A feature-based approach for saliency estimation of omni-directional images

Federica Battisti, Sara Baldoni, Michele Brizzi, Marco Carli

Università degli Studi Roma TRE
Department of Engineering
via Vito Volterra, 62
00146 Rome, Italy

Abstract

Omni-directional imaging records the visual information from any direction with respect to a given view-point. It is gaining consumers' popularity due to fast spreading of low-cost devices both for acquisition and rendering. The possibility to render the whole surrounding space represents a further step towards immersivity, thus providing the user with the illusion of physically being in a virtual environment. The understanding of visual attention mechanisms for these images is a relevant topic for processing, coding, and exploiting such data. In this contribution, a saliency model for omni-directional images is presented. It is based on the combination of low-level and semantic features. The first ones account for texture, viewport saliency, hue and saturation, while the second are used to take into account the impact of the presence of human subjects on the saliency. The proposed model has been tested in the "Salient360! Visual attention modeling for 360° Images" Grand Challenge. The model, the achieved results, and finding/discussions are here presented.

Keywords: Saliency estimation, Human fixation, Omni-directional images, 360° images

1. Introduction

Immersivity strongly depends on successfully fooling several senses, principally sight and hearing. Thanks to this process, it provides the viewers with the feeling of physically being in the place shown by the rendering system. It is generally achieved through the implementation of virtual environments, in which the viewer perceives him/herself as being surrounded by a 3D world. The surrounding world can be computer-generated or can be the rendering of real scenes. In the latter case, specific acquisition devices have to be used. The goal is to obtain an omni-directional image, or video, that allows the visual information to be seen from any direction with respect to a given view-point.

The imaging system may exploit mechanical or optical devices. In the first case, motorized linear or array-based cameras scan the scene resulting in very high-resolution images. In this case, the drawback is time consumption that can be very long in case of high quality scan. When the application scenario does not require very high definition, or when a real time constraint is present, optical solutions are employed. Basically, those are based on the use of mirrors or of special lenses (i.e., fish eye). Nowadays low-cost devices are available for acquiring omni-directional images, i.e. 360° cameras, thus pushing this technology and its application to the consumer market.

The acquired information can be rendered through 2D display, cave, or Head-Mounted Display (HMD). In particular HMD enables egocentric scene viewing, allowing the user to modify the point of view by simply moving his/her head or body, thus increasing the perceived quality of experience. In fact, the user is assumed to be placed in the center of a sphere and by moving the head, he can observe the omni-directional stimula.

The increasing use of this technology opens several issues such as the design of new rendering systems, new applications for exploiting the information, or new compression systems. One of the first aspects that needs to be investigated is the understanding of the modalities in which a human subject explores the omni-directional image, thus defining the salient points in the image.

More generally, saliency estimation refers to the localization of the areas in an image having particular clue for a human observer. This information is generally obtained by exploiting fixation points, that are the points in the visual field that are fixated by the two eyes in normal vision, and for each eye those are the points directly stimulating the fovea [1, 2]. They can be captured by means of eye-trackers, or cameras. The clustering of the fixation points, usually obtained by convolving the fixation points map with a Gaussian kernel, is used to produce the saliency map. The obtained map represents the degree of interest of an observer and can be used in many applications such as quality assessment [3], video surveillance [4], tone mapping [5] or defect detection. A classification of possible applications is presented in [6].

In literature, many efforts have been devoted to model the saliency of images and videos [7, 8]. Proposed methods can be categorized according to the features they rely on: bottom-up approaches are based on low-level local features, like color, intensity, contrast, or orientation [9], while top-down approaches exploit high-level cues like context, semantic, knowledge, expectations or application [10].

However, to the best of our knowledge, very few methods are specific to omni-directional images. An application-based approach is proposed in [11] where a method for extracting visual attention-based features from panoramic (cylindrical) images is used for robot localization. A similar application-based approach is proposed in [12]. In [13] an algorithm exploiting spherical geometry for reducing the geometrical distortions, that may be introduced by the plane mapping, is proposed. This method is based on the use of low-level features as proposed in [14]. In [15], the authors predict salient image regions by taking into account the image exposition time. This allows to understand the influence of this parameter in the resulting saliency map. Performed tests show that duplicating the exposition time does not modify significantly the saliency map. In [16], the authors exploit eye-tracking in a HMD system for gaze analysis. Results suggest that most eye-gaze fixations are rather far away from the center of the viewport. In [17] a method for estimating salient objects in panoramic images is presented. A draft of the saliency map is computed by background estimation and then is refined by computing the contrast only in the surrounding regions. In [15], 2D image features are estimated on a lattice of viewports and the overall saliency map is computed by considering the contribute of each viewport.

