IMPACT OF VISUALISATION STRATEGY FOR SUBJECTIVE QUALITY ASSESSMENT OF POINT CLOUDS

Evangelos Alexiou and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
Emails: FirstName.LastName@epfl.ch

ABSTRACT

Point clouds have recently emerged as a promising and practical solution to code 3D visual information for immersive applications. Among other challenges, objective and subjective quality assessments are still open problems for this type of visual data representation. In this paper, we investigate the impact of already proposed subjective evaluation methodologies in order to assess the visual quality of point clouds in different display environments (e.g. on a desktop versus an augmented reality head-mounted-display) creating different types of experiences to users. Advantages and drawbacks of the above visualization strategies are compared to each other based on a rigorous statistical analysis.

***Index Terms*—** point cloud, rendering, subjective quality assessment

1. INTRODUCTION

As modern information and communication systems progressively support more immersive applications, richer content representations are required in order to increase the engagement of the user, leading to enhanced experiences that stimulate user's senses and better approximate the perception of real-world scenes. Point clouds are expected to be the prevalent representation in several of such advanced applications due to a number of desirable advantages, such as low complexity and high efficiency in capturing, encoding, and rendering of 3D static and dynamic contents. The recent activities of standardization bodies indicate the great deal of interest that has been recently drawn on this type of imaging.

Nowadays, point clouds are typically visualized on either conventional monitors or head-mounted-displays (HMDs). Depending on the type of visualization, different rendering equipment and different levels of interactions are offered to users, which may affect the perception and, by extension, the visual quality of a content. The latter, particularly, is of fundamental importance and is commonly assessed subjectively based on the ITU-R Recommendation BT.500-13 [1]. However, it is not clear whether variations of conventional methodologies proposed by various ITU recommendations, typically limited to passive evaluation, or radically different approaches that exploit the full potentials of advanced content representations, are better to adequately assess the visual quality of immersive 3D contents.

Subjective quality assessment of point clouds has attracted a great amount of interest recently, with the adoption of different test methods, evaluation scenarios, and types of degradations. Zhang et al. [2] performed subjective quality assessment of colored point clouds in a desktop set-up, after applying both geometry and color degradations. Mekuria et al. [3] assessed dynamic colored point clouds that were captured in real-time by multiple Microsoft Kinect sensors. These acquired models represented avatars that were navigating in a virtual environment. The performance of the proposed codec, deployed to encode dynamic contents, was evaluated in this 3D tele-immersive system. Javaheri et al. [4] subjectively assessed the efficiency of point cloud denoising algorithms. For visualization of the denoised contents, the Screened Poisson surface reconstruction algorithm was used. In [5], the same authors conducted subjective quality assessment of colored point clouds, whose geometry was compressed under Octree- and graph-based encoding schemes, while the original color attributes remained uncompressed. The test contents were visualized using cubes as primitives. In both studies, passive visualization was adopted. In our previous work [6], we proposed and compared interactive ways for subjective quality assessment of point cloud geometry, using Double-Stimulus Impairment Scale (DSIS) and Absolute Category Rating (ACR). In [7], we proposed the use of an HMD for subjective evaluation in an augmented reality scenario. In these studies, visual degradations based on Gaussian noise and Octree-based encoding were employed, and the test contents were displayed as collections of points.

To the best of our knowledge, the impact of adopting dif-

This work has been conducted in the framework of ImmersiaTV project under the European Union Horizon 2020 research and innovation program (grant agreement no. 688619) and funded by Swiss State Secretariat for Education, Research and Innovation SERI.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

ferent visualization strategies in subjective quality assessment of point clouds is not reported in the literature. In this paper we tackle this problem by comparing subjective scores collected by two utterly different experimentation settings configured for subjective evaluation of geometry-only point clouds under two types of degradations. Our results reveal that the correlation is highly affected by the type of degradation under assessment.

2. TESTBEDS DESCRIPTION

In this section the equipment that was employed and the configurations settings used in experimental set-ups are briefly described. The experiments were conducted in a test laboratory which fulfills the ITU-R Recommendation BT.500-13 [1] for subjective evaluation of visual contents.

