Face-aware Saliency Estimation Model for 360° Images

Pramit Mazumdar
Department of Engineering
Roma Tre University
Rome, Italy
pramit.mazumdar@uniroma3.it

Giuliano Arru
Department of Engineering
Roma Tre University
Rome, Italy
giuliano.arru@uniroma3.it

Marco Carli
Department of Engineering
Roma Tre University
Rome, Italy
marco.carli@uniroma3.it

Federica Battisti
Department of Engineering
Roma Tre University
Rome, Italy
federica.battisti@uniroma3.it

***Abstract*—In this paper, a saliency estimation technique for omni-directional images is presented. Traditional approaches for estimating 360° image saliency rely on the exploitation of low and high-level image features, along with auxiliary data, such as head movement or eye-gazes. However, the image content plays an important role in saliency estimation. Based on this evidence, in the proposed method low-level features are combined with the detection of human faces. In this way it is possible to refine the saliency estimation based on the low-level features by assigning a larger weight to the regions containing faces. Experimental results on 360° image dataset show the effectiveness of the proposed approach.**

***Index Terms*—Omni-directional images, saliency, face detection, low-level features**

I. Introduction

In recent years, many efforts have been devoted to increase the immersivity feeling of a user watching an image or a video. In fact, the classical 2D image represents a flat projection of the real scene on a 2D planar support. The point of view is fixed, as well as the point of interest or the focused area.

The availability of novel technologies (i.e., plenoptic cameras, omni-directional cameras, or multiple cameras setup) allows a more realistic rendering of the scene giving the user the freedom of changing the point of view or focusing different areas of the scene. In this way, the quality of experience is improved [6], [7], [23]. Among the recording systems, the popularity of omni-directional (or 360°) cameras has been growing as demonstrated by the availability of consumer level, low-cost, acquisition devices. Despite the fact that an omni-directional content is captured, an observer can view only one portion of the scene at a time. The user browses the scene with the movement of the eyes and moves from one region to another by means of head and body movements. The rendering systems designed for 360° images, reproduce this mechanism by showing only one portion of the 360° image (the “viewport”) at a time and then changing the viewport content according to the user’s head movement. As can be noticed, the free head motion results in an experience closer to the real-life viewing behaviour.

With respect to 2D content, a 360° media requires larger storage space, heavier computational speed and larger bandwidth for transmission. Therefore, for compression and processing (e.g., denoising or enhancing) purposes, the understanding of the salient regions in a 360° image is essential. In literature, several methods have been developed for detecting salient regions in classical 2D images. However, studies specifically devoted to estimate saliency in 360° images are limited.

The most straightforward solution has been the application of 2D saliency predictors to 360° images. One example is in [24]. In this work the spherical content is projected to equi-rectangular format. This operation introduces distortions that the authors analyze by considering different interpretations of the equi-rectangular images (i.e., continuity-aware, cube map, and a combination of both).

An extension of 2D classical approaches is in [2] where the authors extract a 2D saliency map together with information on hue and saturation, and combine this information with the detection of the presence of human shapes for refining the overall saliency map. New approaches have been specifically designed for saliency estimation in omni-directional images. Fang et al. in [9] base their approach on the extraction of perceptual features from the CIE Lab color space. A contrast map is fused with a boundary connectivity map for estimating the saliency of omni-directional content. The methods proposed in [1], [10] include a semantic analysis of the scene to account for the content of the image in the task of saliency estimation.

Recently, deep learning schemes have been successfully applied for saliency estimation in panoramic images. He et al. in [14] investigate the use of CNNs (Convolutional Neural Networks) for saliency estimation. Similarly, in [20] the authors exploit CNNs to adapt 2D prediction to omni-directional images.

In this paper, we propose a saliency estimation model for 360° images. It is based on the fusion of features extracted from the visual content at different layers: low-level, high-level, and semantic level features. In more details, information on hue, saturation, luminance, 2D visual saliency and image entropy are combined with a semantic analysis of the content to identify the presence of human faces.

