Point cloud subjective evaluation methodology based on reconstructed surfaces

Evangelos Alexiou[a], Antonio M. G. Pinheiro[b], Carlos Duarte[c], Dragan Matković[d], Emil Dumić[d], Luis A. da Silva Cruz[c], Lovorka Gotal Dmitrović[d], Marco V. Bernardo[b] Manuela Pereira[b] and Touradj Ebrahimi[a]

[a]Multimedia Signal Processing Group, École Polytechnique Fédérale de Lausanne, Switzerland;
[b]Institute of Telecommunications, University of Beira Interior, Portugal;
[c]Department of Electrical and Computer Engineering, University of Coimbra and Institute of Telecommunications, Portugal;
[d]Department of Electrical Engineering, University North, Croatia

ABSTRACT

Point clouds have been gaining importance as a solution to the problem of efficient representation of 3D geometric and visual information. They are commonly represented by large amounts of data, and compression schemes are important for their manipulation transmission and storing. However, the selection of appropriate compression schemes requires effective quality evaluation. In this work a subjective quality evaluation of point clouds using a surface representation is analyzed. Using a set of point cloud data objects encoded with the popular octree pruning method with different qualities, a subjective evaluation was designed. The point cloud geometry was presented to observers in the form of a movie showing the 3D Poisson reconstructed surface without textural information with the point of view changing in time. Subjective evaluations were performed in three different laboratories. Scores obtained from each test were correlated and no statistical differences were observed. Scores were also correlated with previous subjective tests and a good correlation was obtained when compared with mesh rendering in 2D monitors. Moreover, the results were correlated with state of the art point cloud objective metrics revealing poor correlation. Likewise, the correlation with a subjective test using a different representation of the point cloud data also showed poor correlation. These results suggest the need for more reliable objective quality metrics and further studies on adequate point cloud data representations.

Keywords: Subjective Quality Assessment, Point Cloud, Quality Metrics

1. INTRODUCTION

The significant growth of interest in adopting 3D imaging modalities in modern information technologies and communication systems, indicate the need for advanced content representations. Among the alternatives, point clouds have been gaining a large interest from the technological market, as a result of the efficiency and simplicity they offer in capturing, storing and rendering of 3D objects. Considering the current availability of several existing solutions for acquisition and display, a wide range of applications and use cases can benefit from this type of data representation. In particular, a well-documented list has already been identified, and detailed in the JPEG Pleno Point Clouds – Use Cases and Requirements.[1] In this document, the following major classes have been defined: 1) Rendering of content for virtual, augmented and mixed reality technologies; 2) 3D content creation; 3) Medical applications; 4) Construction and manufacturing; 5) Consumer and retail; 6) Cultural heritage; 7) Remote Sensing, GIS; 8) Autonomous vehicles, drones; and 9) Surveillance.

Point clouds are typically represented by huge amounts of data. Thus, efficient and reliable compression solutions are essential. However, the identification of best approaches, or even their improvement, is limited by the lack of adequate solutions for evaluation of the visual quality of a point cloud content. The latter, is currently one of the major challenges and open problems in point cloud imaging. Independently of the type of

Further author information: (Send correspondence to Antonio M. G. Pinheiro.)
Antonio M. G. Pinheiro: E-mail: pinheiro@ubi.pt

data representation, the visual quality of a content is typically evaluated through either subjective or objective quality assessments. Objective quality assessment is performed through computer algorithms that are designed to estimate signal distortions. Subjective quality assessment involves the participation of subjects in experiments in which distorted objects are visualized and their quality is rated. Such experiments are expensive in terms of both cost and time. Thus, the objective quality metrics are frequently used instead, especially, for real-time quality assessment, which is typically part of advanced compression schemes. However, subjective evaluations are necessary to provide the ground truth, against which the objective scores are calibrated and benchmarked.

In this paper we report the results of point cloud subjective evaluation experiments, that took place in three different test laboratories. The point clouds were compressed by octree pruning[2] using the Point Cloud Library (PCL) v1.8.0.[3] The processed point cloud stimuli were then presented and assessed by the subjects after 3D rendering using Screened Poisson surface reconstruction,[4] and displayed as a 3D sequence showing the object from changing point of views. Three different 3D monitors were used according to each testing laboratory availability. The dataset proposed in a previous work[5] was employed to facilitate a comparison between previous work, on our aim to identify statistical differences that may occur by different point cloud rendering and representation approaches.