2.1. Experiment A: Desktop set-up

In this experiment, test subjects were able to visualize the stimuli in a 30-inch Apple Cinema Display with a resolution of 2560x1600 pixels. The test contents were displayed as sets of points and each point was represented by one atomic pixel. Test subjects visualized the contents under evaluation on a graphical user interface deployed in Point Cloud Library (PCL) [8]. Their interactions with the stimuli were performed using the mouse cursor, and their scores were submitted using the keyboard. The color of the graphical user interface background was set to black, while the color of the point clouds was set to white, to increase the contrast while avoiding any distractions. The luminance values of the points and the background were measured on the flat screen as 354 and 0.5 nits, respectively. For more details, the reader may refer to [6].

2.2. Experiment B: Head-mounted-display set-up

In this experiment, test subjects were able to visualize the contents in an iPhone 6S which was used in conjunction with an Occipital Bridge AR headset. The real-world scene was scanned using a wide angle lens of 120-degree field of view, which was attached to the camera of the iPhone. The virtual assets were added on top of the real scenery, defining an augmented reality scenario. The resolution of the phone screen was 326 pixels per inch. The test contents were displayed as collections of points, and each point was represented by an atomic triangle of minimum size. As the atomic triangles were significantly smaller than the dimensions of the virtual model, they were perceived as points by test subjects. The background was a real-world environment with colors involving different shades of grey, while the color of the points was set to white. The models under assessment were placed on top of a test table covered by a medium grey tissue. The luminance values of the points and the test table surface were measured on the screen as 595.28 and 38.91 nits, respectively. At the beginning of each test content evaluation, subjects were asked to stand in front of the test table at the distance of 1 meter. After inspecting the test content from this starting point, they were free to interact by changing their positions in the real world without any constraints. After completing the assessment, subjects provided their scores orally after listening to the list of rating scale to select from. The order of rating scale was provided indentically for every stimulus and every test subject. For more details, the reader may refer to [7].

3. SUBJECTIVE EXPERIMENTS

In this section the design of the subjective quality evaluation experiments of both experimental set-ups is briefly described and motivated. A detailed specification can be found in [6].

3.1. Content preparation

Only the point cloud geometry was assessed in the experiments. *Bunny* and *dragon* are selected from the Stanford 3D Scanning Repository to represent contents with regular geometry and reduced amount of noise. *Cube* and *sphere* are artificially generated and represent synthetic contents with highly regular geometry. Finally, *vase*, is a 3D model manually captured, and constitutes a representative point cloud with irregular structure that can be acquired by low-cost depth sensors.

The target application of such contents involves scenarios where the users may visualize point clouds from the outside and interact by either rotating or moving around them. These use cases typically occur when simple objects are scanned by sensors that provide, either directly or indirectly, a cloud of points to represent their 3D shapes. To form a representative data set, the contents were selected considering the following properties: (a) Simplicity, as it would have been difficult for subjects to clearly perceive a complex scene in the absence of texture. Although simple, the complexity of contents covers a reasonable range. (b) Diversity of geometric structure, as different artifacts may be observed by applying different types of degradations. Thus, test contents used were generated by different means. (c) Similarity of points density, as the visual quality of point clouds is directly affected by the number of points used to represent an object. The contents were also scaled to fit in a bounding box of size 1. In Figure 1 the selected contents and their number of points are illustrated.

3.2. Types of degradations

Two different types of geometric degradations are assessed: (a) Gaussian noise, and (b) Octree-pruning. The first type of distortion is widely used to simulate position errors due to depth sensor imperfections, or errors introduced after stereoscopic triangulation. The coordinates of every point of the content are affected by this type of noise, as its spatial posi-

(a) Bunny (35947) (b) Cube (30246) (c) Dragon (22998) (d) Sphere (30135) (e) Vase (36022)

Fig. 1: Selected contents.

tion is modified in every dimension following a target standard deviation ($\sigma = \{0.0005, 0.002, 0.008, 0.016\}$).

The second type of degradation is based on an Octree, which is a suitable structure to represent a point cloud with known maximum geometric distortion error and is used by several encoding solutions. Octree-pruning is the procedure of removing points after setting a desirable Level of Details (LoD) value, which defines the size of a leaf node. This representation leads to a structural loss with points removal and displacement. In this experiment, the LoD is set appropriately for each content to achieve target percentages (ρ) with respect to the original number of points, allowing an acceptable deviation of $\pm 2\%$ ($\rho = \{30\%, 50\%, 70\%, 90\%\}$).