II. Proposed approach

The proposed Face attention and Low-level feature based omni-directional image Saliency estimation model (FLS) is

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

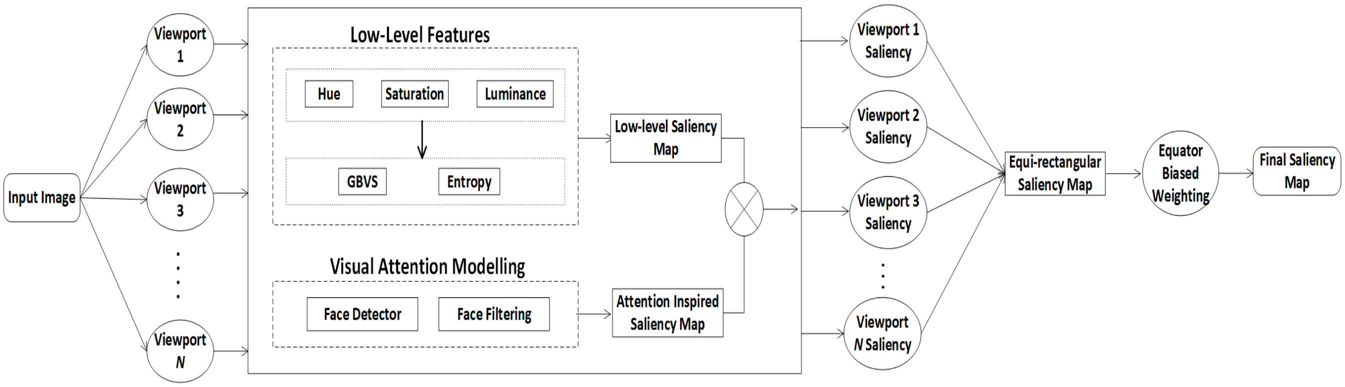


Fig. 1: Proposed Face attention and Low-level feature based omni-directional image Saliency estimation model (FLS).

shown in Figure 1. The following subsections describe each step involved in the FLS model.

A. Viewport Selection

An immersive experience can be obtained by viewing 360° images with a Head Mounted Display (HMD). Each HMD is characterized by a limited Field-Of-View (FOV) that controls the angle and exposure of the 360° content to the viewer. For example, HTC Vive and Oculus DK2 HMDs show 100° of FOV in both horizontal and vertical directions. Thus, a user wearing a HMD is required to move his/her head for exploring the entire 360° content. Similarly, for estimating the saliency of a 360° image, every possible FOV should be considered. The traditional approach partitions the equi-rectangular image into multiple windows (each one representing a possible FOV), estimates the saliency for each window and, finally, re-projects the saliency maps into final equi-rectangular saliency map [2], [18], [22].

The partitioning of an equi-rectangular image into small portions corresponding to the FOVs of a user while exploring a 360° content is a challenging task. This operation is generally performed in two steps. First, angular sampling is performed over the spherical image and each sampled point is considered as the centre of the estimated FOV. Then, windows representing the actual FOV are extracted and an inverse gnomonic projection on a rectangular plane is performed for each pixel. The extracted windows on rectangular plane with fixed width and height are called viewports. As depicted in Figure 1, in this work we first divide the input equi-rectangular image into a number of viewports. Then, each individual viewport undergoes the proposed saliency estimation method. Finally, the estimated saliency maps for each viewport are re-projected to the equi-rectangular plane. A detailed description of the viewport extraction and re-projection technique is presented in [2]. In the following subsections, we describe the proposed saliency estimation technique on each extracted viewport.

B. Low-level Features

Images can be described by exploiting different low-level features such as hue, saturation, edge, brightness, gradients, etc. Such features have been adopted in traditional 2D and, recently, in 360° image saliency estimation systems [9], [21]. In this work, the following low-level features have been used:

- **Hue (H), Saturation (S), Luminance (L):** H and S components provide a basic indication of color and better visual consistency than RGB components [16]. L refers to the overall brightness of an image. Since the human eye is more sensitive to brightness, L acts as a prominent feature for image saliency estimation [3], [8].

- **Graph-Based Visual Saliency (G):** Harel et al. propose a graph computation model for estimating 2D image saliency [12]. The bottom-up saliency estimation technique consists of three steps. First, a linear filtering is performed on the image to compute the feature maps (Gabor, contrast, and luminance). Subsequently, a Markovian based activation map is generated to highlight *unusual* regions in the image. Finally, a normalization step is performed on the generated activation maps. This approach is by far the most popular saliency estimation technique that exploits the low-level features [19], [24], [26].