2. RELATED WORK

Recently, a significant amount of work has been reported in the literature for subjective and objective quality assessment of point clouds. In particular, several objective quality metrics have been proposed and their correlation with subjective quality scores has already been investigated. However, in most such studies, point clouds were displayed as sets of points without applying any surface reconstruction algorithm before the final rendering; notice that the latter reflects a rather common way to consume 3D contents nowadays. In particular, Zhang et al.[6] performed subjective assessment of raw point clouds degraded by geometry and color, after applying uniform noise. However, this type of degradation is not realistic neither for color nor for geometry artifacts. Mekuria et al.[7] proposed a generic and real-time time-varying point cloud codec for 3D immersive video. The subjective quality of the codec performance was evaluated in a realistic 3D tele-immersive system in a virtual 3D room scenario where users are represented and interact as 3D avatars (synthetic content) and/or 3D PC (naturalistic content). Different aspects of the quality were tested, including overall quality of the 3D human rendition, the quality of the colors, the perceived realism of the point cloud, the motion quality, the near/far quality, and the level of immersiveness with the observer comparing the avatar and the point cloud representations. Mekuria et al.[8] provides bitrate results and objective scores for their proposed point cloud codec,[7] which was assessed in the framework of the recent activities of the MPEG standardization committee. The selected data set consists of colored objects and different encoding strategies were considered, providing bitrate results and objective scores. For the subjective evaluations, a passive approach was adopted by obtaining a video from a point cloud renderer using a predefined path for the virtual camera. Raw point clouds were visualized, with each point being displayed as a cube of a fixed size. Although the point size was allowed to be configured, a fixed value was set per content. However, in these studies, no correlation between the subjective and objective scores is reported.

In,[9] point cloud denoising algorithms were subjectively evaluated and the test contents were visualized after applying the Screened Poisson surface reconstruction.[4] A passive assessment was adopted and 2D video sequences were formed, after capturing the resulting mesh objects from different viewpoints by vertical and horizontal rotation. However, the impact of visualizing reconstructed meshes instead of raw point clouds was not investigated. In,[10] subjective quality assessment of colored point clouds was conducted, subject to simple octree and graph-based encoding algorithms. To render the point cloud data, primitive cubes were employed, whose size was adjusted based on the local neighborhood. The resulting test contents were captured from different viewing angles with the virtual camera following a spiral path. The subjects visualized animated 2D video sequences to provide their scores. In[2] and,[11] an interactive approach to subjectively assess geometry-only point clouds in a typical desktop setup was proposed, using the Double-Stimulus Impairment Scale (DSIS) and the Absolute Category Rating (ACR) evaluation methodologies, respectively. In the latter study, the correlation between these two tests was also investigated. The contents under evaluation were degraded after applying Gaussian noise, and octree-based compression. The test contents were presented to the subjects as sets of points. In,[12] a subjective methodology for point clouds using head mounted displays in an augmented reality scenario was

proposed. The subjects visualized raw point clouds and their interaction with the 3D models was performed by physical movements in the virtual reality world environment. More recently,[5] a subjective evaluation in five different independent laboratories with visualization of surface reconstructed point clouds using the Screened Poisson surface reconstruction algorithm.[4] The point clouds were shown to subjects as a video sequence showing an horizontal rotation followed by a vertical rotation around the rendered surface. Several conclusions were reached: the subjective results reveal poor correlation with the most well-known metrics, visualization using a point cloud representation or after surface reconstruction leads to poorly correlated results, and finally, the subjective results were strongly correlated between different laboratories, although different visualization equipment was used. In the current work, we are extending this work[5] by using 3D visualization.

3. SUBJECTIVE ASSESSMENT

In this section the preparation of test contents is described, followed by the design of the subjective evaluations.

Table 1. Parameters used for octree pruning.

	LoD	Number of points	Actual percentage	Target percentage
bunny	-	35947	100.00%	100%
	0.007	32957	91.68%	90%
	0.010	25209	70.13%	70%
	0,012	17763	49.41%	50%
	0.016	10870	30.24%	30%
cube	-	30246	100.00%	100%
	0.015	27541	91.06%	90%
	0.017	20888	69.06%	70%
	0.020	15002	49.60%	50%
	0.025	9602	31.75%	30%
dragon	-	22998	100.00%	100%
	0.008	20847	90.65%	90%
	0.010	16487	71.69%	70%
	0.013	11539	50.17%	50%
	0.017	7026	30.55%	30%
egyptian_mask	-	31601	100.00%	100%
	0.008	28393	89.85%	90%
	0.010	22061	69.81%	70%
	0.013	15790	49.97%	50%
	0.017	9466	29.96%	30%
sphere	-	30135	100.00%	100%
	0.004	27298	90.59%	90%
	0.011	21100	70.02%	70%
	0.015	15168	50.33%	50%
	0.020	8977	29.79%	30%
vase	-	36022	100.00%	100%
	0.007	32454	90.10%	90%
	0.009	25217	70.00%	70%
	0.011	17963	49.87%	50%
	0.015	10693	29.69%	30%
torus (reference)	-	31250	100.00%	100%
	0.005	30566	97.81%	98%
	0.007	27968	89.50%	90%
	0.010	21901	70.08%	70%
	0.012	15715	50.29%	50%
	0.017	9539	30.53%	30%

