https://doi.org/10.2352/ISSN.2470-1173.2019.11.IPAS-271

Depth-based saliency estimation for omnidirectional images

F. Battisti and M. Carli;
Department of Engineering;
Universitá degli Studi Roma Tre, Rome, Italy

Abstract

Visual saliency estimation aims at identifying and localizing the areas of images and videos that are attractive for a human subject. In this work, a novel approach for estimating the visual saliency in omnidirectional images is proposed. It is based on the identification of low-level image feature descriptors (i.e., the presence of texture, edges, etc.) coupled with the information about the local depth of the scene. To evaluate the performances of the proposed method, the estimated saliency map is compared with the available ground truth through two objective metrics: the correlation coefficient and the Kullback-Leibler divergence. The analysis of the achieved results confirms the validity of the proposed approach.

Visual saliency models are developed for estimating the amount of *attention* steered towards different regions of an image by the visual and cognitive system of a human subject [1]. These models may be useful in many research areas such as multimedia analysis and coding, media transmission, or content creation. Many methods for saliency prediction rely on visual attention models. In literature, several approaches have been proposed for traditional 2D imaging systems [2]. The available techniques are mostly based on the estimation of visual features that are combined with different strategies. The features can be classified in low-level (i.e., color, intensity, or orientation) [3, 4] and high level features (i.e., the presence of faces [5], detection of objects in the scene [6], or giving priority to the center of the scene [4]). The final pooling is generally obtained through linear combination [3] or by using learned weights [5, 7]. More recently, models using deep neural networks, trained for object recognition, [8–10], have shown improved capability on predicting visual attention. A detailed analysis of the algorithms presented in the literature for stereoscopic video is in [11].
The recent advent of novel imaging systems enabling the immersivity feeling in the user (e.g. multiview, plenoptic, and ominidirectional cameras), represents a challenge for these methods. In fact, 2D saliency models have been designed for estimating the saliency based on a planar, limited, representation of the scene and this approach is no more valid with the new types of multimedia. In particular, an omnidirectional (or 360°) image collects the visual information in a given point from any direction. It enables the rendering, at a given position, of any viewing angle selected by the user. Many applications can benefit from this technology from media production, to video surveillance and entertainment. The interest on this imaging system is witnessed by the amount of 360°video posted on social networks as Facebook or YouTube (more than 100.000 last year) also thanks to the availability of low cost 360°cameras. The first approaches to saliency estimation for 360°images were based on the application of the methods developed in the 2D domain. These approaches, even if promising, are not able to capture the peculiarity of the new system. In fact, it is useful to underline that while looking at an omnidirectional image, the user can only explore a portion of the spherical image (i.e., the viewport) being however able to modify the position of his viewport. Therefore, the saliency estimation is not limited to a single 2D image but should also consider this peculiarity of 360°images. Recently, a test dataset to support the development of saliency computational models was presented [12] and the challenge "Salient360! Grand Challenge" was proposed in [13]. In [14], a learning-based scene recognition algorithm for predicting the salient regions in a 360°image is proposed. In [15], a model for exploiting 2D saliency detectors in different map projections is proposed and applied for the design of a saliency-based smooth navigation path through the image. In [16], the MPEG-DASH SRD streaming is modified towards the 3D VR environment. In [17], an analysis of viewing behavior and saliency in Virtual Reality is performed and the outcomes applied to other applications e.g., automatic alignment of VR video cuts or saliency-based compression.
In [18], a first saliency model (RM3) was designed. It is based on the combination of low-level (e.g., hue, saturation, texture, intensity, and contrast) and high level (e.g., skin colors, face detection, and number of people) 2D saliency features. An analysis of the achieved results show that the computation of the high level features does not improve significantly the performances of the saliency estimator. For this reason in this paper, in order to build a more reliable saliency estimator, we modify the considered features and we take into account the impact of the distance of the objects in the scene with respect to the observer. The rest of the paper is organized as follows: in Section the proposed method is described, Section presents the tests performed for assessing the performances of the proposed saliency estimator and in Section the conclusions are drawn.

