Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360° contents

Jesús Gutiérrez, Erwan J. David, Antoine Coutrot, Matthieu Perreira Da Silva, Patrick Le Callet
LS2N UMR CNRS 6004
Université de Nantes, France
Email: salient360@univ-nantes.fr

***Abstract*—Virtual Reality (VR) provides the users with new immersive media experiences, offering the possibility to freely explore 360° content. Understanding these new exploration behaviors is crucial for the development of efficient techniques for processing, coding, delivering and rendering omnidirectional content to offer the highest possible Quality of Experience (QoE). Progress has already been made on visual attention (VA) modeling for 360° content. In this paper we briefly review the current status of research on this topic that led us to propose a benchmarking platform for evaluating and comparing the performance of models for saliency and scanpath prediction for 360° content. This paper introduces the 'UN Salient360! benchmark" platform featuring a dataset, a toolbox and a framework for evaluation of different class of models. This online platform can be found in https://salient360.ls2n.fr/.**

***Index Terms*—Omnidirectional content, visual attention, saliency, scanpath, benchmarking, 360° content.**

I. Introduction

An extensive research effort has been made in the last years to develop Visual Attention (VA) models for 2D and even stereoscopic 3D images and videos [1], [2]. In this effort, benchmarking the proposed models is essential to identify their performance and applicability to practical applications such as coding, transmission, and rendering (in the area of multimedia communications) [3][4][5].

The recent emergence and development of Virtual Reality (VR) and 360° content applications has led to the research on VA for omnidirectional content. In fact, some works have already been carried out analyzing VA in this type of content to take into account the novelties that this technology provides to the users, in terms of exploring the represented scene [6][7][8]. Unlike traditional 2D viewing, where users watch a screen with a fixed head position, in 360° content users can freely explore the scene, seeing different content according to their head positions. These novelties affect VA and hinder the direct application of VA models originally developed for traditional technologies.

Furthermore, this new way of consuming audiovisual content increases the need for reliable VA models, in order to obtain efficient approaches for several applications, such as coding [9], streaming [10], foveated rendering, cinematography [11], movie editing [12], and Quality of Experience (QoE) evaluation [7].

Taking into account the current status of VA research for 360° content, benchmarking is required to help on the development and identification of appropriate models. Therefore, this paper introduces a benchmarking platform for saliency and scanpath models for omnidirectional images and videos, including a benchmark dataset, evaluation tools and a framework for a continuous performance evaluation and ranking of the proposed models.

The rest of the paper is structured as follows. Section II provides an overview of the state-of-the-art on VA for 360° content, which led to propose a platform for benchmarking of saliency and scanpaths models for omnidirectional content, described in Section III. Finally, Section IV draws some conclusions.

II. State-of-the-Art on VA for 360° content

Some research has already been done related to VA on VR and 360° content. For instance, a precursory study was carried out by Marmitt and Duchowski [6] focused on analyzing visual scanpaths in VR environments. Then, after the recent development of immersive media technologies (e.g., HMD devices, tracking tools, VR services, etc.) and the emergence of 360° content for entertainment applications, some more works have been published taking into account how users consume omnidirectional content. In this sense, the first approximation is to study how the observers explore this type of content analyzing their head movements, usually by using the head motion trackers integrated within the Head-Mounted Displays (HMD) [13][14]. Therefore, this data could be used to optimize media processing systems (e.g., coding [9], transmission [10]) and applications (e.g., automatic cinematography [11]).

Although head movement could be a valuable proxy for VA in certain cases, eye movements are also crucial to understand how people explore 360° content. Thus, with the development of eye-trackers suitable to be integrated in HMDs, a few studies have been carried out investigating both head and eye movements when visualizing omnidirectional content [8][12],

The work of Jesús Gutiérrez was supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE Programme coordinated by Campus France. The work of Erwan David was supported by RFI Atlanstic2020.

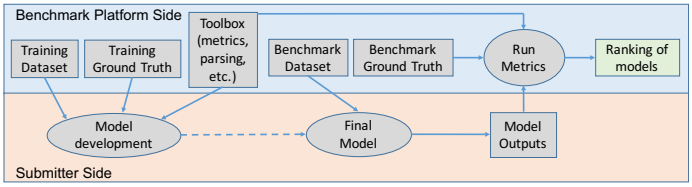

Fig. 1. Graphical overview of the benchmark workflow

and even considering how the outcomes of theses studies can be applied for applications such as quality assessment [7].

A critical aspect on the research on VA for 360° content is the availability of public datasets containing appropriate stimuli and as much information as possible about the exploration behavior of the users. Thus, some datasets have already been published with data from subjective experiments that allowed the collection of tracking data from groups of participants. In this sense, some video datasets are available containing head-movement data from observers [15][16][17][18], while the availability of eye tracking data is much more limited due to the need of specific eye trackers. Nevertheless, some datasets have been published containing both head and eye movements information for 360° images [19][8] and videos [12].