3.3. Subjective evaluation methodology

Since two different types of degradations were assessed, two different sessions were established in every set-up. Each session was launched after a training phase, where the subjects got familiarized with the artifacts caused by the corresponding type of degradation. Furthermore, the subjects learnt how to use the installed equipment to interact with the contents.

The simultaneous DSIS test method with 5-rating impairment scale was selected in both experiments for a side-by-side visualization. The order of ratings was identically provided to the subjects for every stimulus, in each session. Regarding the selection of a test method, although the ACR was found to achieve high discriminative power, the DSIS was chosen as it is more consistent for the identification of the level of distortion [6]. By adopting simultaneous DSIS, the ratings are typically based on relative differences subjects identify between the reference and the degraded content [6]. Thus, the comparison between ratings obtained from the two experiments, reflects the comparison between the relative levels of visual distortions as perceived under the two experimental set-ups.

In both experiments, the position of the reference was known to the subjects. To reduce the contextual effects, the side of the reference was selected randomly and remained fixed across the entire session completed by every subject. The order of stimuli was randomized per session, and consecutive display of the same content was intentionally avoided. No time limitations were imposed to the subjects.

In every session 25 stimuli were assessed, as 5 contents and 4 degradation values were used along with a hidden reference for sanity check. A total of 28 naïve subjects (17 males and 11 females) participated in Experiment A, with every session involving 20 observers. Their age was ranging from 20 to 37 years old (average 27.9). A total of 24 naïve subjects (14 males and 10 females) participated in Experiment B, with every session involving 21 observers. Their age was ranging from 25 to 32 years of age (average 27.7).

4. STATISTICAL ANALYSIS

In this section we describe the statistics, based on which the scores from the two experimental settings are compared.

4.1. Subjective quality metrics

Initially, an outlier detection algorithm based on the ITU-R Recommendation BT.500-13 [1] was applied on all subjective scores for every session, per experiment. Specifically, in Experiment A, no test subject was identified as an outlier for Gaussian noise (i.e., 20/20 scores), whereas 1 outlier was detected for Octree-pruning (i.e., 19/20 scores). In Experiment B, 3 and 1 outliers were found for Gaussian noise (i.e., 18/21) and Octree-pruning (i.e., 20/21), respectively. After discarding outliers, the mean opinion scores (MOS) and the 95% confidence intervals (CIs), assuming a Student's t-distribution, were computed for each test content.

4.2. Comparison between testbeds

Since each type of degradation produces substantially different artifacts, we compare the ratings obtained from Experiment A against the ratings from Experiment B after introducing Gaussian noise and we repeat identical analysis for the Octree-pruning results, separately.

Based on the Recommendation ITU-T P.1401 [9], several fittings were applied to the MOS values obtained from the two different testbeds. The scores collected from one test were considered as ground truth and a fitting function was applied to the scores obtained from the other test, before estimating

the performance indexes. In particular, let us assume Experiment A as the ground truth, with the MOS of the distorted content i being denoted as MOS_A^i. The MOS of the same content from Experiment B is indicated as MOS_B^i. A predicted MOS, indicated as MOS_P, is estimated by applying a fitting function to each pair $[\text{MOS}_A^i, \text{MOS}_B^i]$. In this study, no, linear, and cubic fittings were employed. Then, the Pearson linear correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), the root-mean-square error (RMSE) and the outlier ratio based on standard error (OR) were computed between MOS and MOS_P, for linearity, monotonicity, accuracy and consistency, respectively.

To decide whether the usage of a different testbed leads to statistically distinguishable results, the correct estimation (CE), under-estimation (UE) and over-estimation (OE) percentages were calculated, after a multiple comparison test at a 5% significance level. Let us assume that the scores obtained from the Experiment A are the ground truth. For every distorted content, the true difference MOS_B^i - MOS_A^i between the average ratings from every experiment is estimated with a 95% CI. If the CI contains 0, correct estimation is observed, which indicates that the visual quality of content i is statistically equivalent using both testbeds. If 0 is above, or below the CI, we conclude that the usage of the second testbed leads to under-estimation, or over-estimation of the visual quality of content i, respectively. The same computations are repeated for every content. After dividing the results with the total number of contents, we obtain the correct estimation, under-estimation, and over-estimation percentages.

