SALIENCY DRIVEN PERCEPTUAL QUALITY METRIC FOR OMNIDIRECTIONAL VISUAL CONTENT

Evgeniy Upenik and Touradj Ebrahimi

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

ABSTRACT

The problem of objectively measuring perceptual quality of omnidirectional visual content arises in many immersive imaging applications and particularly in compression. The interactive nature of this type of content limits the performance of earlier methods designed for static images or for video with a predefined dynamic. The non-deterministic impact must be addressed using statistical approach. One of the ways to describe, analyze and predict viewer interactions in omnidirectional imaging is through estimation of visual attention. We propose an objective metric to measure perceptual quality of omnidirectional visual content considering visual attention information.

Index Terms— omnidirectional imaging, virtual reality, visual attention, perceptual quality

1. INTRODUCTION

Omnidirectional visual content is a particular form of immersive multimedia which extends conventional image and video sensations to a three-dimensional space by providing full-spherical coverage of field of view and allowing change-of-sight interactions. This type of content is typically consumed using virtual reality (VR) head-mounted displays (HMD), hand-held devices, and, less frequently, conventional displays of personal computers. Viewers perform interactions by moving their heads, displacing and rotating an accelerometer-equipped device or by means of direct controllers such as computer mice, track-pads and touch-screens.

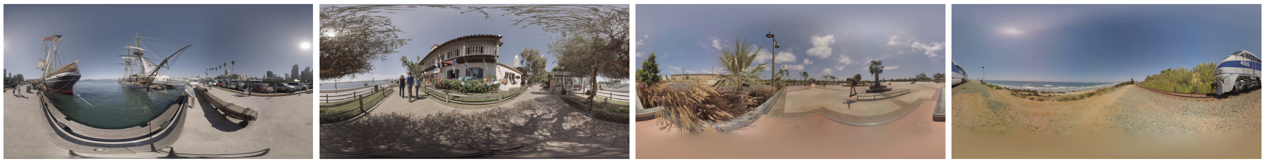
Interactivity is a property of omnidirectional visual content as well as other immersive media which distinguishes them drastically from conventional images and video. It introduces an additional non-deterministic component among the factors influencing perception of this type of content by humans. At any given moment a viewer sees only a subset of the entire omnidirectional image or video frame which is called a viewport. Thus, there may exist a case when an observer does not explore every part of an image. Hence, particular regions acquire more significance and provide higher impact on perceived visual quality, whilst other regions contribute less to it. One way to take this factor into consideration is to collect statistical data of user interactions in order to estimate visual attention or saliency.

Current research on visual attention in omnidirectional images and virtual reality is mainly represented by two trends: one concerns obtaining visual attention information from experimental data involving human viewers, whilst another concentrates on prediction of salient regions using algorithmic approaches. The problem of obtaining visual attention empirically is investigated by researchers in [1–6]. These works provide analysis of eye and head movements during consumption of VR content and propose several methods to process raw experimental data in order to obtain saliency maps. Prediction of salient regions using the algorithmic approach is studied in [7–10] and advocate mostly adaptation of earlier conventional saliency prediction methods described in [11, 12]. Deep learning approaches to predict visual saliency in omnidirectional visual content are presented in [13–15].

State-of-the-art research on perceptual visual quality assessment of omnidirectional content mainly focuses on adaptation of conventional full-reference objective metrics in order to cope with geometrical distortions and spatial entropy redistribution introduced by different representations of such content. A review along with benchmarking results of recently proposed objective quality metrics for omnidirectional visual content is provided by authors in [16, 17]. Among the proposed metrics methodology varies from applying forward-and-backward geometrical mappings as in [18] to different schemes of weighting during pixel-wise comparison as in [19–21]. Croci et al. propose in [22] a framework for perceptual visual quality control in stereoscopic omnidirectional imaging. Their method considers empirical visual attention data to define significance of regions.

In this paper, we propose yet another approach to incor-

This paper reports research performed under the framework of the project Digital Eye: Deep Learning Video Quality Assessment Technology funded by The Swiss Commission for Technology and Innovation (CTI) under the grant 27403.1 PFES-ES.



(a) Harbor (b) KiteFlite (c) SkateboardTrick (d) Train

Fig. 1. Contents selected for experiments.