In this paper a novel model for saliency estimation in omni-directional images is proposed. In more details, viewports are first collected from the omni-directional content and, then, the visual attention is estimated by analyzing low-level and semantic features extracted from each viewport. The strategy of analyzing the viewports instead of the whole panoramic image relies on the fact that, in the exploration of the omni-directional content, the user watches only one portion of it at a time [18].

The rest of the paper is organized as follows: Section 2 details the characteristics of the proposed saliency model, Section 3 includes the system parameters, the details of the adopted image database and a discussion on the characteristics of the proposed model. Finally, in Section 4 the conclusions are drawn.

2. Proposed Method

The proposed saliency model is shown in Figure 1. Each input image (i.e., the equi-rectangular image) undergoes a pre-processing step in which the viewport extraction is performed. Then, high-level (i.e., skin color, faces, and number of people) and low-level features (i.e., hue, saturation, intensity, and contrast) are extracted for each viewport and averaged to obtain a first estimation of the saliency map. This map is then refined by using an equator-prior weighting and a smoothing operation.

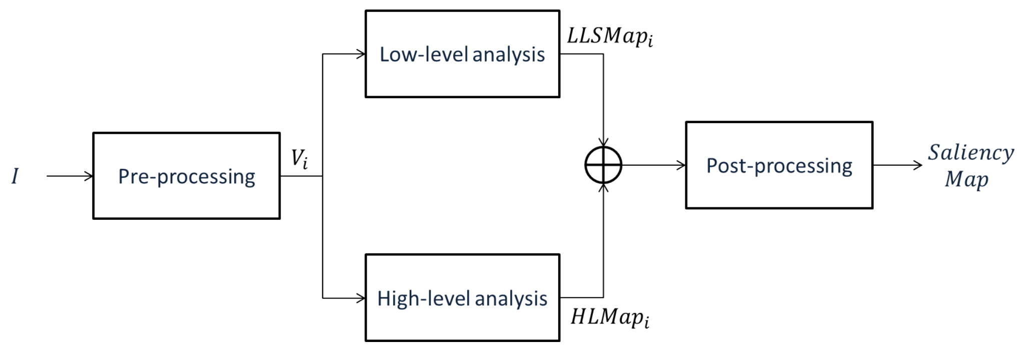


Figure 1: Proposed saliency model

In the following, each step of the algorithm will be detailed.

2.1. Pre-processing

The viewports V_i, with $i = 1, ..., n$, are extracted from the input equi-rectangular image I according to the procedure described in [19], here reported for sake of clarity.

A non-uniform angular sampling of the sphere is performed, with $\Delta\phi$ and $\Delta\theta$ being the horizontal and vertical angular sampling rates and $X_i(\phi, \theta)$ the i^{th} sampling point. Since we assume that the user will change location in the omni-directional content by moving his head, the center of the viewport V_i will correspond to the coordinates of X_i.

To extract each viewport, the position of each pixel of V_i is back-projected into the spherical reference, and then into the equi-rectangular frame. These coordinates are used to interpolate over I.

Let (x, y) be the coordinates of any point M_v in the viewport V_i, whose size is $[V_{width}, V_{height}]$. To represent the inverse gnomonic projection of V_i on the sphere, we define a three dimensional Cartesian coordinate system, whose origin is surrounded by the spherical frame of unitary radius, and place the viewport V_i on the plane tangent to the sampling point X_i (as shown in Figure 2). Let us consider the case in which the sampling point X_i corresponds to the center of the equi-rectangular image. Then, the position of M_v on the aforementioned plane is given by:

$$M_p(x, y, z) = \begin{bmatrix} 1 \\ pxl \cdot (x - \frac{V_{width}}{2}) \\ pxl \cdot (y - \frac{V_{height}}{2}) \end{bmatrix}$$

where pxl is the size of a pixel in V_i, obtained as:

$$pxl = 2\frac{\tan\left(\frac{a}{2} \cdot \frac{\pi}{180}\right)}{V_{width}}$$

where a is the size of the viewport in degrees.

The projection of M_p on the sphere, M_s, is:

$$M_s(x, y, x) = \frac{M_p(x, y, z)}{\|M_p(x, y, z)\|}$$

where $\|M_p(x, y, z)\|$ denotes the L^2 norm of vector $M_p(x, y, x)$.

In the case of any other sampling point, M_s needs to be multiplied by the rotation matrix $R_{\theta,\phi}$:

$$R_{\theta,\phi} = \begin{pmatrix} \cos(\theta)\cos(\phi) & -\sin(\phi) & \cos(\theta)\sin(\phi) \\ \sin(\theta)\cos(\phi) & \cos(\theta) & \sin(\theta)\sin(\phi) \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix}$$

where ϕ and θ are the azimuth and elevation angles of the sampling point X_i on the sphere (Figure 2).