Proposed method

The aim of this work is to design a model for saliency estimation for omnidirectional images based on depth information. The proposed method is based on the fusion of two maps: the depth map and the visual attention map obtained by exploiting selected low-level features of the omnidirectional image. As preprocessing step, viewports of size 1080 x 1920 pixels are extracted from the equirectangular image. The block diagram of the operations performed on each viewport is in Figure 1 and a detailed description of each block is in the following subsections.

Low-level features analysis

As a first step, a low-level analysis is performed to obtain the low-level saliency map. This is obtained as weighted average of three components that are computed after conversion to the HSV color space:

- the H component of the viewport is used to feed a Graph Based Visual Saliency (GBVS) analysis [19]. The output

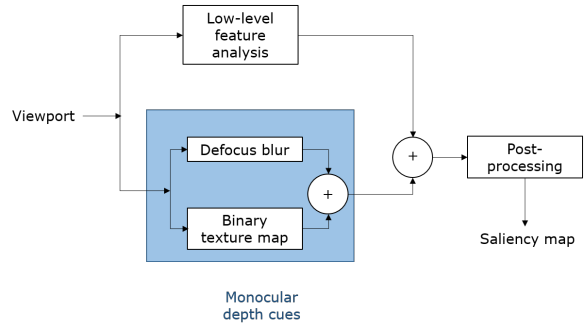


Figure 1. *Proposed saliency map estimation method.*

of this step is an estimate of the human fixations obtained through the creation of activation maps on specific feature channels that are normalized to enhance the importance of the points attracting the human attention;

- a binary texture map extracted from the V component of the viewport by using the multi-channel filtering approach described in [20];
- the S component of the viewport.

The overall low-level saliency map for the omnidirectional image, is obtained by a back projection of the map resulting from the weighted sum of each viewport.

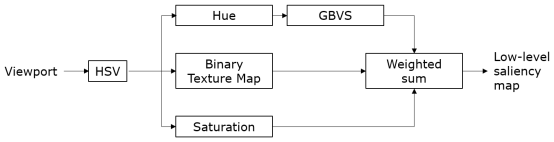


Figure 2. *Block diagram of the low-level feature analysis.*

Monocular depth cues

In order to estimate the depth of each viewport, the monocular cues have been exploited. They can be divided in pictorial and motion based cues. The first ones are based on visual features observed in a static view of a scene while the latter ones exploit the motion of the observer by taking advantage of motion parallax, which relies on the fact that nearby objects apparently move faster in the retinal image than the ones that are more distant. Since in this work we are dealing with still images, only pictorial cues have been taken into account. According to psychophysical studies [21], there are mainly seven pictorial cues supporting depth perception: texture gradient, defocus blur, occlusion, relative heights, linear perspective, shadows, and atmosphere scattering. In this work we have selected texture gradients and defocus blur as pictorial cues. This choice is motivated by: i) we are dealing with static natural images that are usually characterized by the presence of textures, ii) in general, the objects closer to the viewer appear more in focus to those far away, and iii) reducing the number of cues to be computed reduces the computational complexity. In more details:

1. **Texture gradients**, represent the size distortion that affects closer objects more than objects farther away. In fact, the pattern of a textured surface changes as a function of the distance from the observer. At a greater distance, textures get finer and appear smoother.
2. **Defocus blur**, represents the loss of sharpness occurring due to a translation along the optical axis away from the plane or surface of best focus. This is based on the fact that the defocusing of an object causes a depth-dependent blur in the image and that blur inherently keeps the depth information. Usually, in fact, the objects on focus are close up to the observer while the background is more blurred.

The cues obtained by the texture gradient and defocus blur analysis maps are then added up and back projected into an equirectangular image. In order to obtain the final saliency map, after normalization, the low-level saliency map is added to the one obtained by the analysis of the monocular depth cues.