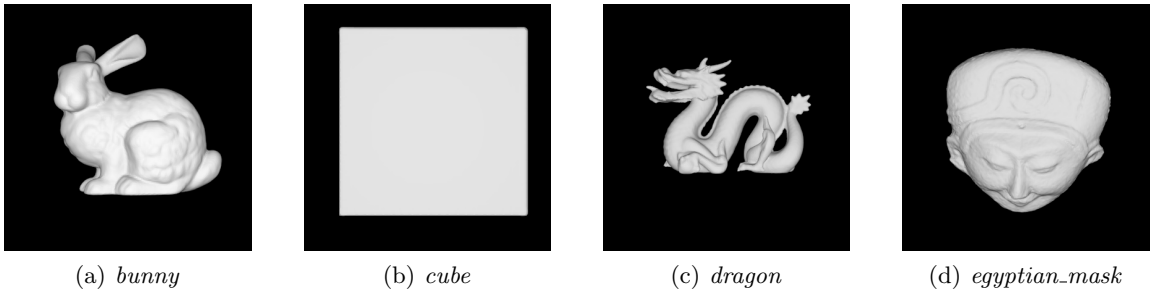
3.1 Content Preparation

In this experiment, a dataset of 7 geometry-only point clouds was used. In particular, *bunny* and *dragon* were selected from the Stanford 3D Scanning repository*. *Egyptian_mask* is a content used in the recent activities of the MPEG standardization committee,[13] and *vase* is an object captured by Intel RealSense R200 in.[2] *Cube* and *sphere* were synthesized using corresponding mathematical formulas, while *torus* was artificially produced in MeshLab†. To ensure that the number of points of every model is in the same order of magnitude, corresponding releases (i.e., *dragon_vrip_res3*) were selected, or sub-sampled versions (i.e., *vase* and *egyptian_mask*) were generated without modifying the original coordinates, thus, maintaining the original geometric structure of the test contents.

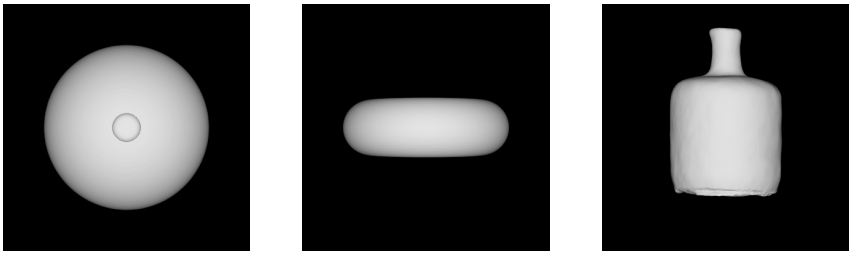
*http://graphics.stanford.edu/data/3Dscanrep/

†http://www.meshlab.net/

The original PCs were compressed by octree pruning, as described in,[11] PCL v1.8.0. To briefly describe this type of degradation, a content is enclosed in an octee structure and by modifying the size of the leaf nodes, which is referred as Level of Details (LoD), the geometric resolution is correspondingly adjusted. For instance, after increasing the LoD, the number of points of the output model naturally decreases. Considering that the octree is the basis for the majority of point cloud compression schemes, this is a simplified approach to produce artifacts after octree-based encoding. To account for a wide range of visible distortions, the target percentages (p) of remaining points after octree pruning were selected as: 90%, 70%, 50% and 30%, allowing a deviation of $\pm$2%. For *torus*, an additional version with 98% of points was also prepared to be used in the training. The number of points for every reference and distorted content, along with the LoD values that were used, can be found in Table 1. More details about the point cloud dataset can be found in.[5]

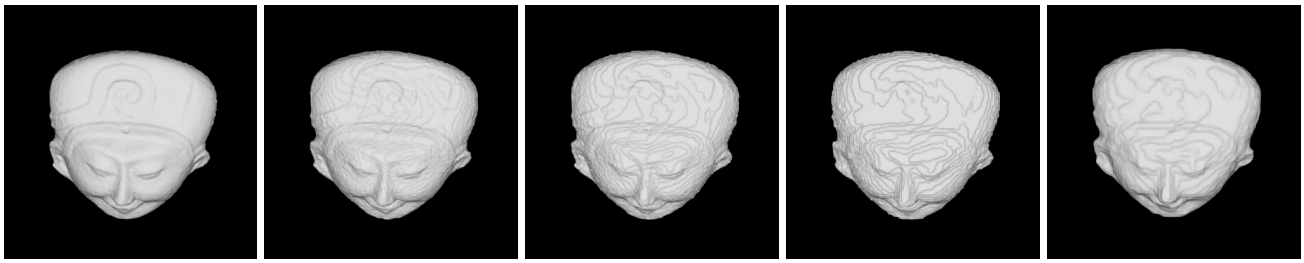


(a) *bunny* (b) *cube* (c) *dragon* (d) *egyptian_mask*



(e) *sphere* (f) *torus* (g) *vase*

Figure 1. Frontal view of each reference mesh.



(a) 100% (b) 90% (c) 70% (d) 50% (e) 30%

Figure 2. Frontal view of *egyptian_mask* for every target percentage.

The original raw point clouds were initially scaled to fit in a bounding box of size 1 and translated at the origin (0, 0, 0). Then, the distorted versions were produced following the aforementioned procedure. The Screened Poisson surface reconstruction algorithm[4] was selected to be used as a pre-rendering approach in order to generate the objects. The CloudCompare[‡] implementation was employed setting an octree depth of 8 and default parameters. To apply this algorithm, normal vectors should exist along with the coordinates of every content. As normal vectors were absent, they were estimated using CloudCompare with default settings, i.e., the radius to identify nearest neighbors was selected automatically and a plane was used as the local surface model.