To obtain the corresponding coordinates in the equi-rectangular image, M_e, the following relation holds:

$$M_e(x, y) = \begin{bmatrix} E_{width} \cdot \left(\frac{ang}{2\pi}\right) \\ E_{height} \cdot \left(\frac{\arcsin(M_s(z))}{\pi + 0.5}\right) \end{bmatrix}$$

where $[E_{width}, E_{height}]$ are the sizes of the equi-rectangular image and ang is given by:

$$ang = \tan^{-1}\left(M_s(y), M_s(x)\right)$$

where the four-quadrant inverse tangent function is used.

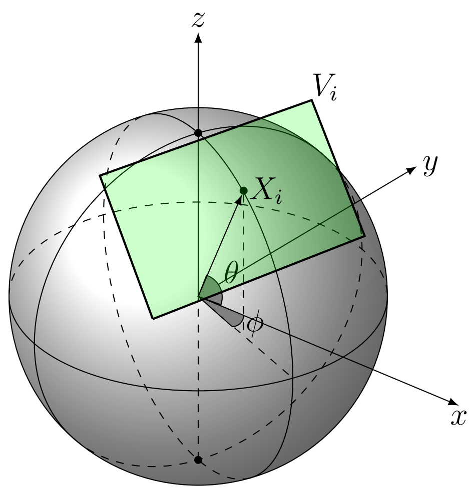


Figure 2: Viewports extraction for an arbitrary sampling point X_i

2.2. Low-level analysis

In the low-level analysis (see Figure 3), each viewport V_i is converted from the Red, Green, and Blue (RGB) color space to the Hue, Saturation, and Value (HSV) one. The latter is characterized by better visual consistency than the RGB one as suggested in [20]. For each viewport V_i, only the Hue (H_i) and Saturation (S_i) components are taken into account. In more details, for each V_i, a first map ($LLSMap_i$) is obtained as combination of the result of the texture analysis performed on V_i with the weighted sum of H_i, S_i and the outcome of the Graph-Based Visual Saliency (GBVS) analysis [21] of the H_i component. This model allows to estimate the human fixations based on the creation of activation maps on specific feature channels that are normalized to enhance the importance of the points attracting the human attention. Moreover, it has been proven that GBVS supports a center bias, by assigning higher saliency values in the center of the image plane. Based on this, $LLSMap_i$ is computed as:

$$LLSMap_i = T_i \cdot W_i \tag{1}$$

where:

- T_i is a binary texture map extracted from V_i by using the multi-channel filtering approach described in [22];
- W_i is obtained as:

$$W_i = \alpha S_i + \beta H_i + \gamma G_i \tag{2}$$

  where α, β and γ are the coefficients of the weighted sum while G_i is the output of the GBVS procedure.

Finally, the overall low-level saliency map, $LLSMap_{tot}$, is obtained by equi-rectangular projection of each $LLSMap_i$.

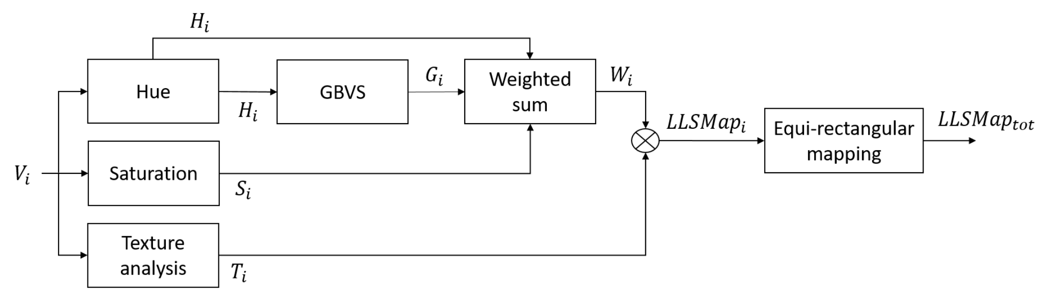


Figure 3: Low-level features analysis

2.3. High-level analysis

The steps performed in the high-level analysis are detailed in Figure 4. The input viewport V_i undergoes two parallel processing: skin and face detection. The former is performed based on the methods presented in [23]. Based on this approach, only the portions of V_i whose color components are inside a pre-defined range are extracted to obtain the map P_i. Face detection is achieved by using the Viola-Jones algorithm [24] to retrieve the map F_i. It is useful to underline that this step is performed using not only the frontal face classification model but also the upper body and profile face classification models. The former detects head and shoulders areas while the latter detects upright face profiles. The maps P_i and F_i are computed for each viewport and they are combined through an equi-rectangular mapping to respectively obtain the maps P_{tot} and F_{tot}. In order to more accurately identify the presence of human subjects, a weighted combination is performed to obtain the fusion map PF_{tot} as:

$$PF_{tot} = \frac{2 \cdot F_{tot} + P_{tot}}{3}. \tag{3}$$

After the regions containing persons have been identified, a people count is performed in order to estimate the number of persons (nP) identified in PF_{tot}. This value, obtained through a blob analyzer, is then used to define the weight w_{people}. Finally, the output $HLMap_{tot}$ is obtained as:

$$HLMap_{tot} = w_{people} \cdot PF_{tot}; \tag{4}$$

In this work, a more relevant weight will be given to regions containing a limited number of subjects rather than a large one. In fact, in the latter case the human subjects in the scene are hardly distinguishable, appearing as a texture and therefore reducing the impact of that region on the saliency.

