10.2352/ISSN.2470-1173.2021.9.IQSP-296

On the Influence of Head-Mounted Displays on Quality Rating of Omnidirectional Images

Abderrezzaq Sendjasni [1,2]***, Mohamed-Chaker Larabi*** [1] ***and Faouzi Alaya Cheikh*** [2]
[1] ***CNRS, Xlim UMR 7252, Université de Poitiers, France***
[2] ***NTNU, Norwegian Colour and Visual Computing Lab, Gjøvik, Norway***

Abstract

In this paper, we propose to study the influence of using head-mounted displays (HMDs) for visual quality rating of omnidirectional images and their impact on the final quality scores. Because of the used display technology, these devices introduce a significant impairment called screen door effect that may alter the quality of experience. Furthermore, the extended use of such a technology may produce cyber-sickness. In this study, a subjective experiment is designed and carried out using various HMDs with various types of content. The statistically analysed results revealed a significant difference between HMDs for quality rating tasks on the overall ratings as well as per individual distortions. These findings will contribute to the development of a reliable protocol for omnidirectional subjective quality assessment, and the constructed database will be used as a ground-truth for quality metrics development.

Keywords: *Omnidirectional images, Perceptual quality, Subjective assessment, Simulator sickness, Head-mounted displays.*

Introduction

Immersive content attracted a significant attention recently, both on the commercial and the research sides. One can notice this in a variety of applications such as entertainment, healthcare, education, and so on. This type of technology provides users with an immersive experience thanks to multiple virtual reality (VR) applications. Omnidirectional images/videos (*a.k.a.* 360-degree) are among the most important content in VR. It is created by taking multiple shots of the scene in all directions at the same time, and then stitching them all together to create a panoramic scene. Such an omnidirectional content is intended to provide users with the ability to adjust the viewing angle within the 360-degree in order to explore the scene. Head-mounted displays (HMDs) made this exploration possible by providing appropriate viewports, offering a higher immersive experience compared to traditional displays. Commercial HMDs differ in terms of field of view (FoV), resolution, magnitude of the screen door effect (SDE), etc. Even if the same content is used, this could vary the immersive experience. Furthermore, discomfort and dissatisfaction of the users caused by possible cyber-sickness induced by the HMD or nature of content can lead to a dodgy assessment. All of these factors may have an impact on the overall quality of the experience (QoE) and the immerssiveness of the user. Determining the impact of such a technology appears important in order to advise for possibility of improvement.

Subjective quality assessment (SQA) is the most reliable method for assessing QoE. It requires a large number of human observers and a significant amount of time and resources. It is typically used to create databases for training and testing objective quality assessment methods. Furthermore, the subjective ratings could be used as a foundation for developing new algorithms. Since SQA must be carefully performed in order to produce accurate results, the international telecommunication union (ITU) and the video quality expect group (VQEG) developed guidelines on how to perform such experiments. Unfortunately, at the time of writing this paper, there is still no guidelines or recommendations on how to conduct a reliable subjective experiments for immersive applications.

Because it is still in its infancy, omnidirectional image quality assessment is facing significant limitations. Currently, there are no extensive studies reporting on the impact of HMDs on perceived quality on the one hand. On the other hand, as HMDs may induce discomfort, the quantification of simulator sickness has only been the subject of a very limited number of studies. Besides the caves, the use of HMDs is unavoidable for quality ratings of immersive application. In the literature, one can find some studies on VQA of omnidirectional content [1, 2, 3, 4, 5] which involve the use of HMDs. In these studies, different devices from different manufacturers are used. In particular, the authors in [1] considered studying the impact of the HTC Vive and the Oculus Rift using different content in terms of resolution (4k, 1080p). They also recorded head movements of the viewers to determine their behaviours. Here the focus was rather on the resolution and not the HMD itself. Similarly, the effect of different resolutions on perceived quality is studied in [2, 5], where HTC Vive and HTC Vive Pro quality ratings are compared. In addition, the study of [5] included the impact of pixel density on the perceived quality. Hence, a high-resolution monitor is used by adjusting the distance between the viewer and the screen to obtain different densities. Here, they demonstrated that with a higher density, quality improves until a saturation at values greater than 60 pixels per degree (PPD), which corresponds to the retina resolvable resolution [6]. In [3], the authors assessed the cyber-sickness caused by high motion omnidirectional content. To do so, they proposed to isolate the camera motion from other factors defining anchors to control the gaze of the viewers. The influence of viewing methods on subjective evaluation is evaluated in [4] including free-viewing, fixed trajectory viewing, and content-dependent viewing modes. The subjective evaluation appears to be strongly affected by the viewing mode. In the above-mentioned studies, neither the SDE nor the impact of different HMDs were considered. Moreover, only a very few studies involving the use of various HMDs, can be found in literature. To the best of our knowledge, there has been no extensive study on the impact of HMDs on the perceived quality of omnidirectional images. Besides, image quality assess-

