produce continuous fixation maps. The first step is to analyze eye movements using one of the methods based on velocity and distance criteria. Methods to obtain fixation locations are described in [8–10]. Typically the next step is to produce a continuous fixation map by applying to

This work has been conducted in the framework of ImmersiaTV project under the European Unions Horizon 2020 research and innovation program (grant agreement no. 688619) and funded by Swiss State Secretariat for Education, Research and Innovation SERI.

fixation locations a Gaussian filter with a certain standard deviation corresponding to the high acuity vision area [11].

In VR environment, in addition to eye movements, observer's head direction must be taken into consideration. One can find studies on eye-head coordination in humans during different tasks in [12, 13]. The main findings in these studies support a hypothesis that the human eye movement range is restricted not physiologically but neurologically and this range is narrower when a subject's head is not fixed. Nonetheless, there is no commonly adopted model for interpretation of eye-head position data in visual attention fixation maps for omnidirectional visual content.

In this paper, we propose a simple approach to treat raw experimental head direction trajectories in virtual reality content. The proposed approach implies three basic steps: First, a computed head angular speed is used to exclude the parts of a trajectory where motion is too fast to fixate viewer's attention. Second, fixations of different subjects are fused together. If needed, this step is preceded by re-sampling track coordinates in order to conform to the equal distribution of points on a sphere. Finally, a Gaussian based filtering is performed to produce continuous fixation maps.

2. EXPERIMENTAL DATA

The data used in the present work has been obtained during a subjective quality evaluation experiment [14] on omnidirectional images. For the current study we selected only the head direction tracks recorded from unimpaired stimuli.

2.1. Experiment

Figure 1 depicts the contents used in the experiment. Observers were asked to assess visual quality of four different omnidirectional images represented in the equirectangular projection and compressed with different quality parameters and different codecs. In particular, viewers were instructed to search for compression artifacts. Overall, 40 subjects participated in the experiment, 25 male and 15 female subjects, between 18 and 32 years old with the average and the median of 24.9 and 24.8, respectively. All participants were tested for correct color vision and visual acuity using Ishihara and Snellen charts respectively.

The experiment was conducted using the testbed for subjective evaluation of omnidirectional visual content proposed in [15]. This software has been developed for iOS and is publicly available for download[1]. During the experiment, subjects were wearing an HMD composed of a VR head-mount with buttons[2] and a mobile device installed inside as a screen. iPhone 6 was used to display the images. The overall resolution of the phone's screen is 1334×750 pixels, which gives 667×750 pixels per eye. The vertical field of view provided by the hardware-software solution is 90 degrees, which corresponds to 8.33 pixels per degree. All the subjects were sitting on a rotatable chair during the experiment.

[1] https://github.com/mmspg/testbed360

[2] https://mergevr.com

(a) Harbor (b) KiteFlite

(c) PoleVault (d) SkateboardTrick

Fig. 1: Omnidirectional images used in the experiment

2.2. Head direction tracks

Raw data of a head direction trajectory contains an array of yaw and $pitch$ coordinates along with their time-stamps. The tracks were recorded for each assessed stimuli, however only the trajectories obtained from unimpaired images have been selected for the current study. Each presented content has head direction tracks from 40 subjects. Two seconds of data in the beginning of each track were dropped. This has been performed in order to compensate the initial head position impact on calculating user gaze fixations.

3. FIXATION LOCATIONS

An angular velocity of observer's head evidently impacts his ability to fixate attention. Although the fact that human visual perception depends on motion is well known, the impact of the ocular-vestibular reflex can decrease its effect. Nonetheless, here we assume there exits a threshold head angular velocity beyond which users are not able to focus their attention on any object. The value of 15 degrees per second has been chosen as the upper boundary for this paper. However, it is a parameter and more experimental data is needed to determine the optimal threshold angular velocity.