- **Entropy (E):** E estimates the information in an image as the frequency of change in pixel intensities. Therefore, entropy can be used to measure the information content of an image. The Shannon entropy is considered in our work and it is computed for each viewport. The entropy of each viewport, V_i, is computed as:

$$E_{V_i} = -\sum_{x} p_x \log_2(p_x) \qquad (1)$$

where, p_x is the probability of occurrence of any pixel x in V_i. The features are computed for each V_i, normalized, and linearly combined. The combination weights (α, β, γ) have been selected to provide higher relevance to selected features. The low-level viewport features can thus be used to compute the low-level saliency map, $Sal_{low}^{V_i}$, as:

$$Sal_{low}^{V_i} = \alpha \cdot (H + S + L)_{V_i} + \beta \cdot G_{V_i} + \gamma \cdot E_{V_i} \qquad (2)$$

The saliency map $Sal_{low}^{V_i}$ is further improved by exploiting the content of viewports. The following subsection depicts the use of human faces as high-level features for saliency estimation.

C. Face Detection

The saliency map may be further improved by using high-level factors, such as recognized objects or the presence of human faces. The latter element plays an important role in guiding visual attention, and thus, the inclusion of face detection into a visual attention model can improve its quality [10].

For the face detection task, we use the TinyFace approach presented in [15]. It is based on a multi-layer hybrid-resolution model for detecting faces with small and large resolution in a scene. The context-aware neural network architecture used in this work employs a very large receptive field so as to account for both types of resolution. First, for an input image a coarse pyramid is built that includes a 2X interpolation of the image. The scaled image is subsequently passed through a ResNet101 CNN architecture [13]. The ResNet101 is trained on 25 templates of the WIDER FACE dataset [25]. Additionally, features from different stages of the network are aggregated to enhance the classification performance. Finally, a non-maximum suppression is performed to get the detected face boundaries for the input image. The published results on FDDB dataset [17] showcase that the approach outperforms state-of-art algorithms for face detection. It performs well for images with large faces as well as for crowd images with numerous small faces. Therefore, we select this face detector for identifying the presence of a person in each viewport and process the detected faces for visual attention modelling.

Based on the evidence that the viewing behaviour of users wearing HMDs follows a Gaussian distribution [5], and that the area falling within -30° to +30° is exhaustively explored, the detected faces lying in this region are given a higher weight.

The face selection approach used in this work can be summarized as follows:

1) for each detected face f, the distance d_f from the center of the viewport is computed;
2) the mean distance d_m of all detected faces from the center is computed for all viewports;
3) the faces in a viewport that lie within the mean distance d_m are considered for visual attention modelling. We set the intensity of pixels with the detected faces to 1 and the remaining pixels to 0.

Figure 2 shows an example of the face detection procedure performed on a viewport. Figure 2 (a) shows the result of the algorithm in [15] while Figure 2 (b) reports the outcome of the proposed modified version of [15] in which only the faces in the salient area around the viewport center are considered.

The output of this module is an attention inspired saliency map, $Sal_{high}^{V_i}$. When this map is available, it is combined with the low-level saliency map ($Sal_{low}^{V_i}$) for each viewport (V_i). The maximum value pixels among the two saliency maps are selected for generating the final estimated saliency map [10]:

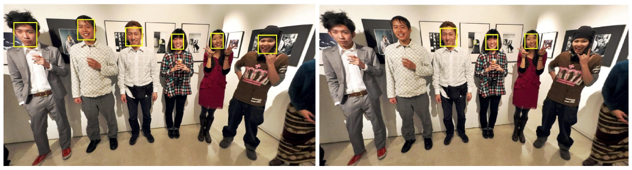


Fig. 2: Left image shows the detected faces using [15]. The image on right shows the faces in the salient area around the viewport center selected by the proposed algorithm.

$$Sal^{V_i} = max(Sal_{low}^{V_i}, Sal_{high}^{V_i}). \tag{3}$$

The obtained saliency maps for all viewports $Sal^{\mathbf{V}}$ (where $\mathbf{V} = V_1, V_2, ..., V_N$ and N is the total number of viewports extracted from the input image) are then re-projected to the equi-rectangular plane.