ment (IQA) requires the use of reliable databases. Currently, the existing ones are lacking diversity in terms of content and cannot be considered as representative of the field. Our first exploration showed that one of the first proposed databases presents a weak correlation with objective quality metrics [7].

In this work, we evaluate the impact of various market HMDs, either related to technology or rendered content, on perceived quality. First, we explore whether the use of different HMDs results in different quality ratings for the same content and conditions. For this, we build a dataset and define a controlled paradigm to conduct subjective experiments for such omnidirectional content. Then, we study the comfort of the viewers by means of the simulator sickness questionnaires. A statistical study of the obtained results is performed so as to compare HMDs and draw conclusions related to quality of experience. In this study, four HMDs are used including Varjo Vr-2 [8], HTC Vive Pro [9], HP Reverb VR and Oculus Quest [10]; each having specific characteristics.

Subjective Quality Assessment

Omnidirectional image databases

The availability of reliable and representative databases is a critical factor in developing image quality models. It allows obtaining accurate and well-generalized IQA models. Particularly, machine-learning (ML) based models, where the performance is only as good as the variety of the available data. Unfortunately, there is a significant lack of omnidirectional image quality databases. Table. 1 summarizes the characteristics of three available ones in terms of number of reference/distorted images, number of subjects participating to the subjective experiments, quality distortion types, and the used HMD. In the following, we discuss each database by providing its characteristics.

Huang et al. [11] : It contains 25 pristine omnidirectional images used to create 12 versions for each one. Four distinct spatial resolutions and three JPEG quality factors (25, 60, and 100) are used to create 300 distorted images. The resolutions are 4k, 2K, and 1080p. The MOS was obtained using absolute category rating (ACR) with 98 subjects participating to the test (53 males and 45 females) which is a very high number compared to state-of-the-art. The quality scale ranges from 0 (Bad) to 100 (Excellent). Each subject rated only three different image contents at the four spatial resolution and three quality factors.

CVIQD2018 [12] : This dataset is composed of 16 omnidirectional images and 528 compressed versions. The compression artefacts are obtained using eleven levels of : 1) JPEG compression with quality factors ranging from 50 to 0, 2) H.264/AVC and H.265/HEVC with quantization parameters from 30 to 50. The authors used the ACR method with a rating scale of 10-levels from the lowest to the highest quality to gather the MOS. The ACR method was adopted with the participation of 20 subjects (14 males and 6 females).

OIQA [13] : It includes 320 distorted omnidirectional images created from 16 reference ones using four distortion types with five levels each. The used distortions include JPEG compression (JPEG), JPEG 2000 compression (JP2K), Gaussian blur (BLUR) and Gaussian white noise (WN). JPEG and JP2K are applied directly on ERPs, while GB and WGN are applied on small blocks individually that are stitched back to ERP. Subjective scores are given in the range from 1 (bad) to 10 (excellent). 20 subjects were involved in the test (15 males and 5 females).

In comparison to Huang et al. [13], only CVIQD [12] and OIQA [13] databases have received attention in the literature. In addition, the study in [7] showed a poor correlation between MOS provided with Huang et al. [11] and objective metrics. This obviously shows the ineffectiveness of certain databases in contrast to others, which may raise questions about their representativeness and reliability.