3.1. Head angular velocity

Observer head position is a vector $[\theta\ \phi]$, where θ and ϕ represent yaw and $pitch$ respectively. Values of yaw and pitch in degrees over time are presented in Figure 2 (top and middle). In order to obtain head angular velocity we compute a first

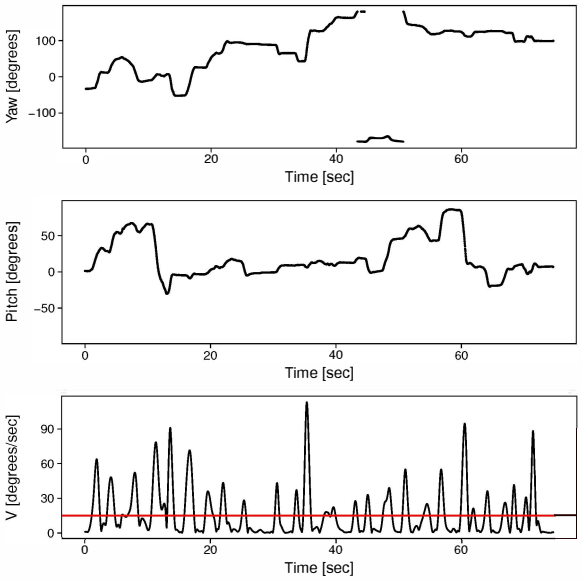
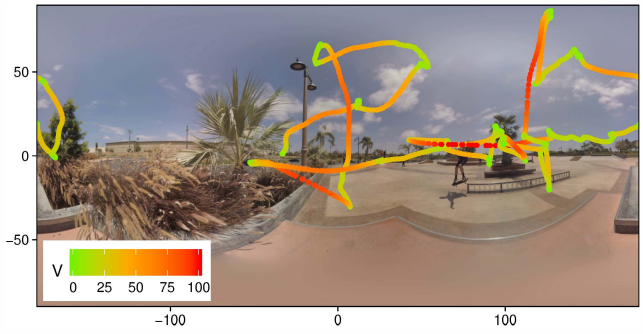
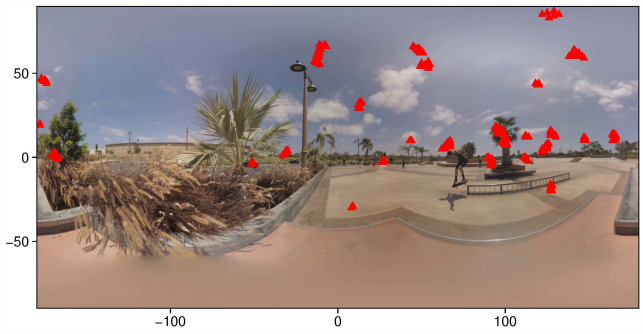


Fig. 2: Yaw (top) and pitch (middle) of viewer's head direction trajectory. Head angular velocity (bottom), red horizontal line depicts the threshold



(a) Head motion trajectory



(b) Viewer's fixations

Fig. 3: A typical head motion trajectory colored with its angular velocity in degrees per second (top), and fixation locations obtained from it (bottom).

order derivative of the following vector:

$$\mathbf{V}_{ang} = \begin{bmatrix} V_\theta \\ V_\phi \end{bmatrix} = \begin{bmatrix} \frac{d\theta}{dt} \\ \frac{d\phi}{dt} \end{bmatrix}$$

Considering only the velocity magnitude, the norm of the vector is taken as:

$$\|\mathbf{V}_{ang}\| = \sqrt{\left(\frac{d\theta}{dt}\right)^2 + \left(\frac{d\phi}{dt}\right)^2}$$

The yaw and pitch data is represented in digital format. Thus we compute a derivative using a standard method of numerical differentiation. For each signal sample the difference with its next value is obtained and divided by the sampling period:

$$s'_n = \frac{s_n - s_{n-1}}{T_{sampl}}$$

Then a 2nd order Butterworth low-pass filter with cutoff frequency of f_c = 2 Hz is applied separately to V_θ and V_ϕ in order to remove digital differentiation noise. We use a forward-backward numerical implementation of the filter to avoid a group delay in the signal [16]. The resulting head angular velocity over time is depicted in Figure 2 (bottom). All the head direction trajectory data with speed above the threshold (red line in Figure 2) is discarded from further analysis.

Figure 3 (a) shows a typical head direction trajectory. The color of the trajectory reflects the head angular velocity. Only the regions colored with green are considered as fixations of attention.

3.2. Equal distribution of points on sphere

There exist cases when after the head angular velocity restrictions, a resulting track requires an additional step of processing before becoming a set of viewers' fixations. Depending on the device used to obtain the raw data, the discrete domain of coordinates can distribute points in a non-equidistant manner on the surface of a sphere. If so, a re-sampling needs to be performed on the data in the following way.