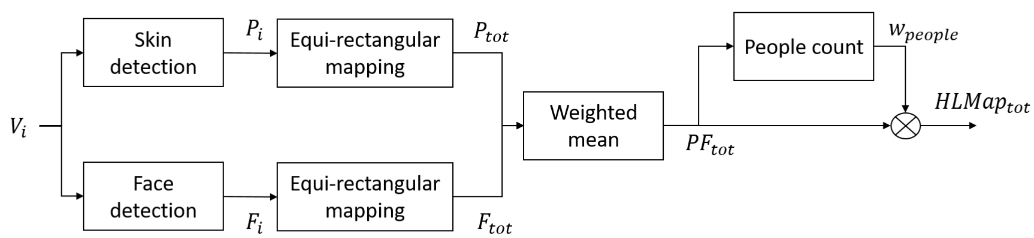


Figure 4: High-level features analysis

2.4. Post-processing

After the computation of $LLSMap_i$ and $HLMap_i$, they are averaged in order to account equally for low-level and high-level features. The obtained fusion map undergoes a post-processing step before returning the overall Saliency Map.

Several studies [25, 26] show that fixations distribution tends to be strongly biased towards the center of the screen when viewing 2D scenes on computer monitors, to the point that a saliency map composed of a centered Gaussian blob has good performances in predicting fixations [27]. This appears to be independent from the distribution of the features in the images [28]. There are different explanations for this behavior: first, objects of interest are often placed by photographers in the center of photographs by exploiting the rule of thirds; second, fixations might be influenced by the setup used to experimentally record eye-tracking data, where users are usually placed in front of

the screen [29].
In the case of omni-directional images, this assumption does not hold completely, since users can explore the whole content by freely moving eyes and head. However, even in this case, a bias towards the central area (i.e., the equatorial area) of the omni-directional image, holds. This bias can be due to the human posture and to the fact that moving the head for looking at a different direction requires more intense movements with respect to the ones required by the eyes [30, 31]. Therefore it is more likely for a subject to first span the visible area with the eyes and then with the head, thus confirming the results in [16]. For this reason, in the proposed method, a weighting window is applied to the estimated saliency map in the equi-rectangular format. As can be noticed in Figure 5, the applied cost function is increasing with the distance with respect to the equatorial line. The cost values are in the range 1 (in the central region) to 1/4 (in the border regions).

Finally a low-pass filtering and a normalization step are performed for obtaining the smoothed final saliency map.

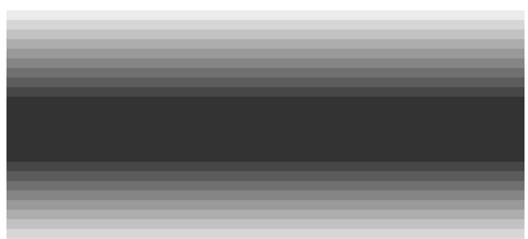


Figure 5: Weighting window

3. Evaluation tests

In the performed tests, the training set has been used for setting up the system parameters while the validation set has been exploited for testing the performances of the proposed approach. In the following, Subsection 3.1 details the procedures used for selecting the system parameters and the adopted values, Section 3.2 describes the adopted database, and Section 3.3 presents the performed tests and the obtained results.

3.1. System parameters

- **Viewport extraction:** in order to perform the extraction of viewports we used a horizontal sampling rate $\Delta\phi = 40°$ and a vertical sampling rate $\Delta\theta = 35°$. Knowing that the HMD used for the test dataset has a resolution of 960x1080 pixels per eye and a total Field-Of-View (FOV) of 100°, we extracted 1920x1080 pixels viewports in order to provide the same FOV and set the value of the size of the viewport in degrees, a, accordingly.

- **Weighting constants:** the values of α, β and γ have been empirically chosen for maximizing the correlation between estimated and ground truth saliency maps in the training dataset. In more details, the adopted normalization values are: $\alpha = 1$, $\beta = 0.1$ and $\gamma = 0.1$. $LLSMap_i$ and $HLMap_i$ are averaged in order to equally account for low-level and high-level features.