To examine whether using a different testbed results in different conclusions for a pair of data points, the correct decision (CD), along with the false ranking (FR), false differentiation (FD) and false tie (FT) classification errors were computed, according to the Recommendation ITU-T J.149 [10]. In particular, let us assume that the subjective scores obtained from Experiment A are the ground truth. The true difference between the ratings of contents i and j, $\text{MOS}_A^i - \text{MOS}_A^j$, with a 95% CI is calculated. Depending on whether 0 lies below, in between, or above the CI, there are three possible categories: (a) i is better than j, (b) i is the same as j, and (c) i is worse than j. This three-way classification is performed for every i and j, with $i \neq j$. This procedure is repeated by computing $\text{MOS}_B^i - \text{MOS}_B^j$ with a 95% CI for every pair of distorted contents, as rated in Experiment B. When the results from this three-way classification for both experiments agree, a correct decision is observed. When the results using the first testbed (i.e., ground truth) say that i is better than j, or i is worse than j, and based on the results from the second testbed i is the same as j, a false tie occurs. This is the least offensive error. When the results from the first testbed say that i is the same as j, and based on the results from the second testbed i is better than j, or i is worse than j, a false differentiation occurs, which is a more offensive error. When the results of the first experiment say that i is better than j, or i is worse than j and the results from the second experiment suggest the opposite, the most offensive error, false ranking, occurs.

One-way and multi-way analysis of variance (ANOVA) were finally performed to determine whether the scores collected from the different experimental set-ups are statistically different, and to identify their influencing factors.

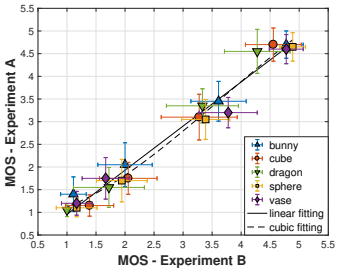
5. RESULTS AND DISCUSSION

In Figures 2 and 3, we illustrate scatter plots comparing the MOS for every test content as rated by subjects in both testbeds, setting the scores of Experiment A and Experiment B as ground truth, respectively. The horizontal and vertical bars associated with every point specify the CIs of the scores collected by the experimental setting indicated by the corresponding label. It is evident that, in both figures, the CIs are larger for Experiment B. In particular, for Gaussian noise and Octree-pruning, the CIs are on average 10.34% and 7.29% smaller for the desktop when compared to the HMD set-up. This behaviour is partially due to the higher level of interactivity offered by the HMD. As the subjects were free to interact with the test contents, not every angle was viewed from every subject and for every stimulus. This may have affected the consistency of the ratings even for the same observer. Moreover, the real environment scenery as a background is an additional factor that could have influenced the perception of the virtual objects. Finally, although no issues were reported by the subjects, the level of discomfort in Experiment B is admittedly higher and could lead to rating inconsistencies across an HMD session. To average such statistical uncertainties, a higher number of subjects is proposed to participate in highly interactive experimental settings.

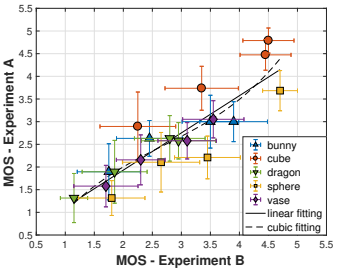
Concerning the results in the presence of Gaussian noise, in Figures 2a and 3a, the linear fitting function achieves an angle of $44.21°$ and $45.13°$, with an intercept of 0.957 and 1.14, respectively. This indicates that although highly correlated, the scores of Experiment A are consistently slightly lower. The strong correlation is verified by the high PCC and SROCC values of Tables 1 and 2. A correct estimation percentage of 100% implies that the MOS of the distorted contents, as rated in both testbeds, are statistically equivalent. The false ranking is 0%, while the false differentiation and false tie percentages are low (below 3.7%).

The results of a multi-way ANOVA performed on the scores for Gaussian noise are depicted in Table 3. It is shown that contents under different levels of distortions are rated as statistically significantly different with a p-value of 0, as can be also clearly noted in Figures 2a and 3a. Statistically significant difference can be also observed in the rating of different contents, as indicated by the p-value of 0.003. Finally, the two testbeds found to be statistically significantly different, albeit with a marginal p-value of 0.041.