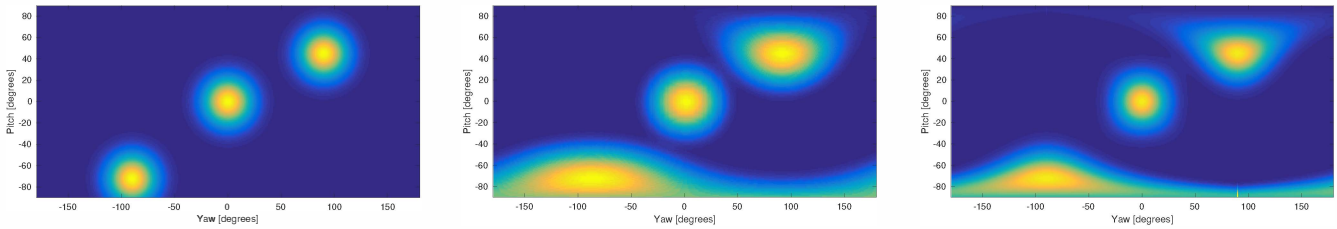
For each latitude level one re-samples the longitude signal $s(n)$ defined on $n \in N$ to the signal $g(m)$ defined on $m \in M$, where $M = N\cos(\phi)$ and $\phi \in (-\pi/2, \pi/2)$.

Resulting head fixation locations for a typical trajectory of one observation are depicted in Figure 3 (b).

3.3. Fusion

Fixation locations obtained from different subjects must be fused together in oder to derive statistical information. One can propose several ways to perform a fusion:

1. Add all the points from each subject as unity values to a resulting fixation set.



(a) Gaussian kernel (b) Gaussian kernel in viewport domain (c) Modified Gaussian kernel

Fig. 4: (a) Gaussian filter applied in equirectangular domain. (b) Gaussian in viewport domain. (c) Modified Gaussian proposed in the paper. Filters are applied to equirectangular image containing three unity points at (-90,-72), (0,0), and (90,45) degrees.

2. Sum-up all the points in cells with specified size producing a weighted set.
3. Only add points if a certain percent of subjects fixated in this particular location or a predefined area around it.

We use the second method to produce fixation locations further in the present work because of its moderate computational complexity.

4. CONTINUOUS FIXATION MAP

Fixation location data does not typically allow to properly depict the regions of visual attention. Because of its discrete nature, this information is not consistent even among human subjects. Indeed, very rarely a person will fixate their attention in the same exact point as another. Thus there is a need to introduce a statistical areas of fixations. For conventional images typically a Gaussian filter is applied to model a human acuity vision region of 1-2 degrees. In case of head direction fixations we assume that the region of possible attention is 30 degrees. Same as in Section 3 this value is a parameter and may be changed after further experimentations.

4.1. Gaussian filter in viewport domain

Omnidirectional content is consumed using an HMD. An observer sees a part of panoramic picture rendered in the viewport. Therefore, to highlight viewing area angle we need to apply Gaussian filter in the viewport domain.

$$G(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}}$$

where u and v are viewport coordinates. However, one normally works with an equirectangular or other panoramic representations of omnidirectional image or video. Thus, in the equirectangular domain, the kernel becomes:

$$G_{eqr}(\theta,\phi) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2(\theta,\phi)+v^2(\theta,\phi)}{2\sigma^2}}$$

Functions $u(\theta,\phi)$ and $v(\theta,\phi)$ are calculated as follows:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & \frac{k}{x} & 0 \\ 0 & 0 & \frac{m}{x} \end{bmatrix} R_Z^{\beta} R_Y^{\alpha} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \Bigg|_{x>0}$$

where k and m are the scaling coefficients for viewport coordinates, R_Z^{β} and R_Y^{α} are rotations for yaw and pitch respectively, and vector $[x\ y\ z]$ represents Cartesian coordinates of a point on the image sphere:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r\sin\phi\cos\theta \\ r\sin\phi\sin\theta \\ r\cos\phi \end{bmatrix}$$

The result of applying kernel $G_{eqr}(\theta,\phi)$ to filter the image directly in the equirectangular format is shown in Figure 4 (b).

Another approach to perform Gaussian smoothing in an equirectangular picture is to apply the filter in the rendered viewport and then project it back. However, the drawback of this method is the interpolation noise added during the transformations.

4.2. Modified Gaussian kernel in equirectangular domain

The method of filtering proposed in subsection 4.1 is computationally very heavy. To simplify the calculations we propose a modified Gaussian kernel.

$$G_{mod}(x,y) = \frac{1}{2\pi\sigma_y^2} e^{-\frac{x^2}{2\sigma_x}} e^{-\frac{y^2}{2\sigma_y}}$$

where

$$\sigma_x = \frac{\sigma_y}{\cos(\phi)}$$

and σ_y is a constant value. In the denominator of normalization coefficient we use σ_y^2 instead of $\sigma_x\sigma_y$ to prevent the change of the amplitude with x.