Finally, some efforts have already been done towards the development of VA models for omnidirectional content. In particular, it is worth pointing out the organization of the "Salient360!" Grand Challenge at ICME'17 that supported the development of saliency and scanpaths models for 360° images, providing a dataset with eye and head tracking data and an evaluation of the proposed models. This activity resulted in the publication of several models for saliency maps considering only head movements, saliency maps taking also into account eye movements, and scanpaths for head and eye movements [20].

III. Benchmarking VA models for 360° content

The proposed framework aims at a continuous benchmark of VA models for 360° content, extending the concept of the MIT saliency benchmark [21] (for 2D images). The proposed web-based platform (https://salient360.ls2n.fr/) is based on the following points:

- Accepted models: Four types of VA models are addressed, both for images and video, aiming at predicting: 1) saliency maps derived from the head movements only, 2) saliency maps derived from the head movements as well as from eye movements within the viewport, 3) scanpaths of eye gaze (i.e., considering head and eye-movement), and 4) scanpaths of head-gaze (i.e., center of the viewport as a succession of head-positions).
- Benchmark dataset: A benchmark dataset is used for evaluating the performance of the models (see details in Subsection III-A). The 360° stimuli (images and videos) are made available, although the corresponding ground-truth data of head and eye movements is kept secret for fair benchmarking (avoiding the possibility of training models with this information). In this sense, training datasets (totally different form the benchmarking dataset) are also provided by the organizing team containing both the stimuli and the eye and head movement data.
- Benchmark tools: A toolbox is provided to compute metrics for comparing saliency maps and scanpaths to assess the performance of the models. A detailed description of those metrics is provided in Subsection III-B. It is worth noting that this is a first approach, so future findings of the research community regarding metrics on this topic will be considered.
- Workflow: The process of benchmarking VA models is based on asking the interested people to run their models in the benchmark dataset of stimuli, and provide their results (saliency maps and/or scanpaths) to the benchmark platform. The performance evaluation is then carried out by the organizing team (according to the benchmark metrics), and a ranking for each type of model is published on the website.

A. Generation of the benchmark dataset

The ground-truth data, for the considered four types of models consist on saliency maps (in equi-rectangular format) and scanpaths (sequences of fixations). This ground truth is generated from the head and/or eye movements of the participants in a subjective test, who freely explore the benchmark stimuli using an HMD with an embedded eye-tracker.

The raw tracking data is firstly processed to generate the ground-truth saliency maps and scanpaths. In particular, the eye-tracking data is firstly parsed to identify fixations and saccades. This process should be carried out in the spherical domain to avoid other projection issues [7]. Then, once this classification is made, the fixations are used to generate the saliency maps and the scanpaths based on eye and head data. For head-only saliency maps and scanpaths, the head-gaze is set in the center of the viewport, and head trajectories are defined by sampling the raw head-tracking data coming from the HMD.

B. Evaluation metrics

1) Comparing saliency maps: To benchmark the models that estimate saliency maps (both from head-only data and head and eye data), the outputs of the models are compared to ground-truth saliency maps. Several studies have already been published proposing and analyzing metrics for comparing saliency maps for conventional images [22][4], while for videos usually a frame-by-frame comparison is carried out, and the values are pooled to obtain a global score [23]. In particular, the metrics considered to compare saliency maps are: Linear Correlation Coefficient (LCC), Kullback-Leibler divergence (KLD), Normalized scanpath saliency (NSS), Area under the curve (AUC), and Information Gain (IG) [5].

To apply these metrics to omnidirectional content, it is important to take into account distortions caused by projection issues. In particular, since equi-rectangular saliency maps are considered in the proposed benchmarking platform, the stretching at the poles of this image format should be taken into account. The most appropriate way to do this is to take

the saliency values of the equi-rectangular saliency maps that correspond to the points resulting from an uniform sampling of the sphere [19]. Then, the aforementioned metrics could be used over these saliency values.

2) Comparing scanpaths: Several metrics have also been proposed for comparing scanpaths in traditional content [24][25][26]. However, their performance and reliability seem to be very dependent on the specific case, meaning that they are probably not as established as the metrics for comparing saliency maps [22].

Taking this into account, the metric proposed by Jarodzka *et al.* [27] has been considered for benchmarking scanpaths (both models based on head-only movements and based on head and eye movements). In essence, this metric allows a comparison between the sequence of fixations from two scanpaths in terms of similarity aspects, such as spatial proximity between the fixations, difference in their direction and magnitude, and temporal proximity.