To further investigate the impact of adopting a different visualization strategy to assess contents subject to Gaussian

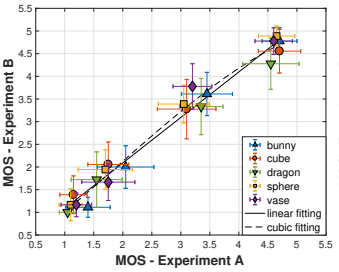


(a) Gaussian noise

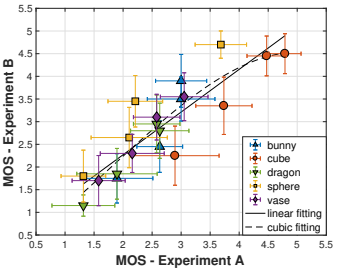


(b) Octree-pruning

Fig. 2: Scatter plots with Experiment A as ground truth.



(a) Gaussian noise



(b) Octree-pruning

Fig. 3: Scatter plots with Experiment B as ground truth.

Table 1: Performance indexes considering the scores collected in Experiment A as the ground truth.

Gaussian noise

	PCC	SROCC	RMSE	OR	CE	UE	OE	CD	FR	FD	FT
No fitting	0.9884	0.9590	0.2336	0.30	100%	0%	0%	94.21%	0%	3.68%	2.11%
Linear fitting	0.9884	0.9590	0.2102	0.25	100%	0%	0%	94.21%	0%	3.68%	2.11%
Cubic fitting	0.9904	0.9590	0.1921	0.15	100%	0%	0%	96.32%	0%	0.53%	3.16%

Octree-pruning

	PCC	SROCC	RMSE	OR	CE	UE	OE	CD	FR	FD	FT
No fitting	0.8790	0.8723	0.5435	0.60	90%	0%	10%	73.68%	0%	17.89%	8.42%
Linear fitting	0.8790	0.8723	0.4536	0.45	95%	0%	5%	72.11%	0%	15.26%	12.63%
Cubic fitting	0.8852	0.8723	0.4425	0.45	100%	0%	0%	76.32%	0%	14.21%	9.47%

Table 2: Performance indexes considering the scores collected in Experiment B as the ground truth.

Gaussian noise

	PCC	SROCC	RMSE	OR	CE	UE	OE	CD	FR	FD	FT
No fitting	0.9884	0.9590	0.2336	0.25	100%	0%	0%	94.21%	0%	2.11%	3.68%
Linear fitting	0.9884	0.9590	0.2136	0.15	100%	0%	0%	94.21%	0%	2.11%	3.68%
Cubic fitting	0.9904	0.9590	0.1946	0.15	100%	0%	0%	94.21%	0%	2.63%	3.16%

Octree-pruning

	PCC	SROCC	RMSE	OR	CE	UE	OE	CD	FR	FD	FT
No fitting	0.8790	0.8723	0.5435	0.45	90%	10%	0%	73.68%	0%	8.42%	17.89%
Linear fitting	0.8790	0.8723	0.4858	0.45	95%	5%	0%	73.68%	0%	8.42%	17.89%
Cubic fitting	0.8907	0.8723	0.4633	0.35	100%	0%	0%	74.74%	0%	6.84%	18.42%

Table 3: Multi-way ANOVA for Gaussian noise scores.

Source	SS	DF	MS	F	p
Testbed	1.60	1	1.60	4.20	0.041
Content	6.17	4	1.54	4.04	0.003
Distortion level	1389.72	3	463.24	1214.28	0
Error	286.50	751	0.38		
Total	1683.99	759			

noise, a multi-way ANOVA grouped per level of distortion revealed that the two testbeds are statistically different at a 5% significance level only for $\sigma = 0.002$ ($p = 0.0198$), while for the other σ values, no statistical differences were found ($p > 0.05$). Thus, it can be concluded that for the same distortion level, all contents are rated similarly, which is confirmed by the MOS indicated in Figures 2a and 3a. Finally, a multi-way ANOVA grouped per content, shows that *sphere* is the only point cloud for which the two testbeds are considered as statistically distinguishable ($p = 0.0156$). Hence, the statistical difference between the testbeds seems to be strictly linked to content *sphere* and to target level of distortion $\sigma = 0.002$.