Post-processing step

As well known, the distribution of the fixations tends to be strongly focused on the center of the scene, independently from the distribution of the features in the images. This is because the photographers tend to place the objects of interest in the center of the picture and because the fixations might be influenced by the setup used to record eye-tracking data, which usually places the user in front of the scene [22]. With omnidirectional images this is not completely true because the user is able to explore the entire content with free movements of the eyes and of the head. However, performed tests show that there is a strong bias towards the central area, which is the area around the equatorial line of the sphere [23]. For this reason, we apply a weighted window to the estimated saliency map in the equirectangular format. The window's weights increase with the distance from the equatorial line, raising from 1 in the central region to 1/4 in the border regions. Finally a low-pass filtering and a normalization step are performed for smoothing the overall saliency map.

Experimental test

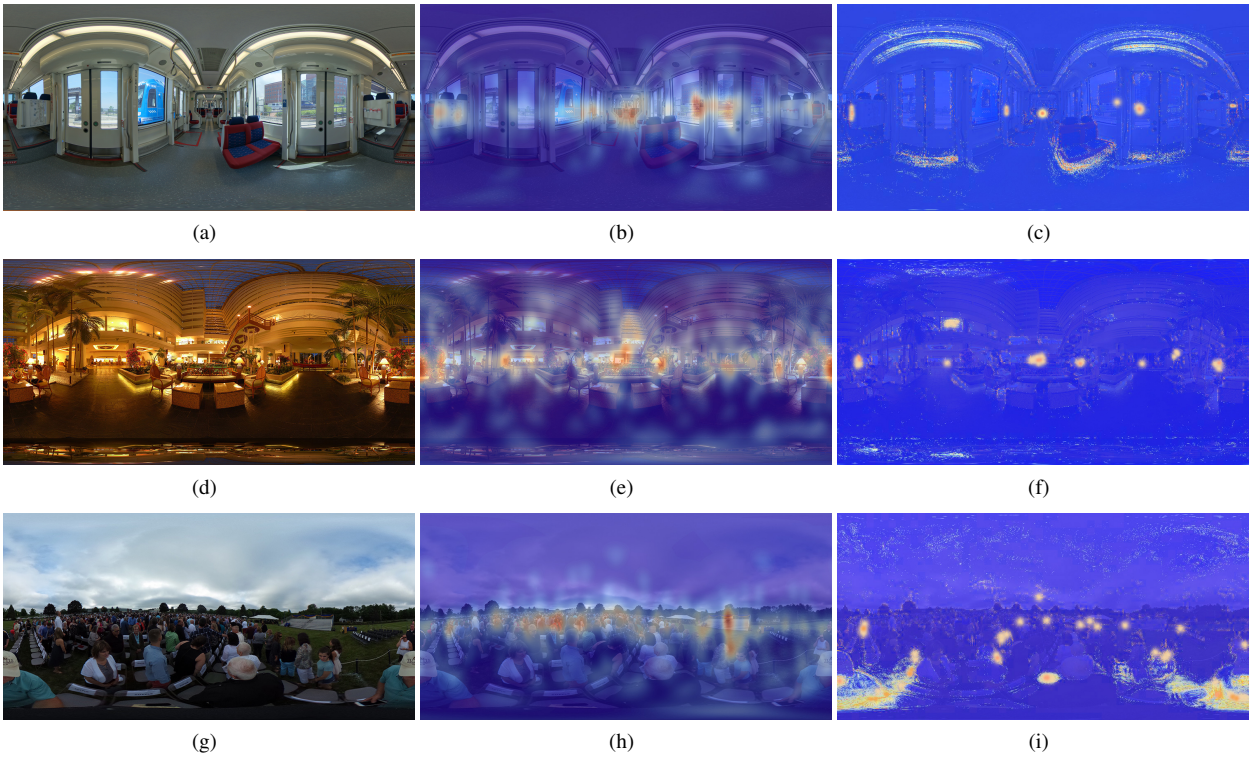
Image dataset

The test images used for assessing the performances of the proposed method have been selected from the evaluation database of the "Salient360! Grand Challenge", created for the head only model [12]. The dataset consists of 25 equirectangular images (with size from 3000 x 1500 to 12000 x 6000 pixels) with the corresponding ground truth saliency map. The dataset includes a wide range of shooting conditions, locations (i.e., buildings, hotels, museums, theaters, markets), and subjects (i.e., animals, natural landscapes, people, vehicles) and urban environment. Furthermore, indoor and outdoor sets with different lighting conditions are included in the dataset.