D. Equator Biased Weighting

In order to account for the fact that users wearing HMD tend to look at contents close to the equator and rarely along the periphery [5], an equator biased weighting on the face enhanced low-level equi-rectangular saliency map is performed in accordance with [2], to generate the final saliency map. To create a smooth equi-rectangular saliency map, we additionally perform normalization and apply a Gaussian low-pass filter, to obtain the final FLS$_{map}$.

III. Experimental results

To verify the effectiveness of the proposed method, experiments are performed on the 360° image dataset presented in the Grand Challenge “Salient360!” organized at the IEEE International Conference on Multimedia and Expo (ICME) 2017 [11], [22]. The dataset contains 85 equi-rectangular images and their corresponding saliency maps collected through subjective tests. The performances of the proposed FLS model are then compared with one baseline approach (GBVS) [12] and five state-of-art works: JU [9], RM3 [2], and TU1, TU2, and TU3 [24]. The Correlation Coefficient (CC) [4] and the Kullback-Leibler Divergence (KLD) [4] are used for comparing the performance of the considered approaches with respect to the available ground truths. The CC depicts the strength and direction of a linear relationship between the estimated and the ground truth saliency maps. Its value ranges between -1 to +1, where a higher value depicts a better saliency estimation. Whereas, the KLD is, in the present context, the difference between the distribution of pixel intensities in the estimated saliency map and the ground truth. A lower KLD indicates better estimation of image saliency.

The parameters and thresholds used during the performed experiments are detailed in the following. During the viewport extraction, the horizontal and vertical sampling rates are set at 40° and 35°, respectively. The dimension of viewports are fixed at 1920 × 1080 pixels. We empirically set the

TABLE I: (a) Results of CC and KLD averaged over all images in the dataset [11]. (b) and (c) are the list of best and worst performing images with FLS approach, respectively. The results are presented on basis of best and worst CC and KLD for the proposed FLS model.

(a) Results on overall dataset.

Model	**CC↑**	**KLD↓**
FLS	**0.63**	**0.57**
TU1 [24]	0.62	0.75
TU2 [24]	0.56	0.64
RM3 [2]	0.52	0.81
JU [9]	0.57	1.14
TU3 [24]	0.44	1.09
GBVS [12]	0.41	0.97

(b) Best performing images.

Metric	**P26**	**P81**	**P8**	**P17**
CC	0.79	0.76	0.76	0.75

Metric	**P32**	**P26**	**P25**	**P5**
KLD	0.25	0.27	0.28	0.3

(c) Worst performing images.

Metric	**P21**	**P63**	**P60**	**P11**
CC	0.44	0.44	0.44	0.45

Metric	**P96**	**P57**	**P4**	**P68**
KLD	2.04	1.13	1.02	0.83

tuning factors α, β and γ at 0.2, 0.2 and 0.6, respectively (Equation (2)). The neural network architecture and related parameters/weights for face detection are set according to [15].

For sake of clarity, in the following we briefly describe the saliency estimation approach in the state-of-art works with which we compared the proposed FLS. The JU approach [9] combines only prominent low-level features such as texture, contrast, edge boundary connectivity, hue, saturation and luminance for detecting salient regions in an image. A combination of low and high level features are used in RM3 [2]. The basic idea of our proposed FLS model is based on the work presented in RM3. They consider the presence of a person in an image as a driving criteria for saliency estimation. Their work is divided into three modules, combination of hue and saturation channels, GBVS on the hue component and lastly, include presence of person as the high-level feature by detecting skin tone, face and number of persons. Startsev et al. [24] propose three saliency estimation models ensemble of deep networks eDN (TU3), saliency attentive model SAM (TU2) and a combination of eDN, SAM, and GBVS (TU1).