  The parameter w_{people} is set according to:

$$w_{people} = \begin{cases} 0.3 & if\ nP \geq 10 \\ 0.6 & if\ 5 \leq nP < 10 \\ 0.9 & if\ nP < 5. \end{cases} \tag{5}$$

  During the face detection step, the area of the smallest detectable object is set to 100x100 pixels and during the people count step, the minimum blob area detectable by the blob analyzer is set to 6000 pixels.

- **Gabor filters:** the adopted approach [22] is based on the use of a bank of Gabor filters in order to almost uniformly cover the spatial-frequency domain. In this work four orientation values were adopted [0°, 45°, 90°, 135°] and increasing values of radial frequency, raising with step 1 octave from $\sqrt{2}$ to the hypotenuse length of the input image, have been taken into account.

3.2. Image database

The test dataset [32] is composed by three sub datasets (one per model type: head only, head+eye, scanpath). In this work, only the head motion-based saliency model is considered, and the relevant dataset is composed by 20 training images (with size from 4000x2000 to 16000x8000 pixels) and 25 evaluation images (with size from 3000x1500 to 12000x6000 pixels), with the corresponding ground truth. The database includes several subjects such as vehicles (i.e. cars, public transport), urban environment (i.e. squares, buildings, supermarkets, museums theaters, hotels), landscapes, animals, and people. Moreover, for the provided images, different lighting conditions can be found.