[‡]https://www.danielgm.net/cc/

Then, on the same tool, the normals were oriented using a Minimum Spanning Tree of 6 nearest neighbors.

Table 2. Specific camera properties in Blender for different render types

	Stereoscopic	**Nvidia Stereoscopic**	**Texture-plus-depth**
Focal length	60	60	60
Aspect ratio	X: 2, Y: 1	X: 1, Y: 1	X: 1, Y: 1
Resolution per left/right view (pixels)	480x1080	960x1080	960x1080
Overall frame size (pixels)	1920x1080	2 x 1920x1080	3840x1080
Depth settings, Map Value box	-	-	offset -1.75, size 0.35, min 0, max 255
Convergence plain distance	1.95 m	1.95 m	-
Interocular distance, smaller parallax	-	0.02 m	-
Interocular distance, bigger parallax	0.035 m	-	-

In order to obtain 3D video sequences for further subjective assessment, Blender[§] was used with different parameters, to produce stereoscopic or texture-plus-depth based rendering. Every object was centered (set origin - geometry to origin). The camera and illumination source (lamp) were set at the position $(x, y, z) = (0, 0, 4)$. The properties of the camera setup can be found in Table 2. The lamp properties were: point type source; only diffuse shading; energy 1.7; inverse square falloff; falloff distance equal to 30. For the generation of still images to form the 3D video sequences, we used an empty object attached to a point cloud (empty is child object and point cloud is parent object), while the camera and the lamp were attached to an empty object (camera and lamp are child objects and empty is parent object). The empty object was rotated with linear rotation speed around Y-axis of 1° per frame (360 frames for a total rotation), and afterwards around X-axis with -1° per frame (360 frames for a total rotation), producing 720 frames overall. The reference meshes of every selected content are shown in Figure 1, whilst in Figure 2 the *egyptian_mask* is presented for every degradation level. Different setup for camera produces different final 3D videoframe, Figure 3. Since each participating laboratory had different types of 3D displays, contents were generated for the three corresponding display types: texture-plus-depth, stereoscopic and Nvidia stereoscopic setups. Furthermore, interocular distance parameter or disparity between left and right view was set to 2 different values, giving smaller or bigger parallax, as per Table 2. The methodology and the display types and specifications that were used in every test laboratory are described in Section 3.2. Finally, the sequences of 720 frames per object were encoded at 30 fps, which gives 24 second-long video sequences. Finally, the 3D video sequences were encoded with H.264/AVC lossless compression, in an .mp4 container using FFmpeg tool.

3.2 Evaluation Methodology

The subjective experiments were conducted in three laboratories: University of Beira Interior (UBI), Covilhã, Portugal; University of Coimbra (UC), Coimbra, Portugal; and University North (UNIN), Varaždin, Croatia. The conditions of every test environment were adjusted to follow the ITU-R Recommendation BT.500-13.[14] The equipment used in each laboratory is described in Table 3. A passive subjective methodology was followed, with the subjects visualizing the generated 3D video sequences in a customized video player, and providing their scores using a customized interface, either during or after finishing observing each individual test stimulus.

The DSIS simultaneous test method was adopted with a 5-level impairment scale (1 - *very annoying*, 2 - *annoying*, 3 - *slightly annoying*, 4 - *perceptible, but not annoying*, 5 - *imperceptible*), including a hidden reference for sanity check. Thus, both the reference and the degraded stimuli were simultaneously shown to the observer side-by-side, and every subject rated the visual quality of the distorted with respect to the reference stimulus, which were clearly annotated. To avoid biases, in half of the individual evaluations, the reference was placed on the right and the degraded content on the left side of the screen, and vice-versa for the remaining half of the evaluations.

The evaluation protocol allowed a free viewing (FV) scenario. That is, after the initial position, every subject was free to move closer or further from the screen in a pre-determined range. This approach was established to allow for compensation of perspective effects, as different objects could be perceived as having different volume. For instance, from a fixed distance between the observer and the screen, the *dragon* appears to be smaller than