Experimental results

To evaluate the performances of the proposed method, for each image of the test database, the estimated saliency map is compared with the provided ground truth saliency map. The performances of the proposed method are assessed by computing the Linear Correlation Coefficient (CC) and the Kullback-Leibler Divergence (KL). A lower value for the KL divergence represents higher similarity. Table 1 compares the proposed model with the solutions presented at the Salient360! Challenge [13] that resulted to be the best performing ones (Wuhan University (WU) and Zhejiang University (ZU)) and with the previous version of the proposed algorithm (RM3). These numbers were provided by the organizers of the challenge. It is worthful to notice that only the aggregate numbers are available. In this table the performances are evaluated in terms of average CC and KL values. As can be noticed, the proposed method shows increased performances with respect to the previous one. To further evaluate the method, an in-deep analysis of its performances has been carried out by analyzing the content of each image.

The content analysis highlights that the proposed method



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3. *Left column: omnidirectional image, central column: corresponding ground truth map, right column: estimated saliency map for SRCs P1 (a-c), P16 (d-f), and P69 (g-i).*

Method	CC	KL
ZU	0.69	0.44
WU	0.71	0.51
RM3	0.52	0.81
Proposed	0.56	0.64

Table 1: CC and KL mean scores obtained on the evaluation dataset by the proposed method with respect to the RM3, ZU, and WU methods.

has the best performances with images characterized by good lighting conditions, large sections of homogeneous colors (i.e., images P94 and P8) ad present features that can be easily captured with the adopted monocular cues (i.e., image P71 reported in Figure 4 or images P1 and P16, shown in Figure 3). In more

(a)

Figure 4. *Figure P71.*

details, in image P71 it can be noticed that the part of the crowd closer to the observer is more detailed and less dense than the farther one. This situation is perfectly captured by the texture gradient detection. The proposed algorithm also shows good performances for images containing aisles and thus outlining the depth element. In Table 2, the results for the best performing cases are reported. As can be noticed, for these images the results are better than the aggregated average values obtained in the challenge (as reported in Table 1).

Metric	P1	P8	P16	P26	P71	P73	P94
CC	0.64	0.66	0.58	0.69	0.57	0.69	0.70
KL	0.51	0.50	0.36	0.36	0.56	0.35	0.45

Table 2: CC and KL scores for the proposed method in the best performing cases.

The performed analysis highlights the presence of some weak points. For example, in case of images P69 (see Figure 3 and P96, the bad lighting conditions represent an issue for monocular depth estimation (the CC values for those images are 0.48 and 0.44 respectively).

Concluding remarks

In this contribution, a novel method for visual saliency estimation for omnidirectional images is presented. It is based on the fusion of information typical of 2D images (i.e., texture, edges) with information about the depth of the scene. The performances of the method, evaluated by comparing the estimated saliency maps with respect to the available ground truths, confirm the validity of the proposed approach while maintaining a low computational complexity. However, the system is sensible to conditions of low luminance of the scene and the study of this aspect is the base for ongoing work.

Acknowledgements

The authors would like to thank Chiara Scotumella for her support in developing the proposed saliency estimation method.

References

[1] 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, Jan 2013.

[2] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," *arXiv preprint arXiv:1510.07748*, 2015.

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

[4] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior." in *BMVC*, vol. 6, no. 7, 2011, p. 9.

[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, IEEE 12th international conference on*, 2009, pp. 2106–2113.

[6] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Computer Vision, IEEE International Conference on*, 2011, pp. 914–921.