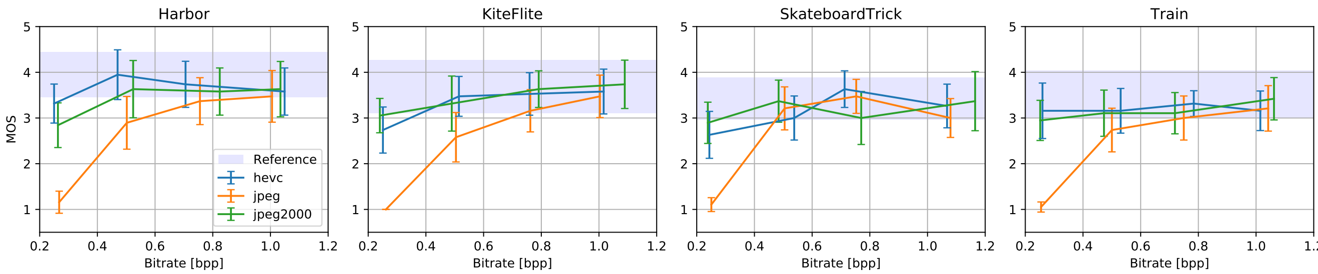


Fig. 2. Subjective mean opinion scores (MOS) with 95% confidence intervals. The area filled with transparent purple color depicts the 95% confidence interval of the hidden reference.

porate visual attention data into a full-reference objective perceptual visual quality measurement.

2. VISUAL ATTENTION WEIGHTED METRIC

In this section, we propose an objective perceptual visual quality metric which takes into account ground-truth viewer's visual attention information in order to make image quality assessment selective with respect to regions of interest.

As a base for our method we choose to use Peak Signal to Noise Ratio (PSNR) metric because it is widely accepted, its implementation is simple, and its performance is satisfactory to test our hypothesis. We define a ground-truth image as $I(i,j)$, where $i = 0, 1, ..., H, j = 0, 1, ..., W$, with W and H being dimensions of the image. The impaired image is defined as $\hat{I}(i,j)$. Thus, PSNR is described by the following equation:

$$PSNR = \frac{MAX_I^2}{MSE}$$

where

$$MSE = \frac{\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(I(i,j) - \hat{I}(i,j) \right)^2}{H * W}$$

and MAX_I is the maximum possible value of pixel intensity of the assessed image, e.g. for an 8-bit image it equals 255.

Given that sufficient amount of empirical data of head movements is available for an assessed omnidirectional image, one can obtain a visual saliency map using a method described in [1].

Hence, the saliency map can be defined as:

$$h_{i,j} \in [0,1], i = 0, 1, ..., H, j = 0, 1, ..., W$$

where each pixel of $h_{i,j}$ provides a visual attention value for each corresponding pixel of $\hat{I}(i,j)$. The saliency map $h_{i,j}$ can be obtained independently for different degradation levels of impaired images. This issue is further addressed in Section 3.3.

Visual saliency map is used to compute a saliency-weighted mean square error MSE_{VA} which contributes to PSNR equation as a denominator.

$$MSE_{VA} = \frac{\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(I(i,j) - \hat{I}(i,j) \right)^2 h_{i,j}}{\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} h_{i,j}}$$

Therefore, a Visual Attention PSNR (VA-PSNR) is defined as:

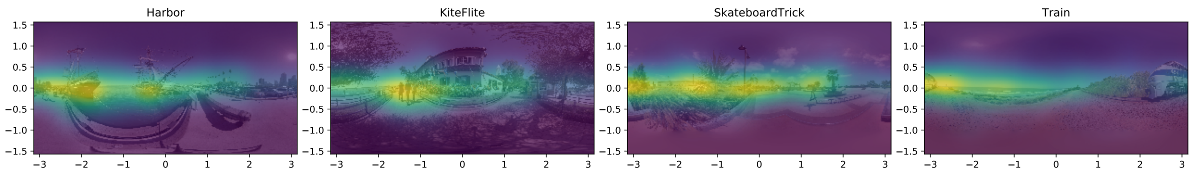
$$PSNR_{VA} = \frac{MAX_I^2}{MSE_{VA}}$$

VA-PSNR allows comparison of two omnidirectional images regardless of the projection (equirectangular, cubic, etc.) they are represented in, provided that both are represented in the same.

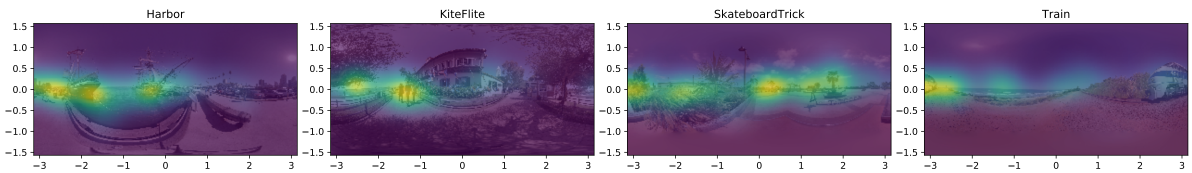
The source code and data are publicly available on-line at: `https://github.com/mmspg/saliencymetric360`