The Proposed 360-IQAD Database

The selected content to create our database is composed of twenty images, from which 240 distorted versions are created. First, we chose omnidirectional images from the joint video exploration team (JVET) test sequences [14] and the SUN360 database [15]. In addition, to account for synthesized content related to VR, four scenes have been added to the dataset. The used images are given in Fig. 1. As it can be seen, the images represent a variety of content types, including indoor and outdoor natural scenes, as well as synthesized ones. Often, datasets are constructed without paying attention to the diversity of the content. In our case, we account for two important characteristics, *i.e.* spatial complexity and colourfulness. Spatial information (SI) index represents an indicator of edge energy, giving an idea about the complexity of an image. Colourfulness information (CFI) is a perceptual indicator of the variety and intensity of colours in the image. SI and CFI are calculated according to ITU-T P.910 [16] recommendations and the metric described in [17], respectively. The CFI versus SI plot of the pristine images shown in Fig. 2 aims at demonstrating the spatial and colour diversity of the selected images. One can notice that the used images span over the range of CFI and SI values. This shows the diversity of the content provided in the database.

Once selected, the images are distorted using the JPEG compression, Gaussian blur (GB), and Gaussian noise (BN). For each distortion type, four levels are applied to cover the perceived quality range from annoying to imperceptible. The levels are purposefully chosen in such a way that the perceived difference between them is obvious for observers. Thereby, 12 distorted images are created per pristine one.

Subjective Assessment Protocol

In order to construct a reliable database, the selection of subjective protocol is of paramount importance. Unfortunately, there are no guidelines for conducting experiments for immersive applications. In our case, we built the test by relying on the ITU recommendation ITU-BT.500 [18]. Hence, the adopted protocol is depicted in Fig. 3. It is scrupulously followed by each observer. First, the observer is screened for visual acuity and colour blindness in order to collect reliable scores. Then he is asked to complete a simulator sickness questionnaire (SSQ) before beginning the test. For this aim, we used the virtual reality sickness questionnaire (VRSQ) proposed in [19]. The VRSQ consists of nine questions in which the observer is asked to rate the severeness of nine symptoms on a four scale (None: 0, Slight: 1, Moderate: 2, Severe: 3). Individual symptoms are classified into three categories: oculomotor agitation (O), disorientation (D) and to-

Table 1: Description of state-of-the-art omnidirectional image databases.

Database	Ref images	Distorted images	Distortion type (Distortion level)	Number of subjects	HMD
Huang et al. [11]	25	300	JPEG (3) / Down-sampling (4)	98	HTC Vive
CVIQD [12]	16	528	JPEG (11) / AVC (11) / HEVC (11)	20	HTC Vive
OIQA [13]	16	320	JPEG (5) / JPEG2000 (5) / GB (5) / WGN (5)	20	HTC Vive