[§]https://www.blender.org/

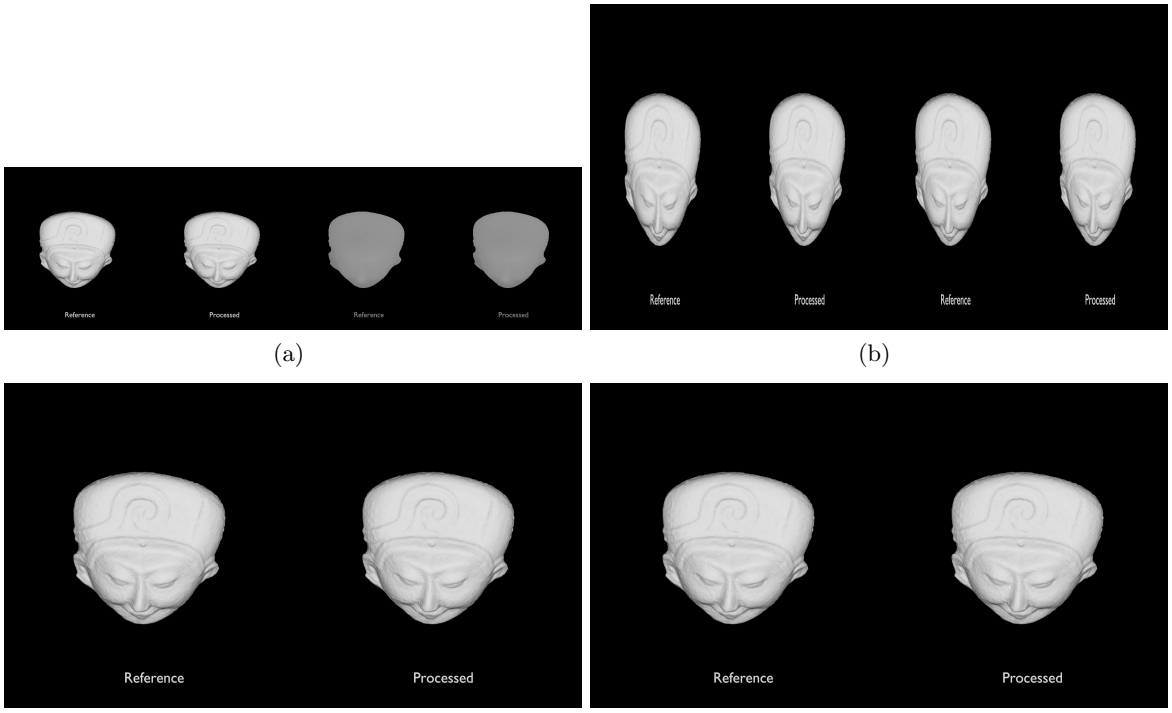


Figure 3. Different setup for *egyptian_mask* object, frame 1, hidden reference is right-processed object: a) texture-plus-depth, b) side-by-side, stereoscopic frame with combined left and right view, c) Nvidia stereoscopic, left view, d) Nvidia stereoscopic, right view

Table 3. Equipment and subjects information per laboratory.

		UBI	**UC**	**UNIN**
Equipment	Monitor	LG 47LA860V	Asus VG278HR	Dimenco DM504MAS
	Inches	47”	27”	50”
	Resolution	1920x1080	1920x1080	3840x2160
	View Distance	1.2 m (FV)	0.8 m (FV ±20 cm)	2.15 m (FV ±10 cm)
	Used player	MPV	Nvidia 3D Vision Video Player	Dimenco 3D Player
	Render type	Stereoscopic (Passive)	Nvidia Stereoscopic (active)	Texture-plus-depth (auto-stereoscopic)
Subject Info	Males	14	10	9
	Females	6	10	14
	Overall	20	20	23
	Year span	21-40	20-54	19-57
	Average age	28.5	27.8	24
	Outliers	0	1	1

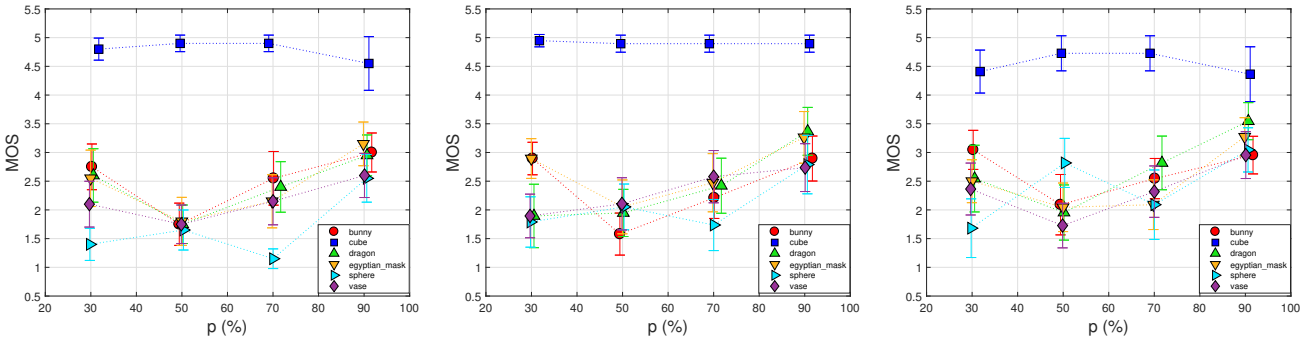
the *sphere*, due to the different ratio between height and length. In every test laboratory, the initial position and the viewing range were different, as different equipment was used. The exact values are reported in Table 3. It should be also noted that in the texture-plus-depth evaluation methodology, observers had to be static, due to the used auto-stereoscopic 3D display which has defined distance from the screen, 2.15m in tested case, as well as 28 different views that can generate unnatural 3D view if the observers are wrongly placed.

At the beginning of each individual evaluation, a training session took place in order to familiarize the subjects with the visual representations under assessment. The *torus* was selected for this purpose and, hence, it was excluded from the actual testing. The training was performed using 3 animated video sequences that represented 3 different levels of degradation in order to indicatively illustrate the range of visible distortions.