3. SUBJECTIVE EXPERIMENTS

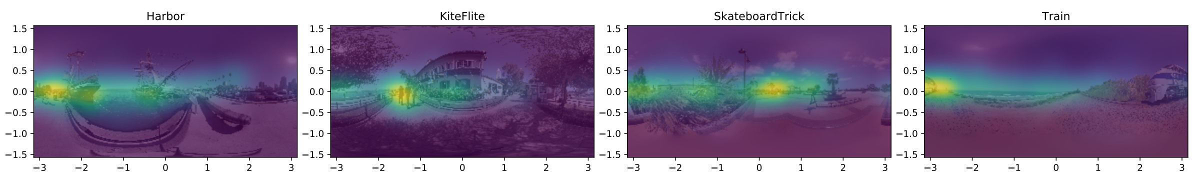
Two independent content viewing sessions were conducted. Participants were divided in two disjoint groups: one was



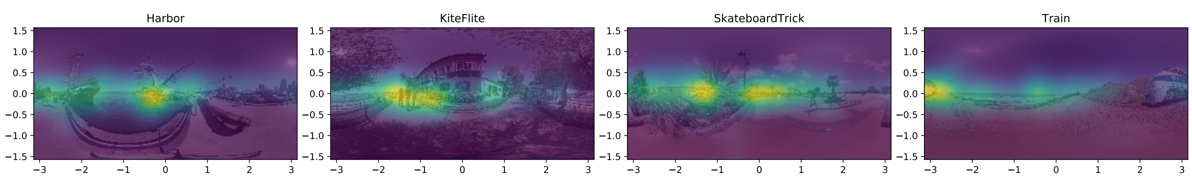
(a) evaluation experiment: all quality levels



(b) evaluation experiment: stimuli rated by subjects as high quality



(c) evaluation experiment: stimuli rated by subjects as low quality



(d) exploration experiment

Fig. 3. Visual attention heatmaps obtained from experimental data.

asked to evaluate omnidirectional images according to visual quality, whilst another performed free exploration with a dummy task to assess aesthetic value of the pictures. It is interesting to observe that although other datasets [23] have been proposed, they are not task dependent.

3.1. Dataset and Equipment

Four still images extracted from test sequences of MPEG omnidirectional video dataset were selected for the experiments as depicted in Figure 1. The contents were compressed using three different codecs, namely JPEG, JPEG 2000, and HEVC Intra-frame. The software used was the same as in [16] with the quality parameters specified in Table 2. Original images were downscaled to 5760×2880 pixels before compression in order to comply with technical requirements of the display.

Experiments were conducted with the help of a testbed for subjective evaluation of omnidirectional content proposed in [17] which is publicly available for downloading[1]. Participants were observing stimuli using a head-mount[2] with a mobile device acting as a screen. Galaxy S7 Edge SM-G935F was used. The resolution of the device is 2560×1440 pixels. During the experiments, subjects were sitting on a rotating chair. All subjects passed color vision and visual acuity tests.

3.2. Evaluation and Exploration

During an evaluation experiment subjects were assessing omnidirectional images following the methodology called Absolut Category Rating with Hidden Reference (ACR-HR). They were asked to rate stimuli on the five-level scale 5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, and 1 - Bad. 19 subjects participated in the evaluation session, among which 9 were females, with an overall median age of 24.5. Results of subjective assessment are presented in Figure 2.

Exactly the same set up as for evaluation was used in an exploration experiment. However, subjects were asked to evaluate aesthetic value of the pictures and only uncompressed stimuli were used. Their subjective scores were dis-

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

[2]https://mergevr.com

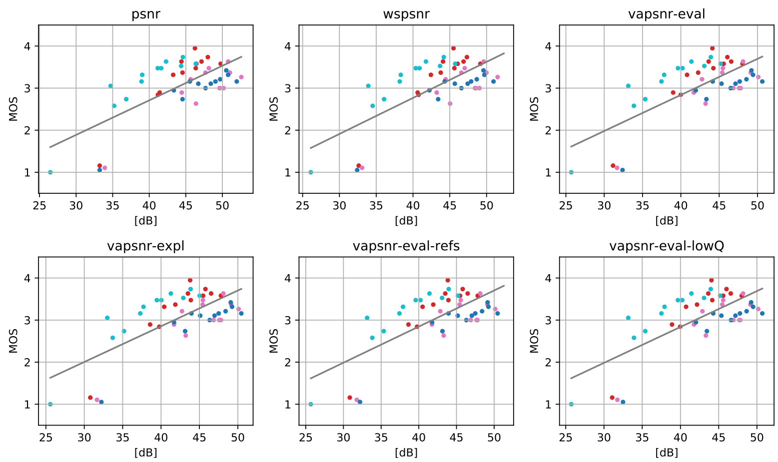


Fig. 4. Mapping of objective scores to subjective ratings. Grey line depicts linear fitting. Different colors represent different contents: blue - Train, red - Harbor, cyan - SkateboardTrick, magenta - KiteFlite.