Regarding the results after Octree-pruning, in Figures 2b and 3b the linear fitting function achieves an angle of $39.38°$ and $43.27°$ with an intercept of 1.23 and 1.63, respectively. Based on the performance indexes of Tables 1 and 2, the selection of a different testbed may lead to different conclusions regarding the visual quality of compression-like artifacts. In particular, the PCC and SROCC values decrease, while the RMSE and OR coefficients remarkably increase with respect to the Gaussian noise case. A correct estimation below 100% indicates that there is a percentage of distorted contents for which the MOS values are statistically distinguishable, with the subjects over-estimating the visual quality using the HMD set-up. The false ranking remains at 0%, but high percentages of false differentiation in Table 1 and false tie in Table 2 suggest that the desktop may not differentiate two stimuli, as opposed to the HMD setting.

The differences in rating trends are confirmed by the multi-way ANOVA results of Table 4, which prove that every

influencing factor is statistically significantly different, with $p < 0.001$. A multi-way ANOVA performed on scores clustered per level of distortion shows that the two testbeds are statistically different with $p = 0$ at a 5% significance level for $\rho = 70\%$ and $\rho = 90\%$. This implies that the visual quality of the contents that are most severely impacted by Octree-based compression, may be reasonably undifferentiated by the usage of a different experimental set-up, as opposed to contents with higher visual quality. When the scores are clustered per content, the two testbeds found to be statistically different at a 5% significance level for *bunny* ($p = 0.043$), *cube* ($p = 0.0107$), *sphere* ($p = 0$) and *vase* ($p = 0.008$). These results verify the scoring trends in Figures 2b and 3b.

Table 4: Multi-way ANOVA for Octree-pruning scores.

Source	SS	DF	MS	F	p
Testbed	10.42	1	10.42	14.56	0.0001
Content	236.79	4	59.20	82.78	0
Distortion level	410.59	3	136.87	191.38	0
Error	551.37	771	0.72		
Total	1209.18	779			

Finally, a one-way ANOVA was performed on the stimuli, separately for Gaussian noise and Octree-pruning, to better understand differences between the two testbeds as the sole influencing factor. Results show that the set-ups are not statistically significantly different for Gaussian noise ($p = 0.3959$), whereas for Octree-pruning, they found to be different with statistical significance ($p = 0.0095$). Considering the high performance indexes depicted in Tables 1 and 2 along with the marginal p-value of 0.041 of the multi-way ANOVA, it seems that the perception of visual distortions that can be more naturally anticipated by human visual system, such as Gaussian noise, is not affected by the application of radically different visualization strategies. On the contrary, other types of degradations that involve point removal, structured displacement and elimination of high frequency components of the underlying model, are assessed differently with high statistical significance when employing different equipment, as expressed by the p-value of 0.0001 of the multi-way ANOVA. Another factor that could influence the subject ratings could be the smaller resolution of the phone screen, which may differently affect the level of fidelity of a content subject to a different degradation type.

6. CONCLUSION

In this paper we investigated the impact of adopting different visualization strategies of different degrees-of-freedom for quality assessment of point clouds. Our results expose that different rating trends are observed under the usage of different equipment as a function of the degradation type under assessment. In particular, although in the presence of Gaussian noise, scores obtained from the desktop and the HMD set-ups were found as statistically equivalent in a strict sense, in the case of Octree-pruning, the testbeds are statistically distinguishable. This study suggests that visual quality assessment of point clouds should be conducted using the target equipment for consumption, to ensure high prediction accuracy.

7. REFERENCES

[1] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union, Jan. 2012.

[2] J. Zhang, W. Huang, X. Zhu, and J. N. Hwang, "A subjective quality evaluation for 3D point cloud models," in *International Conference on Audio, Language and Image Processing*, Jul. 2014.

[3] R. Mekuria, K. Blom, and P. Cesar, "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, Apr. 2017.

[4] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "Subjective and objective quality evaluation of 3D point cloud denoising algorithms," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, Jul. 2017.

[5] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "Subjective and objective quality evaluation of compressed point clouds," in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2017.

[6] E. Alexiou, E. Upenik, and T. Ebrahimi, "On the performance of metrics to predict quality in point cloud representations," in *Proceedings of SPIE*, Aug. 2017, vol. 10396 of *Applications of Digital Image Processing XL.*

[7] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards subjective quality assessment of point cloud imaging in augmented reality," in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2017.

[8] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation*, May 2011.

[9] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, Jul. 2012.

[10] ITU-T J.149, "Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)," International Telecommunication Union, Mar. 2004.