In Table I (a), the average values of CC and KLD metrics are reported. The obtained results clearly depict that the proposed FLS model outperforms the compared saliency estimation approaches for both values of CC and KLD. The high-level features that act on local regions exploit the content information of the image. GBVS does not exploit the high-level features for saliency estimation. The JU approach does not involve any high-level feature for saliency estimation. However, high-level features are very important for both 2D and 360° images. The approach in RM3 exploits the Viola-Jones face detector which fails to detect tilted or turned faces. Moreover, they are also prone to illumination variance and detection of faces in a crowd. In this direction, we exploit a neural network architecture for detecting faces. The ensemble of networks eDN employed in TU3 is trained using the salient and non-salient regions. However, no specific learning module is involved for identifying the objects in the image. The LSTM-based CNN used in saliency attentive model SAM (TU2) also has a similar limitation. However, the combination of eDN, SAM and GBVS, over the continuity-aware image projection (TU1) is found to be outperforming other existing approaches (TU2, TU3, RM3, JU and GBVS), and is close to our approach.

The proposed model (FLS) outperforms other models mainly based on two aspects: the inclusion of entropy as low-level feature and the modelling of user attention by considering detected faces. Entropy accounts for the unexpectedness in an image and thus helps us to understand the most diverse regions in the image. This diversity is in regard to the pixel intensities and hence is considered as a low-level feature.

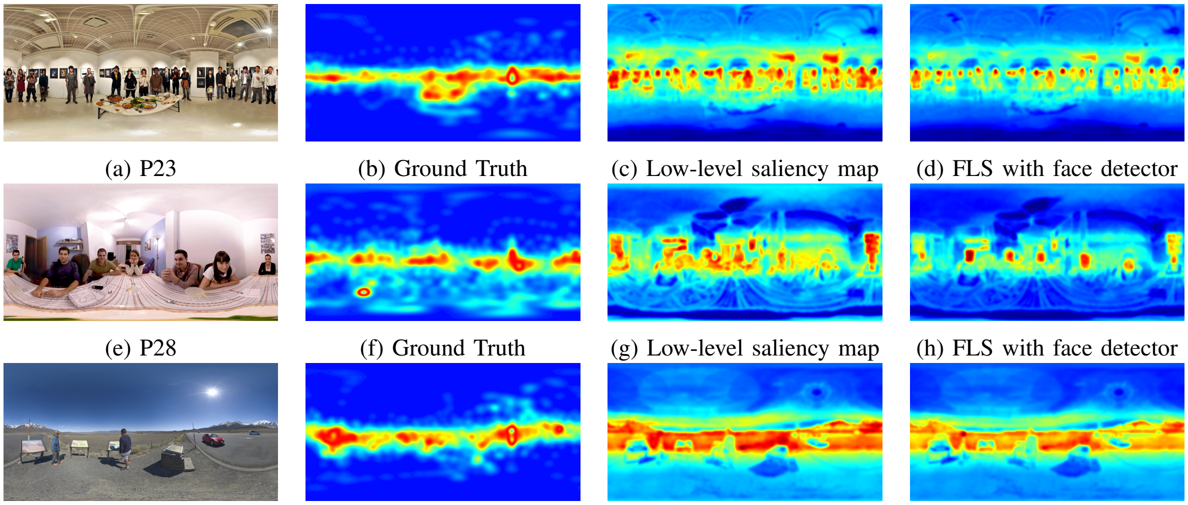
The presence of a person in the scene attracts attention of viewers [2], [10]. Therefore, detecting faces and weighting them based on their impact on user attention helps to improve the overall saliency estimation. The improvements in saliency estimation obtained when considering the presence of human subjects can be noted in Figure 3. The detection of faces allows to improve the performances of the proposed algorithm with respect to the case in which only the low-level features are exploited. Table I (b) and (c) show the best and worst performing images on the proposed FLS model.

From the performed analysis it results that, for some images, the estimated saliency map differs from the ground truth. In these cases, the scenes contain high-level features, different from faces, representing important clues for the human attention. To cope with this issue, in future works, other relevant features might be considered to improve the estimation performances.

Another important observation from the worst performing images is that, detecting saliency of images under low light is very challenging.

IV. Conclusions

In this contribution, a face attention and low-level feature based omni-directional image saliency estimation model is presented. The approach unifies low and high-level features for estimating 360° image saliency. The hue, saturation and luminance feature channels are combined with GBVS and entropy for estimating a low-level saliency map. Subsequently, attention of users while viewing an omni-directional image is modelled by detecting the presence of faces in the scene. Moreover, only the detected faces that impact the overall visual saliency are taken into account. The final saliency map is generated by taking the maximum of the low-level saliency map and the face saliency map. Experiments performed on the ”Salient360!” 360° image dataset show that the proposed approach outperforms the existing saliency techniques.