	PSNR	WS-PSNR [19]	VA-PSNR Eval	VA-PSNR Expl	VA-PSNR Eval-Refs	VA-PSNR Eval-LowQ
PLCC	0.6959	0.7106	0.7107	0.7074	**0.7114**	0.7083
SRCC	0.3706	0.4131	0.4131	0.4075	**0.4163**	0.4080
KRCC	0.2706	0.2976	**0.3012**	0.2904	0.2976	0.2958

Table 1. Standard performance indexes. Pearson linear correlation coefficient (PLCC), the Spearman rank correlation coefficient (SRCC), and Kendall rank correlation coefficient (KRCC). Bold text shows the best result per index.

Codec	Harbor	KiteFlite	Skateboard..	Train
JPEG	9,53,79,87	4,23,54,73	8,71,87,93	8,65,85,92
JPEG 2000	41,44,46,47	35,39,42,44	44,47,49,51	43,46,48,50
HEVC-I	32,27,24,21	37,30,26,23	29,23,21,18	30,24,21,19

Table 2. Quality "Q" parameters used to encode images.

carded and only head direction tracks were collected. Exploration sessions had 17 participants, of which 10 were females, with an overall median age of 24.3.

3.3. Visual Attention and Quality

Head direction tracks were collected from both evaluation and exploration experiments. They were processed according to the method described in [1] in order to produce saliency maps. Additionally, raw visual attention data from evaluation sessions were grouped into three categories: all tracks, tracks from stimuli which have Mean Opinion Scores (MOS) lying withing the 95% confidence interval of hidden reference, and with MOS lower then 3.0. The resulting saliency maps are depicted in Figure 3.

4. VALIDATION AND DISCUSSION

The proposed method is essentially an extension of PSNR. Thus, it was benchmarked against other PSNR-based metrics. VA-PSNR and other metrics were computed for all the stimuli using each set of saliency maps described in Section 3.3.

Standard performance indexes were calculated (Table 1) after applying linear fitting to the data as it depicted in Figure 4. Notably, VA-PSNR-Refs computed using saliency maps from high quality evaluation stimuli outperforms VA-PSNR-Expl, VA-PSNR-Eval, and VA-PSNR-lowQ computed using maps from exploration sessions, from all evaluation tracks, and from low quality evaluation stimuli tracks respectively.

The proposed method requires empirical visual saliency data and it can be applied in post-production of cloud services where, after a certain time from the moment of initial release, sufficient amount of data can be collected and used *a posteriori* to estimate quality during re-compression of the content which can be beneficial for saving bandwidth.

5. CONCLUSION

In this paper, we proposed a new method called VA-PSNR which estimates perceptual quality of omnidirectional content considering visual attention. We validated our method against subjective MOS and benchmarked it against state-of-the-art objective metrics. VA-PSNR shows better performance when compared to alternative approaches based on PSNR.

6. REFERENCES

[1] E. Upenik and T. Ebrahimi, "A Simple Method to Obtain Visual Attention Data in Head Mounted Virtual Reality," in *IEEE International Conference on Multimedia and Expo 2017*, Hong Kong, 2017.

[2] 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*, vol. 24, no. 4, pp. 1633–1642, 2018.

[3] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, and Y. Wang, "A subjective study of viewer navigation behaviors when watching 360-degree videos on computers," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[4] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

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

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

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

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

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

[10] J. Ling, K. Zhang, Y. Zhang, D. Yang, and Z. Chen, "A saliency prediction model on 360 degree images using color dictionary based sparse representation," *Signal Processing: Image Communication*, vol. 69, pp. 60–68, 2018.

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

[12] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?," in *Computer Vision ECCV 2016*, 2016, pp. 809–824.

[13] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 videos," 2018, pp. 488–503.

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

[15] A. Borji, "Saliency prediction in the deep learning era: An empirical investigation," *arXiv:1810.03716 [cs]*, 2018.

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

[17] E. Upenik, M. Rerabek, and T. Ebrahimi, "Testbed for subjective evaluation of omnidirectional visual content," in *2016 Picture Coding Symposium (PCS)*, 2016.

[18] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Proceedings Volume 9970, Optics and Photonics for Information Processing X*, 2016, vol. 9970.

[19] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[20] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[21] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2015, pp. 31–36.

[22] S. Croci, S. Knorr, L. Goldmann, and A. Smolic, "A framework for quality control in cinematic VR based on voronoi patches and saliency," in *2017 International Conference on 3D Immersion (IC3D)*, 2017, pp. 1–8.

[23] J. Gutiérrez, E. J David, A. Coutrot, M. Da Silva, and P. Le Callet, "Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–3.