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

[8] M. Kümmerer, T. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[9] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[10] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *Transaction on Image Processing*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016. [Online]. Available: https://doi.org/10.1109/TIP.2016.2602079

[11] Y. Fang, C. Zhang, J. Li, J. Lei, M. P. D. Silva, and P. L. Callet, "Visual attention modeling for stereoscopic video: A benchmark and computational model," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4684–4696, Oct 2017.

[12] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: ACM, 2017, pp. 205–210. [Online]. Available: http://doi.acm.org/10.1145/3083187.3083218

[13] U. of Nantes and Technicolor, "Salient360!: Visual attention modeling for 360 images grand challenge," http://www.icme2017.org/grand-challenges/.

[14] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *European Conference on Computer Vision*. Springer International Publishing, 2016, pp. 809–824.

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

[16] M. Hosseini and V. Swaminathan, "Adaptive 360 vr video streaming based on mpeg-dash srd," in *2016 IEEE International Symposium on Multimedia (ISM)*, Dec 2016, pp. 407–408.

[17] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, and G. Wetzstein, "Saliency in VR: how do people explore virtual environments?" *CoRR*, vol. abs/1612.04335, 2016. [Online]. Available: http://arxiv.org/abs/1612.04335

[18] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli, "A feature-based approach for saliency estimation of omni-directional images," *Signal Processing: Image Communication*, vol. 69, pp. 53 – 59, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092359651830242X

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

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

[21] E. Goldstein, *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education, 2010.

[22] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision research*, vol. 47, no. 19, pp. 2483–2498, 2007.

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

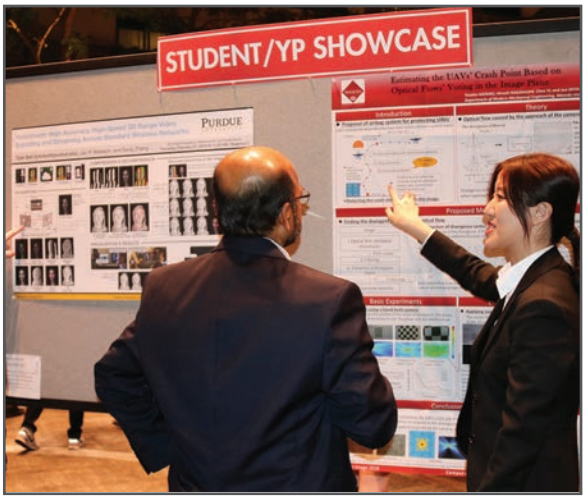
JOIN US AT THE NEXT EI!

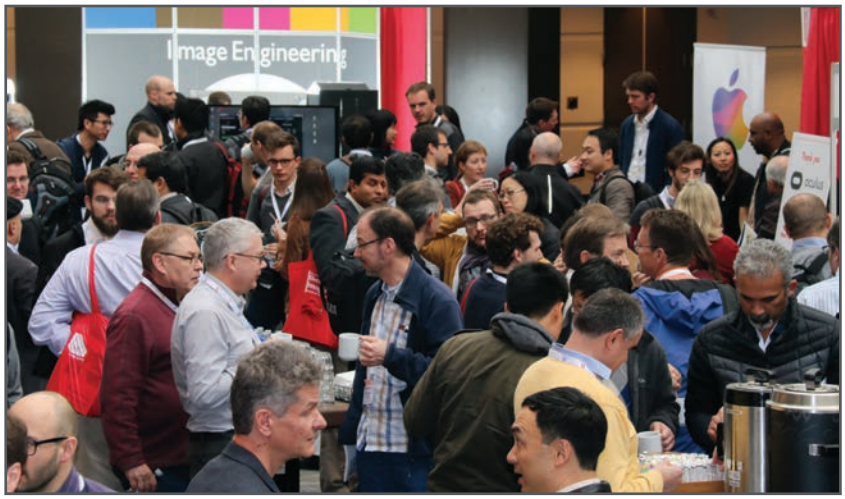
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org