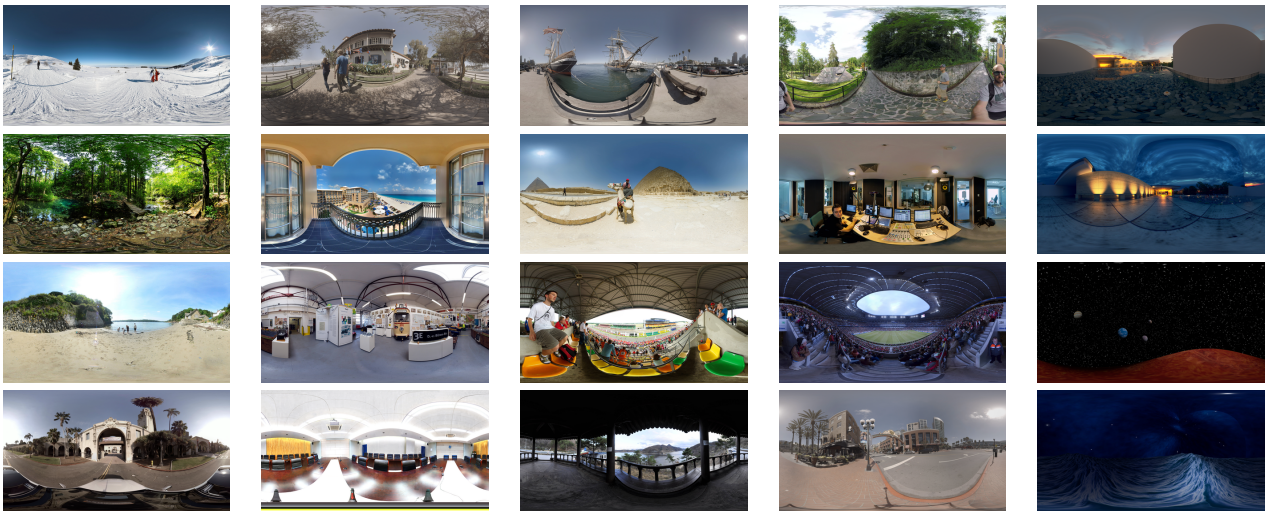

Figure 1: Pristine images in the proposed database. First to fourth rows are images taken from JVET and SUN360. Fifth row are created as synthesized images.

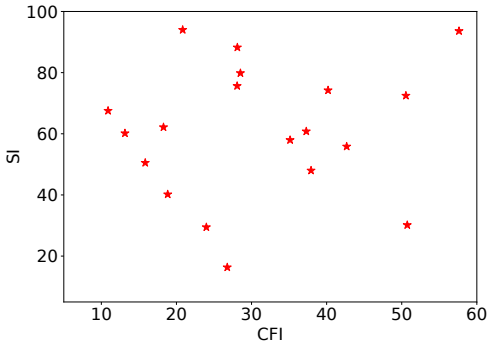

Figure 2: Colourfulness index (CFI) versus spatial information (SI) plot of the selected pristine images for the construction of the database.

tal score (TS). The use of SSQ prior to beginning the test serves as the observer's initial state and is used to compare the progression of the symptoms described in the VRSQ. It is known that conducting a subjective test in the morning or the afternoon may lead to a different assessment because of the psycho-visual state of the observer. The observer in then trained on a few samples of omnidirectional images with perceptual qualities corresponding to those used in the test. The training is designed to familiarize the observer with the task at hand, and get familiar with VR environment, since not all observers are VR or HMD users. Samples used in this session are for training purpose only and are discarded from the experiment results.

After the training session, the first session starts by asking the observer to rate the quality of the impaired omnidirectional images using a five-point quality scale ranging from 5 (excellent) to 1 (bad), following the ACR method. This quality scale should be sufficient to cover the quality levels used in the constructed database and where the maximum quality corresponds to the pristine images. In the first session, the observer rates a hundred and thirty images, corresponding to a duration of 32.5 min (130 samples $\times 15s$). This duration is reasonable as the test deals with images only (*i.e.* there are no motions as in the case of videos). In addition, the observer can stop the test any time based on his condition. After the first session, the observer fill out another SSQ so to record his state after experiencing VR for approximately half an hour. After a sufficient break, the second session takes place with the remaining images from the database. Finally, at the end of the second session, another SSQ is filled out. In order to collect reliable results, we ensured that all observers followed the exact same protocol. The images playlists are randomly constructed, and each observer watch a random one in order to avoid rating biases.

The observers were recruited from our university, and they are all naive. Due to the current sanitary situation (Covid-19 pandemic), running subjective experiments becomes very challenging. In our case, the experiment with four HMDs lasts for about six hours for a single observer. This is why, the results exposed in this paper are based on eight valid observers.

The HMDs considered in this study are from different manufacturers, and have specific characteristics each. Table 2 summarizes the ones that may contribute to the quality assessment task.

Table 2: Characteristics of the considered HMDs.