Figure 4 (c) shows an equirectangular image filtered using kernel $G_{mod}(x,y)$. As can be seen by comparing Figures 4 (b) and (c), Gaussian filter in the viewport domain and Modified Gaussian kernel give visually similar results.

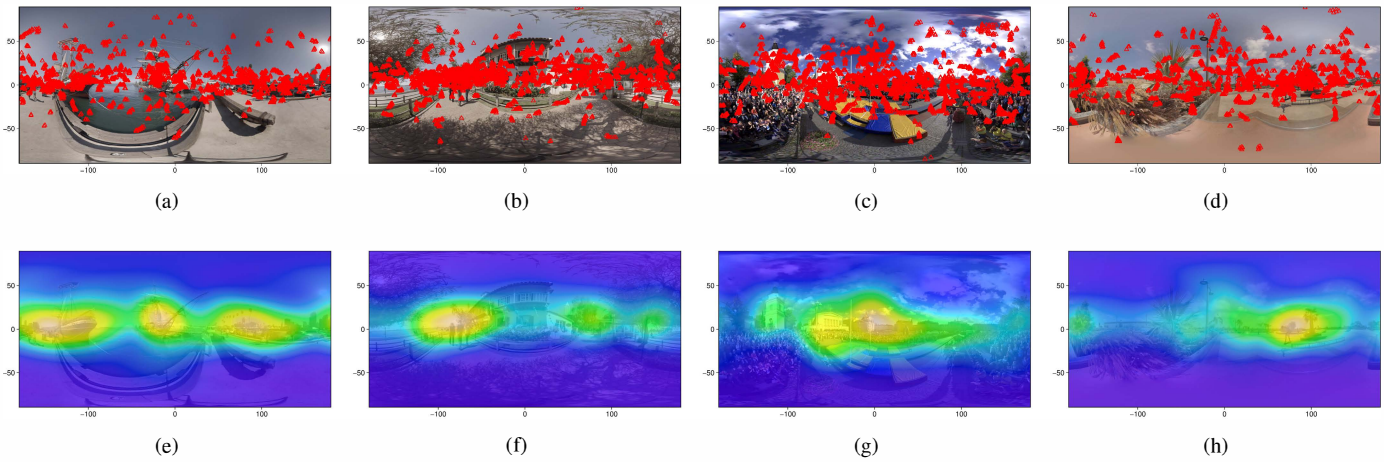


Fig. 5: Fixation locations (top row) and continuous fixation maps (bottom row) computed for the contents.

4.3. Generic statistical kernel in equirectangular domain

Faced with a lack of statistical data on eye-head relative movements, we assumed a Gaussian distribution of eye fixations around the center of a viewport. However, if we have such statistics it can be applied to form a kernel in the viewport domain:

$$K \equiv f(u, v)$$

where $f(u, v)$ is a probability density function on $(u, v) \in \mathbb{R}^2$, which can be estimated from statistical frequency distribution of eye fixations in the viewport by applying a regression to its two-dimensional histogram $m_{i,j}$ with k^2 the number of bins:

$$f(u,v)\Big|_{\substack{u=(i-k/2)w \\ v=(j-k/2)w}} \cong \frac{m_{i,j}}{\sum\limits_{i,j\in\mathbb{N}} m_{i,j}}$$

where $i, j \in \mathbb{N}$ are the indexes of each histogram bin, and $w \in \mathbb{R}^+$ is the bin width. The histogram is calculated as:

$$m_{i,j} = \sum_{\substack{(i-1-k/2)w<u_p\leq(i-k/2)w \\ (j-1-k/2)w<v_p\leq(j-k/2)w}} X[u_p, v_p]\Bigg|_{\substack{i\in[1,k] \\ j\in[1,k]}}$$

where $X[u_p, v_p]$ is the relative frequency distribution of fixation locations $(u_p, v_p) \in \mathbb{R}^2$, which are determined as a shift from the viewport center for $p \in [1, M]$, $p \in \mathbb{Z}$ and M is finite. The number of bins must be chosen according to one of the criteria described in [17, 18] depending on the distribution law.