An overall of 30 scores were obtained per evaluation session, considering that each subject assessed 6 test contents degraded in 4 distinct levels along with the hidden references. An outlier detection algorithm based on ITU-R Recommendation BT.500-13[14] was applied to the collected scores, and the ratings of the identified outliers were discarded. Then, the mean opinion scores (MOS) and the 95% Confidence Intervals (CIs), assuming a Student's t-distribution were computed. In Table 3, equipment details, observer information and the number of outliers per test laboratory are reported.

4. RESULTS

In Figure 4, the MOS of the test contents are presented as a function of the level of degradation. The caption of each sub-figure indicates the test laboratory where the subjective scores were obtained.



(a) UBI (b) UC (c) UNIN

Figure 4. Subjective scores against degradation values per laboratory.

As can be observed, the subjective ratings for the *cube* remain high, independently of the distortion level. For the majority of the rest of the contents, the MOS is increasing as the degradation level is decreasing, excluding the lowest value ($p = 30\%$) where we observe stable or even higher ratings. This comes as a result of the working principle of the selected surface reconstruction algorithm. In particular, due to the significant reduction of points of the distorted contents, smoother functions were automatically employed by the algorithm to reconstruct the corresponding meshes. Note, also, that octree pruning leads to structural point removal without considering the underlying characteristics of the objects. By applying smoother fitting functions, although there is fidelity loss, the visual degradations that correspond to the geometric artifacts introduced by the octree structure are decreasing, which is preferred by the subjects. An example can be seen in Figure 2, where the impairment of the content of Figure 2 (e) can be considered as visually less annoying with respect to the content of Figure 2 (d). Finally, for *sphere*, we observe an increase in MOS values from $p = 30\%$ to $p = 50\%$ and a decrease at $p = 70\%$. This is because surface smoothing was observed for $p \leq 50\%$ for this content.

4.1 Correlation between Subjective and Objective Scores

In this section we present the results of benchmarking of the state-of-the-art objective quality metrics for point clouds. In particular, the current objective quality metrics can be distinguished into two classes: (i) point-to-point (p2point), and (ii) point-to-plane (p2plane). The first class is based on the geometric distance between a reference and a distorted point, while the second, is based on the error obtained after projecting a reference point to the normal vector that corresponds to a distorted point. For both objective metrics, the algorithm to compute the total error of a point cloud is identical. In particular, for every point of the content under assessment, a nearest neighbor is identified in the reference point cloud and an individual error is computed, based on the selected metric. Then, the Mean-Squared-Error (MSE), or the Hausdorff distance are typically applied to the individual errors to get a total degradation value. This procedure is commonly repeated by setting both the distorted and the original contents as a reference. The maximum error value is kept, which is referred to as the symmetric error. Finally, the Peak-to-Signal Noise Ratio (PSNR) values were also computed as defined in,[15] using a factor of 1 in the numerator. The PSNR is defined as the maximum distance of the nearest neighbors of the reference content divided by the squared error value (MSE, or squared Hausdorff). Considering every

combination of (a) objective quality metrics, (b) approaches to compute the total geometric error, and (c) PSNR values, we have a total of 8 objective metrics that were used in this study. The objective scores were computed using the software v.0.11.[15]

Similarly to,[5] the objective scores are computed on the point clouds before and after the surface reconstruction, given that a polygonal mesh is a list of vertices with associated faces. In addition, due to the outlier behavior of *cube* which can be seen in Figure 4, we present the performance of metrics after including and excluding the scores of this content. Finally, to use the p2plane metric, the existence of normal vectors is required. For point clouds before surface reconstruction, the estimated normal vectors that were used to reconstruct the corresponding meshes were employed, as described in Section 3.1, while for point clouds after surface reconstruction, the normal vectors that are naturally associated to the produced mesh, were used.

For the benchmarking of objective quality assessment tools, the subjective ratings are commonly set as the ground truth. Let us define the result of execution of a particular objective metric as a Point cloud Quality Rating (PQR). A predicted MOS for a test content is obtained after a fitting function on every [PQR, MOS] pair. In this study, the monotonic cubic function was selected for regression analysis. Then, following the Recommendation ITU-T P.1401,[16] to assess the performance of every objective metric, the Pearson Correlation Coefficient (PCC), the Spearman Rank Order Correlation Coefficient (SROCC), the Root-Mean Squared Error (RMSE) and the Outlier Ratio (OR) are computed between the subjective and the predicted MOS, to account for linearity, monotonicity, accuracy and consistency, accordingly.