HMD	Resolution per eye	FoV	PPD
Varjo Vr-2	1920×1080	87°	60
HTC Vive Pro	1440×1600	110°	13.09
HP Reverb VR	2160×2160	114°	18.94
Oculus Quest	1600×1440	100°	14.4

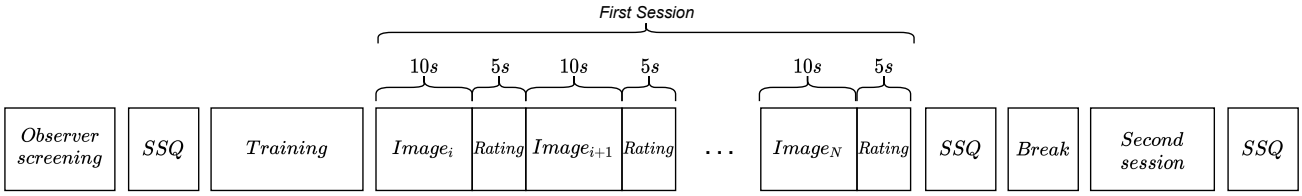

Figure 3: Illustration of the adopted subjective assessment protocol.

Results and Discussion

In the following, we provide and analyse the subjective quality evaluation results. First, we investigate whether there is a substantial difference in terms of quality ratings between HMDs on the one hand and particular distortions on the other hand. The last part of this section will concentrate on analysing the simulator sickness questionnaire results.

Effects of HMD On The Rating

It is known that, the use of different devices for SQA may result in different outcomes since each device has unique properties. In our case, the device is the HMD. It is critical to establish whether such a difference is substantial. Especially when it may have an impact on the overall QoE. To that purpose, various questions are framed in order to determine the impact of using HMDs for omnidirectional image SQA. In this study, these questions are roughly summarized as follows:

- Would the use of various HMDs result in a different rating?
- What is the inter-observer difference?
- Is the impact of a single distortion the same regardless of the used HMD?
- Which HMD offers a better quality?
- What about comfort and cyber-sickness?

Thanks to a statistical analysis of the obtained scores, we aim to find answers to the above questions. The histograms of the gathered rating scores of all HMDs are shown in Fig. 4. We can clearly observe that, the ratings are distributed across the five perceptual quality scales.

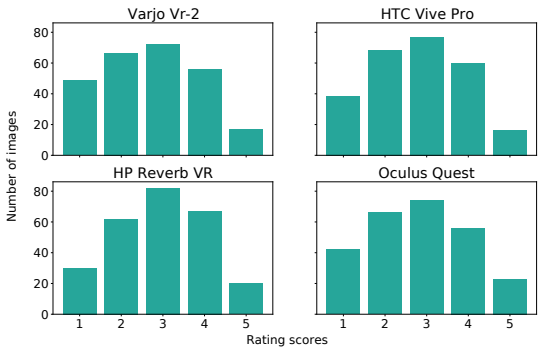

Figure 4: Histograms of the rating scores obtained by the used HMDs.

$$SOS(x)^2 = -ax^2 + 6ax - 5a \tag{1}$$

Prior to investigating the effect of HMDs on ratings, we calculated the standard deviation of opinion scores (SOS) using the SOS hypothesis described in [20], which is defined by Eq. 1. The SOS parameter a quantifies the uncertainty ratio among observers on a scale of 0 to 1. It reflects the inter-observer reliability. A value of 0 denotes a full agreement among all observers, and 1 indicates a maximum variance. Table. 3, provides the a value regarding the obtained scores by the four HMDs. As it can be seen, the SOS parameter for all HMDs is around $0 < a < 0.06$. Based on the description given above, this interval of values demonstrates an inter-observer agreement and reliability of approximately 90%. This observation substantiates the overall efficacy of the constructed experiments and the adopted procedure.

Table 3: SOS parameter a of all HMDs' rating scores.

Varjo Vr-2	HTC Vive Pro	HP Reverb VR	Oculus Quest
0.0361	0.0414	0.0338	0.0504

In order to statistically assess the impact of HMDs on the quality rating, we analyse the variance between the obtained MOSs. The following are the null hypothesis H_0 and the alternative one H_1:

H_0: There is no significant difference between the four HMDs.

H_1: At least one HMD is significantly different from the others.