Normally, for a given stimuli, several scanpaths are obtained coming from each participant in the experiment for generating the ground truth. Thus, to benchmark scanpath models, the models should provide as many scanpaths as contained in the ground-truth, so that all of them would be compared. Then, the results are averaged to obtain a global score. With this aim, an optimizer algorithm is used to make a one-to-one match between the scanpaths so the best-matches among all of them are considered.

Since this metric was originally proposed for still 2D images, some modifications were made to fit it to this specific benchmark case. In particular, to deal with 360° content, the orthodromic distance is used instead of the euclidean distance proposed in the original metric. In addition, to deal with video content, the temporal alignment of the compared scanpaths used in the original metric is avoided, since temporal misalignment is meaningful when comparing scanpaths in dynamic content [27].

IV. Conclusion

The main contributions of this paper are twofold. On one side, the paper provides an overview of the status of the research on VA for 360° content, especially covering exploratory experiments and datasets with tracking data. On the other side, and as a consequence of this current status, a benchmarking platform is introduced for evaluating and comparing the performance of models for saliency and scanpath prediction for omnidirectional content, which includes a dataset, tools and a framework for model evaluation.

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, 2013.

[2] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2058–2067, 2013.

[3] T. Judd, F. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," *MIT Technical Report*, vol. 1, pp. 1–7, 2012. [Online]. Available: http://dspace.mit.edu/handle/1721.1/68590

[4] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" pp. 1–24, Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.03605

[5] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency Benchmarking: Separating Models, Maps and Metrics," Apr. 2017. [Online]. Available: http://arxiv.org/abs/1704.08615

[6] G. Marmitt and A. T. Duchowski, "Modeling visual attention in vr: Measuring the accuracy of predicted scanpaths," in *Eurographics 2002, Short Presentations*, Saarbrücken, Germany, Sep. 2002, pp. 217–226.

[7] Y. Rai, P. Le Callet, and P. Guillotel, "Which saliency weighting for omni directional image quality assessment?" in *Proc. International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

[8] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018.

[9] M. Yu, H. Lakshman, and B. Girod, "A Framework to Evaluate Omnidirectional Video Coding Schemes," in *IEEE International Symposium on Mixed and Augmented Reality*, Sep. 2015, pp. 31–36.

[10] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," *IEEE International Conference on Communications*, 2017.

[11] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2Vid: Automatic Cinematography for Watching 360° Videos," no. 1, Dec. 2016.

[12] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Trans. on Graphics*, vol. 36, no. 4, p. 47, 2017.

[13] B. Hu, I. Johnson-Bey, M. Sharma, and E. Niebur, "Head movements during visual exploration of natural images in virtual reality," in *Proc. Annual Conf. on Information Sciences and Systems*, Mar. 2017, pp. 1–6.

[14] E. Upenik and T. Ebrahimi, "A simple method to obtain visual attention data in head mounted virtual reality," in *IEEE International Conference on Multimedia & Expo Workshops*, Jul. 2017, pp. 73–78.

[15] X. Corbillon, F. De Simone, and G. Simon, "360-Degree Video Head Movement Dataset," *Proc. ACM on Multimedia Systems Conference*, pp. 199–204, Jun. 2017.

[16] C. Wu, Z. Tan, Z. Wang, and S. Yang, "A Dataset for Exploring User Behaviors in VR Spherical Video Streaming," *Proc. ACM on Multimedia Systems Conference*, pp. 193–198, Jun. 2017.

[17] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "360 Video Viewing Dataset in Head-Mounted Virtual Reality," *Proc. ACM on Multimedia Systems Conference*, pp. 211–216, Jun. 2017.

[18] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams, "A Public Database of Immersive VR Videos with Corresponding Ratings of Arousal, Valence, and Correlations between Head Movements and Self Report Measures," *Frontiers in Psychology*, vol. 8, no. Dec, Dec. 2017.

[19] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. ACM Multimedia Systems Conference*, Taipei, Taiwan, Jun. 2017.

[20] Salient360, "Special issue," *Signal Processing: Image Communication*, 2018, to appear.

[21] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

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

[23] 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, Sep. 2007.

[24] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behavior Research Methods*, vol. 47, no. 4, pp. 1377–1392, Dec. 2015.

[25] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of Scores, Datasets, and Models in Visual Saliency Prediction," in *IEEE International Conference on Computer Vision*, Dec. 2013, pp. 921–928.

[26] T. C. Kübler, "Algorithms for the comparison of visual scan patterns," Ph.D. dissertation, Eberhard-Karls-University Tübingen, 2016.

[27] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multidimensional scanpath similarity measure," in *Proc. Symposium on Eye Tracking Research & Applications*. ACM, 2010, pp. 211–218.