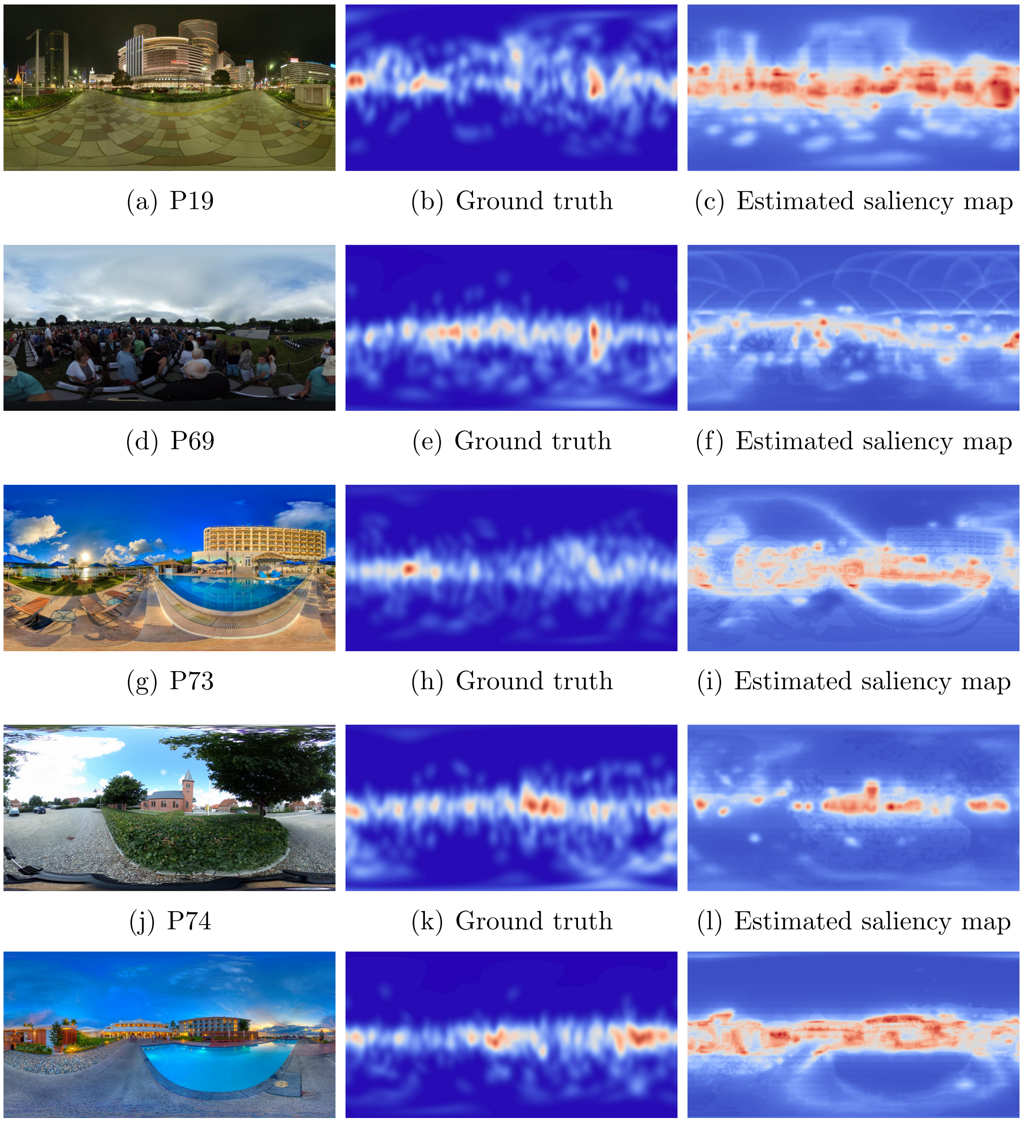
3.3. Experimental results

In order to assess the performances of the proposed method (indicated in the following as RM3), the tools provided by the "Salient360! Visual attention modeling for 360° Images" Grand Challenge, detailed in [32, 33], have been used to compare the estimated saliency map with the available ground truth. For evaluating the effectiveness of the proposed method with respect to other methods, we computed the Correlation Coefficient (CC) and KL Divergence (KLD) between the estimated saliency map and the ground truth saliency map of the images in the validation set. In Figures 6-7 the estimated saliency maps giving the best and the worst results are shown together with the original image and the corresponding ground truth. The corresponding CC and KLD values, compared with the ones obtained by the best performing algorithms in the challenge, Wuhan University (WU) [34] and Zhejiang University (ZU) [35], are reported in Tables 1-2, respectively. As can be noticed from Table 1, the CC or KLD values are comparable with the best performing ones. In some cases the CC value is closer to the best one while in others the KLD value is closer. This difference is due to the statistical measures considered by the two similarity functions.

	RM3		**WU [34]**		**ZU [35]**	
Image	**CC**	**KLD**	**CC**	**KLD**	**CC**	**KLD**
P19	0.63	0.57	0.68	0.46	0.75	0.51
P69	0.68	0.71	0.65	0.90	0.73	0.38
P73	0.63	0.51	0.84	0.20	0.86	0.19
P74	0.67	0.78	0.78	0.59	0.73	0.52
P79	0.65	0.67	0.78	0.43	0.72	0.43

Table 1: CC and KLD scores for the proposed method in the best performing cases compared with the benchmarks.

In order to carry out a more complete analysis of the proposed method, we evaluated its performances in terms of strengths and weaknesses on the entire dataset. In fact, by considering not only the validation but also the training set, it is possible to have a better understanding of the behavior of the proposed algorithm for different types of content.



(a) P19 (b) Ground truth (c) Estimated saliency map

(d) P69 (e) Ground truth (f) Estimated saliency map

(g) P73 (h) Ground truth (i) Estimated saliency map

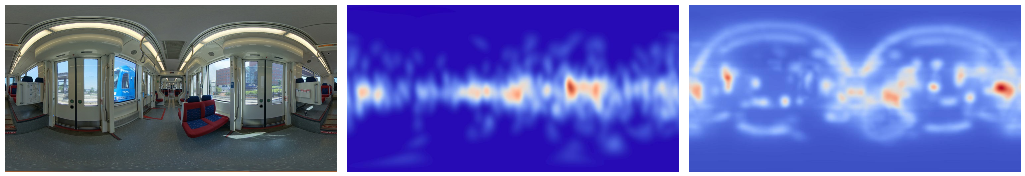
(j) P74 (k) Ground truth (l) Estimated saliency map

(m) P79 (n) Ground truth (o) Estimated saliency map

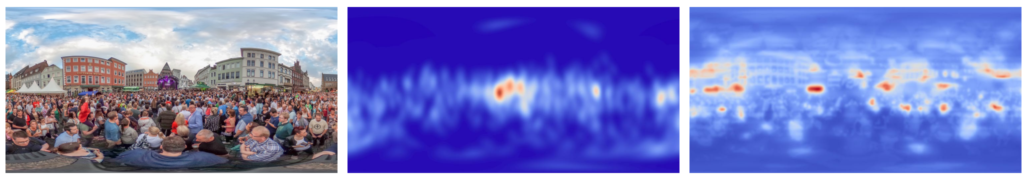
Figure 6: Left column: best performing images in the entire dataset. Center column: saliency map ground truth. Right column: saliency map estimated by the proposed algorithm

3.3.1. Model strengths

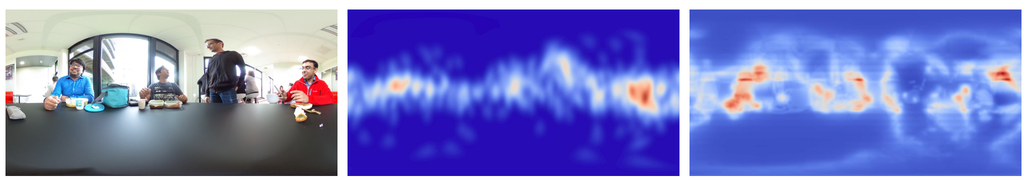
From the analysis performed on the images in the dataset, it is possible to highlight some characteristics of the algorithm that allowed to obtain results