To analyse the variance, the use of ANOVA [21] is a good choice. However, the ANOVA assumes that the sample data is normality distributed. Therefore, a normality check is performed, and the probability distribution for each HMD is illustrated in Fig. 5. The formula used for the theoretical quantiles (horizontal axis of the probability plot) is Filliben's estimate [22]. Looking at the plots, we see an upward sloping linear relationship. Deviations by the dots from the line can be observed around both extremities. The sample data (*i.e.* MOS) partially fits the diagonal line, which shows a deviation from the expected normal distribution. This demonstrates that the distribution of the gathered MOSs is not perfectly normal but close. Based on this observation and in order to reliably analyse the variance, a non-parametric test is applied in addition to ANOVA. Here, the Kruskal-Wallis H-test [23] is used.

The ANOVA showed a $p-value$ of 0.035 while Kruskal a $p-value$ of 0.038, leading to the rejection of H_0, implying that the HMD has a statistically significant influence on the quality ratings. One possible explanation could be the screen door affect explained previously. This observation contrasts with the results stated in [1], which found that the effect of HMDs is not significant compared to other factors. Since a statistical difference is found, we further analyse specific differences between HMDs. A post hoc test [24] is performed in this case, and a significance plot is provided in Fig. 6. It appears that, the source of the identified differences is significant between HP Reverb VR and HTC

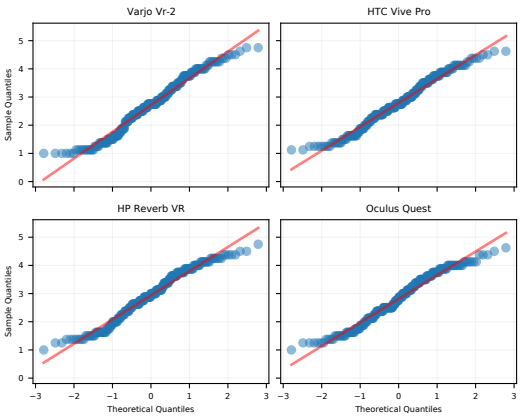

Figure 5: Probability plot of MOS against the normal distribution quantiles.

Vive Pro, and more with Varjo Vr-2. While Oculus Quest provides no significant difference with any of the selected HMDs. This demonstrates that the variance occurs from multiple HMD and not just a single one.

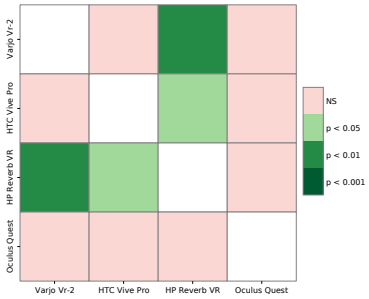

Figure 6: Pairwise multiple significance plot between HMDs for the overall MOS. NS stands for no significance.

In addition to the analysis of variance on the overall scores, we performed an analysis of variance using the Kruskal-Wallis H-test between HMDs per individual distortions, with the aim of evaluating the effect of a single distortion independently of the used HMD. A $p-value$ of 0.023, 0.274, and 1 are achieved for JPEG, GB, and GN, respectively. In this case, observers noticed a difference between HMDs for JPEG but not for the remaining distortions. One may question the link between JPEG artefacts and the SDE, presenting some similarities in terms of distortion type (*i.e.* blocking artefacts). This observation backs up the previous one about the differences on the overall ratings. Additionally, we looked into the differences regarding the JPEG distortion, the significance difference is depicted in Fig. 7. One can notice that the difference here is between Varjo Vr-2 and Oculus Quest, as well as with HP Reverb VR. Compared to the differences on the overall MOSs, Varjo Vr-2 is one of the common source. The significant difference in PPD (*see* table 2) between this HMD and the other ones, which greatly contributes to visual quality, may explain such result.

We examined the MOS obtained for all HMDs to determine which HMD provides the best quality, and how the MOS per individual distortions is distributed. A box plot of MOSs from all images per single distortion is depicted in Figure 8. Overall, we can notice that MOS for a certain distortion level are mostly within a limited range. This backs up the findings of the SOS parameter, which was previously discussed. One can also notice that, compared to JPEG and GN, GB was frequently rated as bad (1) and poor (2). Especially, levels 3 and 4 where the means falls in the same range. This clearly shows that the observers were annoyed by such a distortion regardless of the used HMD. For GN, the MOS mostly falls in the same range for level 2, 3 and 4, as if the observers did not perceive much difference between these levels. Particularly with HTC Vive Pro and Varjo Vr-2. In terms of which HMDs provides a better quality, we can observe that with Varjo VR-2 and HP Reverb VR more MOS greater than 3.5 were given. This suggests that these two offer a better quality, which can be related to their resolution and PPD (*see* Table. 2).