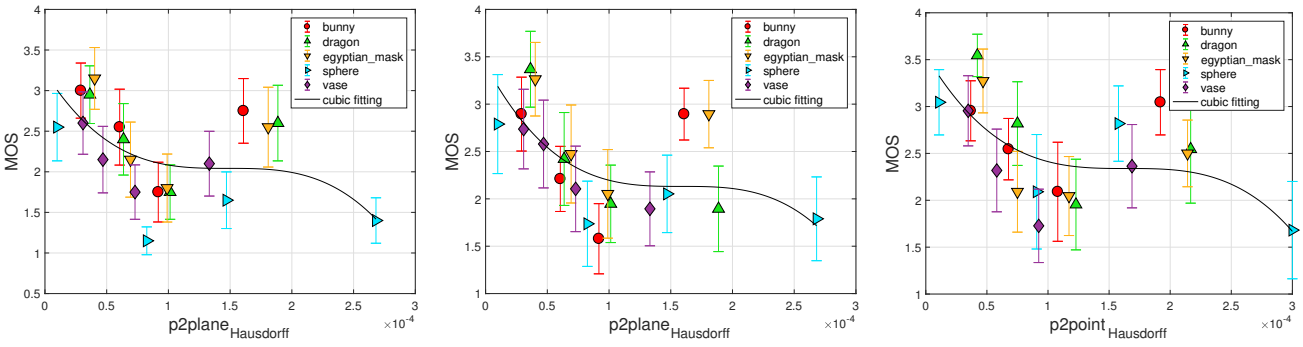
Table 4. Benchmarking results for the best-performing objective metric per test laboratory and tested case.

Before surface reconstruction

		Objective metric	**PCC**	**SROCC**	**RMSE**	**OR**
UBI	With *cube*	p2plane-MSE	0.655	0.132	0.829	0.625
	Without *cube*	p2plane-Hausdorff	0.592	0.463	0.450	0.500
UC	With *cube*	p2plane-MSE	0.644	0.071	0.822	0.625
	Without *cube*	p2plane-Hausdorff	0.635	0.545	0.407	0.300
UNIN	With *cube*	p2plane-MSE	0.599	0.090	0.731	0.583
	Without *cube*	p2point-Hausdorff	0.644	0.498	0.401	0.300

After surface reconstruction

		Objective metric	**PCC**	**SROCC**	**RMSE**	**OR**
UBI	With *cube*	p2plane-MSE	0.735	0.078	0.744	0.708
	Without *cube*	p2point-Hausdorff	0.556	0.541	0.464	0.500
UC	With *cube*	p2plane-MSE	0.710	0.046	0.758	0.583
	Without *cube*	p2plane-MSE	0.492	0.622	0.459	0.350
UNIN	With *cube*	p2point-MSE	0.673	0.231	0.675	0.542
	Without *cube*	p2plane-MSE	0.326	0.515	0.496	0.400



(a) UBI (b) UC (c) UNIN

Figure 5. Cubic fitting to benchmark the best-performing objective metric per test laboratory, excluding *cube*.

In Table 4, the objective quality metrics with better performance are presented per test laboratory and tested case. As can be observed, the performance indexes are generally low and the correlation is improved when the ratings for *cube* are excluded from the computations. Furthermore, by computing the objective scores before

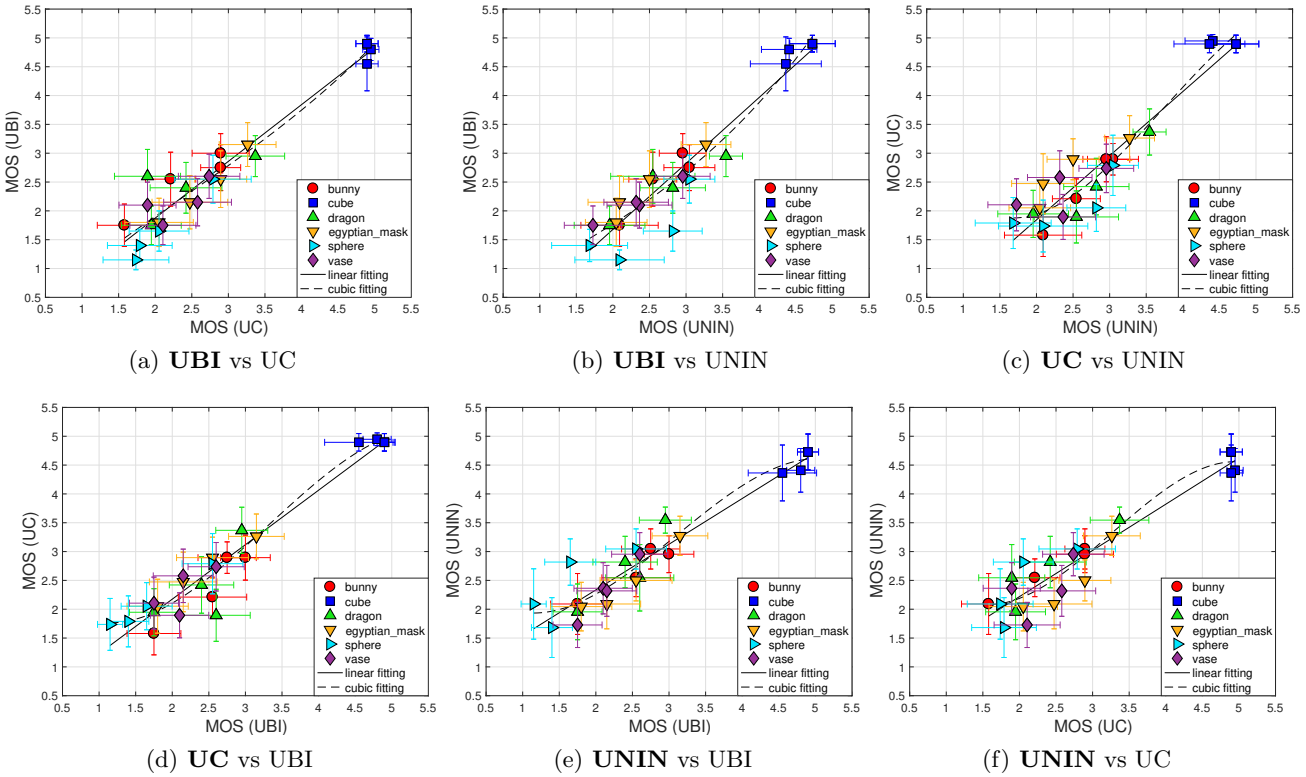
surface reconstruction leads to better correlation results. In Figure 5, we present the best-performing objective metric per laboratory; that is obtained before surface reconstruction and excluding *cube*.