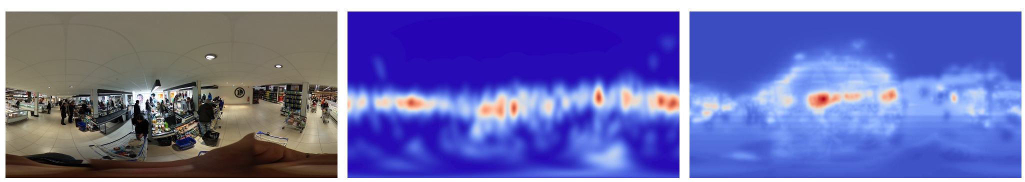
(a) P1 (b) Ground truth (c) Estimated saliency map



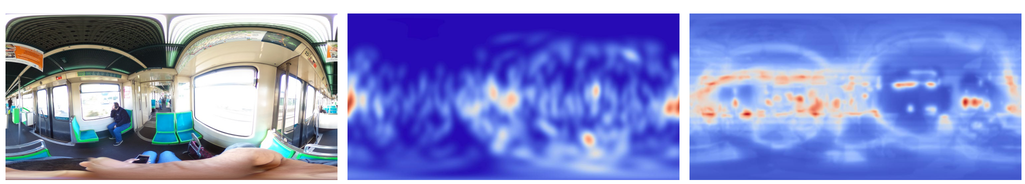
(d) P71 (e) Ground truth (f) Estimated saliency map



(g) P94 (h) Ground truth (i) Estimated saliency map



(j) P96 (k) Ground truth (l) Estimated saliency map



(m) P98 (n) Ground truth (o) Estimated saliency map

Figure 7: Left column: worst performing images in the entire dataset. Center column: saliency map ground truth. Right column: saliency map estimated by the proposed algorithm

similar to the best performing algorithms on the dataset:

- Use of equator-prior based weighting: the advantage of including this step in the proposed algorithm can be noticed for stimula P24 and P73,

	RM3		WU [34]		ZU [35]	
Image	**CC**	**KLD**	**CC**	**KLD**	**CC**	**KLD**
P1	0.49	0.95	0.71	0.58	0.65	0.52
P71	0.16	1.9	0.63	0.81	0.56	0.47
P94	0.41	1.03	0.76	0.45	0.79	0.34
P96	0.52	1.24	0.68	0.87	0.80	0.40
P98	0.30	0.75	0.41	0.56	0.36	0.77

Table 2: CC and KLD scores for the proposed method in the worst performing cases compared with the benchmarks.

in Figure 8(b) and (f) respectively. In these images, the horizon line corresponds to the equatorial line and thus the proposed model can effectively estimate the saliency of the scenes;

- Good performances in poor lighting conditions: this is achieved by including in the salient regions only areas containing mid-level values of saturation (S in the range [0.36-0.7]). An example is for stimula P10 and P69, respectively shown in Figure 8(a) and (d). As can be noticed, the method is able to detect saliency even in darker areas without overlaying regions belonging to the border between differently illuminated areas;

- High-level feature extraction: this step increases the weight in the saliency map estimation of the areas containing human subjects. This can be noted for stimulus P28 (Figure 8(c)) where the people around the table are successfully recognized.

3.3.2. Model weaknesses

The analysis of the performances of the proposed algorithm on the dataset, allowed to highlight some weak points that need a further improvement of the system:

- Indoor scenes with presence of distributed light sources: this is the case of Figures 8(a) and (f). A reason for such poor behavior might be the fact that, after conversion in the HSV color space, the V component is discarded. This operation might hinder the right handling of this type of light source;

- Difficulty in people detection in crowded environments: this problem could be solved by changing the adopted technique for discriminating the presence of skin. An example of this problem is in Figure 8(e), in which the presence of the crowd hinders the face detection algorithm. In this case, the proposed algorithm detects the presence of faces, however giving too importance to these areas. A more accurate skin selection procedure could reduce the number of false alarm.


(a) P10


(b) P24


(c) P28


(d) P69


(e) P71


(f) P73

Figure 8: Samples of images in the dataset.

4. Conclusions and comments

In this contribution, a method for visual saliency estimation for omnidirectional images is presented. The proposed method relies on the extraction of a set of viewports from the equi-rectangular image. From each of them, a first estimation of the saliency map is obtained by analyzing the high-level (i.e., skin color, faces, and number of people) and low-level features (i.e., hue, saturation, intensity, and contrast). These intermediate maps are then fused according to pre-defined weighting coefficients and finally refined by using an equator-prior weighting and a smoothing operation. The proposed model has been tested in the "Salient360! Visual attention modeling for 360° Images" Grand Challenge achieving good results especially in images containing human subjects and in case of poor light conditions. Future work will be devoted to a deeper investigation of the impact of image characteristics (e.g., contrast, illumination) on the performances of the system thanks to the use of larger databases.