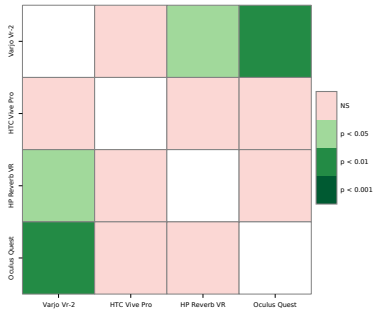

Figure 7: Pairwise multiple significance plot between HMDs for the JPEG distortion. NS stands for no significance.

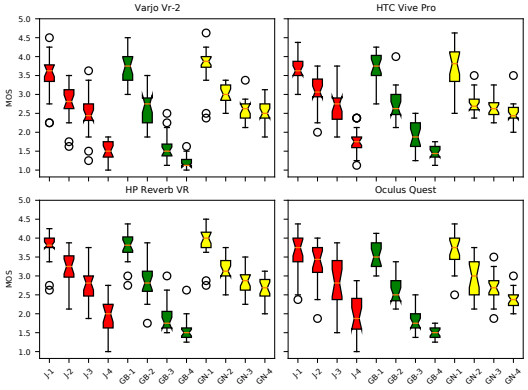

Figure 8: Box plot of MOS per level of distortion, where J-*, GB-* and GN-* stand for JPEG, Gaussian Blur and Gaussian Noise with 4 different levels.

Simulator Sickness Assessment

We computed the simulator-sickness scores, as mentioned previously, to measure the sickness level caused by each of the used HMDs. The scores are grouped by total scores (TS), oculomotor (O), and disorientation (D) as described in the VRSQ [19]. The VRSQ is derived from SSQ [25] where 9 symptoms are selected among 16. Scores for TS, O and D are calculated using the method in VRSQ [19], where the TS is the average score of O and D. Here, a score around 40 is considered severe. Fig. 9 shows the histograms of the simulator-sickness scores obtained for the

selected HMDs.

For this experiments, we focused on analysing which HMD causes higher sickness in terms of TS, O and D. In comparison to the others, the Varjo VR-2 received the highest overall scores, while, the Oculus Quests received the lowest. Two explanations could convey these results. First, the weight of Varjo VR-2, reported by the observers as being high. Then, the double displays composing this HMD with two different resolutions, often requiring an adapted content. One can also see that the observers are more prone to oculomotor symptoms compared to those for disorientation and even the TS. The length of the sessions where the observers are subject to very close displays may be a reason for this. As the oculomotor involves eye strain, difficulty focusing, and fatigue, which can be increased with more exposition to VR.

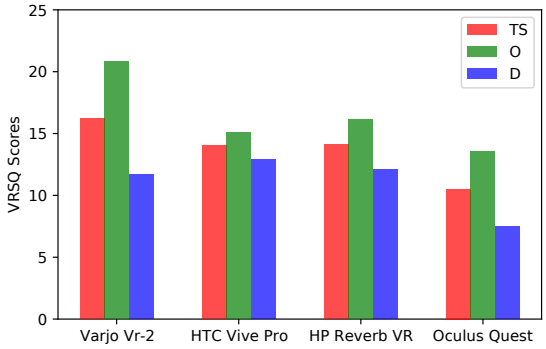

Figure 9: Simulator-sickness scores for the considered HMDs in terms of total scores (TS), oculomotor (O), and disorientation (D).