(a) P23 (b) Ground Truth (c) Low-level saliency map (d) FLS with face detector

(e) P28 (f) Ground Truth (g) Low-level saliency map (h) FLS with face detector

(i) P62 (j) Ground Truth (k) Low-level saliency map (l) FLS with face detector

Fig. 3: First column shows images with faces that can attract visual attention. Second column shows the corresponding ground truth saliency maps. Third column shows the low-level saliency map without using visual attention modelling. Fourth column shows the estimated saliency by the proposed FLS model.

REFERENCES

[1] A. Azaza, J. Weijer, A. Douik, and M. Masana. Context proposals for saliency detection. *Computer Vision and Image Understanding*, 174:1 – 11, 2018.

[2] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli. A feature-based approach for saliency estimation of omni-directional images. *Signal Processing: Image Communication*, 69:53 – 59, 2018.

[3] S. Biswas, S. A. Fezza, and M.-C. Larabi. Towards light-compensated saliency prediction for omnidirectional images. In *International Conference on Image Processing Theory, Tools and Applications*, pages 1–6. IEEE, 2017.

[4] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

[5] Y. Ding, Y. Liu, J. Liu, K. Liu, L. Wang, and Z. Xu. Panoramic image saliency detection by fusing visual frequency feature and viewing behavior pattern. In *Pacific Rim Conference on Multimedia*, pages 418–429. Springer, 2018.

[6] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang. Perceptual quality assessment of omnidirectional images. In *International Symposium on Circuits and Systems*, pages 1–5. IEEE, 2018.

[7] T. Ebrahimi. Quality of multimedia experience: past, present and future. In *International Conference on Multimedia*, pages 3–4. ACM, 2009.

[8] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin. Saliency detection for stereoscopic images. *IEEE Transactions on Image Processing*, 23(6):2625–2636, 2014.

[9] Y. Fang, X. Zhang, and N. Imamoglu. A novel superpixel-based saliency detection model for 360-degree images. *Signal Processing: Image Communication*, 69:1–7, 2018.

[10] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[11] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images. *Signal Processing: Image Communication*, 69:35 – 42, 2018.

[12] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[14] S. He, A. Borji, Y. Mi, and N. Pugeault. What catches the eye? Visualizing and understanding deep saliency models. *Computing Research Repository*, abs/1803.05753, 2018.

[15] P. Hu and D. Ramanan. Finding tiny faces. *Computing Research Repository*, abs/1612.04402, 2016.

[16] C. Huang, Q. Liu, and S. Yu. Regions of interest extraction from color image based on visual saliency. *The Journal of Supercomputing*, 58(1):20–33, 2011.

[17] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[18] B. John, P. Raiturkar, O. Le Meur, and E. Jain. A benchmark of four methods for generating 360 saliency maps from eye tracking data. In *International Conference on Artificial Intelligence and Virtual Reality*, pages 136–139. IEEE, 2018.

[19] P. Lebreton and A. Raake. GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. *Signal Processing: Image Communication*, 69:69–78, 2018.

[20] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic. SalNet360: Saliency maps for omni-directional images with CNN. *Signal Processing: Image Communication*, 69:26 – 34, 2018.

[21] N. Murray, M. Vanrell, X. Otazu, and C. Parraga. Saliency estimation using a non-parametric low-level vision model. In *Conference on Computer Vision and Pattern Recognition*, pages 433–440. IEEE, 2011.

[22] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Conference on Multimedia Systems*, pages 205–210. ACM, 2017.

[23] A. Singla, S. Fremerey, W. Robitza, and A. Raake. Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays. In *International Conference on Quality of Multimedia Experience*, pages 1–6. IEEE, 2017.

[24] M. Startsev and M. Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Processing: Image Communication*, 69:43 – 52, 2018.

[25] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Conference on Computer Vision and Pattern Recognition*, pages 5525–5533. IEEE, 2016.

[26] Y. Zhu, G. Zhai, and X. Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018.