References

[1] I. Murakami, Fixational eye movements and motion perception, Progress in Brain Research 154 (2006) 193 – 209. Visual Perception.

[2] S. Lee, A. Bovik, Fast algorithms for foveated video processing, IEEE Transactions on Circuits and Systems for Video Technology 13 (2003) 149–162.

[3] Q. Huynh-Thu, M. Barkowsky, P. L. Callet, The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives, IEEE Transactions on Broadcasting 57 (2011) 421–431.

[4] Y. Tong, F. A. Cheikh, F. F. E. Guraya, H. Konik, A. Trémeau, A Spatiotemporal Saliency Model for Video Surveillance, Cognitive Computation 3 (2011) 241–263.

[5] M. Narwaria, M. Perreira Da Silva, P. Le Callet, R. Pepion, Tone mapping based HDR compression: Does it affect visual experience?, Image Commun. 29 (2014) 257–273.

[6] M. Mancas, O. Le Meur, Applications of Saliency Models, Springer New York, New York, NY, pp. 331–377.

[7] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013) 185–207.

[8] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, C. W. Lin, A video saliency detection model in compressed domain, IEEE Transactions on Circuits and Systems for Video Technology 24 (2014) 27–38.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.

[10] A. Oliva, A. Torralba, M. Castelhano, J. M. Henderson, Top-down control of visual attention in object detection, in: IEEE International Conference on Image Processing, volume 1, pp. 253–256.

[11] A. Bur, A. Tapus, N. Ouerhani, R. Siegwar, H. Hiigli, Robot navigation by panoramic vision and attention guided features, in: 18th International Conference on Pattern Recognition (ICPR'06), volume 1, pp. 695–698.

[12] G. Schillaci, S. Bodiroža, V. V. Hafner, Evaluating the effect of saliency detection and attention manipulation in human-robot interaction, International Journal of Social Robotics 5 (2013) 139–152.

[13] I. Bogdanova, A. Bur, H. Hugli, Visual attention on the sphere, IEEE Transactions on Image Processing 17 (2008) 2000–2014.

[14] C. Koch, S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, Human Neurobiology 4 (1985) 219–227.

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

[16] Y. Rai, P. L. Callet, P. Guillotel, Which saliency weighting for omni directional image quality assessment?, in: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6.

[17] B.-J. Han, J.-Y. Sim, Saliency detection for panoramic landscape images of outdoor scenes, Journal of Visual Communication and Image Representation 49 (2017) 27 – 37.

[18] F. Jabar, J. Ascenso, M. P. Queluz, Perceptual analysis of perspective projection for viewport rendering in 360 images, in: 2017 IEEE International Symposium on Multimedia (ISM), pp. 53–60.

[19] FOXEL, Libgnomonic, `https://github.com/FoxelSA/libgnomonic`, 2017.

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

[21] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, MIT Press, Cambridge, MA, USA, 2006, pp. 545–552.

[22] A. K. Jain, F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, in: 1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings, pp. 14–19.

[23] S. Thakur, S. Paul, A. Mondal, S. Das, A. Abraham, Face detection using skin tone segmentation, in: 2011 World Congress on Information and Communication Technologies, pp. 53–60.

[24] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2004) 137–154.

[25] B. W. Tatler, R. J. Baddeley, I. D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vision Research 45 (2005) 643 – 659.

[26] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113.

[27] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, G. W. Cottrell, Sun: A bayesian framework for saliency using natural statistics, Journal of Vision 8 (2008) 32.

[28] B. W. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of Vision 7 (2007) 4.

[29] O. L. Meur, P. L. Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Research 47 (2007) 2483 – 2498.

[30] S. M. Safavi, S. M. Sundaram, A. H. Gorji, N. S. Udaiwal, P. H. Chou, Application of infrared scanning of the neck muscles to control a cursor in human-computer interface, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 787–790.

[31] A. Al-Rahayfeh, M. Faezipour, Eye tracking and head movement detection: A state-of-art survey, IEEE Journal of Translational Engineering in Health and Medicine 1 (2013) 2100212–2100212.

[32] Y. Rai, J. Gutiérrez, P. Le Callet, A dataset of head and eye movements for 360 degree images, in: Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys'17, ACM, New York, NY, USA, 2017, pp. 205–210.

[33] J. Gutiérrez, E. David, Y. Rai, P. Le Callet, Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360° still images, Signal Processing: Image Communication (2018).

[34] J. Ling, K. Zhang, Y. Zhang, D. Yang, Z. Chen, A saliency prediction model on 360 degree images using color dictionary based sparse representation, Signal Processing: Image Communication (2018).

[35] P. Lebreton, A. Raake, GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images, Signal Processing: Image Communication (2018).