Moving to the equirectangular domain can be performed as in Subsection 4.1:

$$K_{eqr}(\theta, \phi) = K(u(\theta, \phi), v(\theta, \phi))$$

A filter with the kernel $K_{eqr}(\theta, \phi)$ can be applied to fixation locations directly in the equirectangular domain.

5. RESULTS AND DISCUSSION

We apply the proposed approach to compute fixation locations and continuous fixation maps as interpretation of the raw experimental data described in Section 2. A head angular velocity threshold equal to 15 degrees per second is used. The Gaussian filtering is performed using $\sigma = 15$ in the base function. In order to fuse individual fixation locations, the points are summed up in cells of 1×1 degree. The modified Gaussian kernel $G_{mod}(x, y)$ is used to filter the data in the equirectangular domain. Figure 5 shows the fixation locations and the continuous fixation maps for four contents used in the experiment.

In the present work we apply Gaussian filtering in equirectangular projection. However, the proposed approach can be easily generalized to cope with other panoramic representations of omnidirectional visual content, such as cubic mapping and other convex polyhedron projections. Only the calculation of $u = u(x, y)$ and $v = v(x, y)$ must be changed to comply with a new projection.

In more theoretically oriented work [6,7], authors develop a mathematical model for Gaussian filtering in the geometry of the two-dimensional surface of a sphere. We consider these to be unnecessary complications, due to the fact that an observer sees only a rendered rectilinear viewport of an omnidirectional content and not the entire image. Thus applying vision range models in the viewport geometry appears to reflect better user experience and perception.

Head motion information is typically available without any additional cost during rendering of omnidirectional visual content in VR environments. For instance, during broadcast-

ing, a content provider can obtain anonymized head direction trajectory statistics of consumers. This information can be further used to adapt compression parameters when adaptive coding is applied. An example of such an adaptive coding method has been proposed in [19] for conventional images.

6. CONCLUSION

In this paper we have described a simple model to obtain fixation locations and continuous fixation maps from head direction trajectories for virtual reality content. The model incorporates analysis of a head angular velocity and provides the idea of a generic solution to produce continuous fixation maps for omnidirectional images represented in panoramic projections. Those fixation maps obtained from head position data can be a suitable first order approximation when eye tracking data is not available.

Furthermore, we applied the above approach to the raw experimental data and obtained the visual attention results for four omnidirectional images as a proof of concept.

Future work will focus on quantification of the approximation error when compared to true gaze detectors data. It may concern also refining the parameters of a threshold head angular velocity and statistical distribution of attention in a viewport by conducting additional subjective experiments. More specifically, generic statistical kernel for data smoothing can be estimated practically from experimental data.

7. REFERENCES

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*, no. 188, pp. 115–141, Springer Netherlands, 1987.

[2] J. J. Clark and N. J. Ferrier, "Modal control of an attentive vision system," *IEEE International Conference on Computer Vision*, 1988.

[3] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[4] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[5] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," *MIT Technical Report*, 2012.

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

[7] I. Bogdanova, A. Bur, H. Hügli, and P.-A. Farine, "Dynamic visual attention on the sphere," *Computer Vision and Image Understanding*, vol. 114, no. 1, pp. 100–110, 2010.

[8] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pp. 71–78, ACM, 2000.

[9] A. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Springer Science & Business Media, 2007.

[10] K. Holmqvist, M. Nystrm, R. Andersson, R. Dewhurst, H. Jarodzka, and J. v. d. Weijer, *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[11] O. L. Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251–266, 2013.

[12] D. Guitton and M. Volle, "Gaze control in humans: Eye-head coordination during orienting movements to targets within and beyond the oculomotor range," *Journal of neurophysiology*, vol. 58, no. 3, pp. 427–459, 1987.

[13] A. Doshi and M. M. Trivedi, "Head and gaze dynamics in visual attention and context learning," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 77–84, 2009.

[14] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in *9th International Conference on Quality of Multimedia Experience QoMEX 2017*, June 2017.

[15] E. Upenik, M. Rerabek, and T. Ebrahimi, "A testbed for subjective evaluation of omnidirectional visual content," in *32nd Picture Coding Symposium PCS*, Dec 2016.

[16] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 988–992, 1996.

[17] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.

[18] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.

[19] V. Hosu, F. Hahn, O. Wiedemann, S.-H. Jung, and D. Saupe, "Saliency-driven image coding improves overall perceived jpeg quality," in *32nd Picture Coding Symposium PCS*, Dec 2016.