Conclusion

In this paper, we present a detailed evaluation of the impact of HMDs on the quality rating of omnidirectional images. The provided analysis revealed a statistically significant difference between the used HMDs. This difference is mostly related to the distinct characteristics of each HMD. Especially the SDE, which can be confused with the distortions on the viewed scenes, and may lead to a dodgy assessment. This contrast with previous observations in the literature. Furthermore, a significant difference on specific distortions was also observed with JPEG compared to GN and GB. The source of such difference was found between multiple HMDs supporting the observations regarding the device induced influence. Additionally, the simulator-sickness assessment revealed that the use of some HMDs lead to a higher simulator sickness scores compared to others, and oculomotor related symptoms induce significantly higher scores when compared to disorientation. A further analysis including additional factors is planned in order to provide a holistic assessment regarding the use of HMDs for subjective quality ratings.

Acknowledgments

This work is funded by the Nouvelle-Aquitaine's regional council under project SIMOREVA360 2018-1R50112 and project CPER/FEDER e-immersion.

References

[1] 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 *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, Erfurt, Germany, 2017.

[2] F. Hofmeyer, S. Fremerey, T. Cohrs, and A. Raake. Impacts of internal hmd playback processing on subjective quality perception. In *Electronic Imaging*, pages 219–1, Burlingame, California, USA, 2019.

[3] P. Perez, N. Oyaga, J. J. Ruiz, and A. Villegas. Towards systematic analysis of cybersickness in high motion omnidirectional video. In *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3, Cagliari, Italy, 2018.

[4] W. Zhang, W. Zou, F. Yang, L. Lévêque, and H. Liu. The effect of spatio-temporal inconsistency on the subjective quality evaluation of omnidirectional videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4055–4059, Brighton, United Kingdom, 2019.

[5] W. Zou, L. Yang, F. Yang, Z. Ma, and Q. Zhao. The impact of screen resolution of hmd on perceptual quality of immersive videos. In *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, London, United Kingdom, 2020.

[6] CV. Curcio and KA. Allen. Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1):5–25, 1990.

[7] A. Sendjasni, MC. Larabi, and FA. Cheikh. On the improvement of 2D quality assessment metrics for omnidirectional images. In *Electronic Imaging*, pages 287–1, Burlingame, California USA, 2020.

[8] Varjo vr-2 pro. `https://varjo.com/products/vr-2/`. Online; accessed 20 November 2020.

[9] Vive pro. `https://www.vive.com/eu/product/vive-pro/`. Online; accessed 20 November 2020.

[10] Oculus quest. `https://www.oculus.com/quest/features/`. Online; accessed 20 November 2020.

[11] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao. Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression. *IEEE Transactions on Image Processing*, 27(12):6039–6050, 2018.

[12] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai. A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In *IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–6, Vancouver, BC, Canada, 2018.

[13] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang. Perceptual Quality Assessment of Omnidirectional Images. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, Florence, Italy, 2018.

[14] W. Sun and R. Guo. Test sequences for virtual reality video coding from letinvr. *Joint Video Exploration Team (JVET) of ITU-T SG*, 16, 2016.

[15] J. Xiao, K. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In

IEEE Conference on Computer Vision and Pattern Recognition, pages 2695–2702, 2012.

[16] TU Berlin Pierre Lebreton. Siti. `https://vqeg.github.io/software-tools/quality%20analysis/siti/`. Online; accessed 30 November 2020.

[17] D. Hasler and S. Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. International Society for Optics and Photonics, 2003.

[18] ITUR BT. 500-14. BT. 500: Methodologies for the subjective assessment of the quality of television images. *International Telecommunications Union: Geneva, Switzerland*, 2019.

[19] H. Kim, J. Park, Y. Choi, and M. Choe. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied ergonomics*, 69:66–73, 2018.

[20] T. Hoßfeld, R. Schatz, and S. Egger. Sos: The mos is not enough! In *Third International Workshop on Quality of Multimedia Experience*, pages 131–136, Mechelen, Belgium, 2011.

[21] ER. Girden. *ANOVA: Repeated measures*. Number 84. Sage, 1992.

[22] JJ. Filliben. The probability plot correlation coefficient test for normality. *Technometrics*, 17(1):111–117, 1975.

[23] WH. Kruskal and WA. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[24] M. Terpilowski. scikit-posthocs: Pairwise multiple comparison tests in python. *The Journal of Open Source Software*, 4(36):1169, 2019.

[25] RS. Kennedy, EB. Lane, KS. Berbaum, and MG. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.