4.2 Comparison between Subjective Scores from different Labs

In this section we study and analyze inter-laboratory correlation results. In particular, following the Recommendation ITU-T P.1401,[16] several fitting functions were applied to compute the PCC, SROCC, RMSE and OR coefficients. Specifically, no fitting, linear fitting and monotonic cubic fitting functions were applied as regression models. Furthermore, to determine whether statistically equivalent subjective scores are obtained in the three labs, the Correct Estimation (CE), Under Estimation (UE), and Over Estimation (OE) percentages were calculated after a multiple comparison test at a 5% significance level. Finally, to identify if a pair of data points can lead to different conclusions, the False Ranking (FR), False Differentiation (FD), False Tie (FT) and Correct Decision (CD) percentages were computed, based on the Recommendation ITU-T J.149.[17] Notice, that, since the scores of a particular laboratory cannot be used as the ground truth, for every pair of universities (X, Y), we set the subjective scores of university X as the ground truth benchmarking the scores of university Y, and vice versa.

Table 5. Performance indexes to compare subjective scores between different labs (Bold text represents the ground truth).

	PCC	SROCC	RMSE	OR	CE	OE	UE	CD	FR	FD	FT
UBI *vs* UC	0.965	0.888	0.287	0.125	100%	0%	0%	89.493%	0%	2.536%	7.971%
UC *vs* UBI	0.971	0.888	0.256	0.125	95.833%	0%	4.167%	89.493%	0%	2.174%	8.333%
UBI *vs* UNIN	0.960	0.893	0.308	0.167	100%	0%	0%	89.493%	0%	2.536%	7.971%
UNIN *vs* UBI	0.958	0.893	0.263	0.083	100%	0%	0%	85.507%	0%	1.449%	13.043%
UC *vs* UNIN	0.963	0.843	0.292	0.292	100%	0%	0%	90.580%	0%	1.087%	8.333%
UNIN *vs* UC	0.948	0.843	0.292	0.125	100%	0%	0%	86.594%	0%	0.725%	12.681%



(a) **UBI** vs UC (b) **UBI** vs UNIN (c) **UC** vs UNIN

(d) **UC** vs UBI (e) **UNIN** vs UBI (f) **UNIN** vs UC

Figure 6. No, Linear and Cubic fitting, to evaluate the correlation between pairs of laboratories (Bold text represents the ground truth).

The performance indexes were very similar for every regression model we tested, but slightly better for cubic function. Thus, the latter was used to compute the coefficients reported in Table 5. Moreover, in Figure 6,

we present scatter plots along with every type of fitting, that indicate the correlation between every pair of universities. Note that for the same pair, the corresponding sub-figures are arranged vertically.

Based on our results, strong correlation is noted for every combination of different laboratories. In general, the PCC and SROCC coefficients are high, whilst the RMSE and OR values remain low, indicating high accuracy and consistency of the ratings, respectively. The False Ranking, which is the most severe type of error remains 0%, while the Correct Decision is above 85%, which indicates that for a big percentage of pairs of contents, the same conclusions can be drawn by two test labs. In UNIN we observe higher CIs with respect to the other two test laboratories; specifically, the CIs in UNIN are on average 12.30% and 8.63% larger than UBI and UC, respectively. This explains why the False Tie percentages are high when the UNIN scores are set as ground truth. Based on the performance indexes, there is a small percentage of contents that is under-estimated in UBI when compared to the corresponding scores in UC. However, the highest deviation in terms of rating behavior is observed between UNIN and UC. In particular, when the UC scores are set as ground truth, the linear fitting function achieves an angle of 48.05° with a Y-intercept of -0.38, while by setting the UNIN scores as ground truth, an angle of 38.71° and a Y-intercept of 0.62 is achieved. This, essentially, indicates that in UNIN lower visual quality objects are consistently rated higher, while higher visual quality objects are rated lower with respect to UC. It should be considered, though, that except of the different 3D setup that was adopted, in this pair of universities we have the most diverse display specifications. Specifically, in UNIN a 50" display with 4K resolution was used, whilst in UC a 27" monitor with FHD resolution was installed.

4.3 Comparison between Subjective Scores for Mesh visualization in 2D and 3D monitors

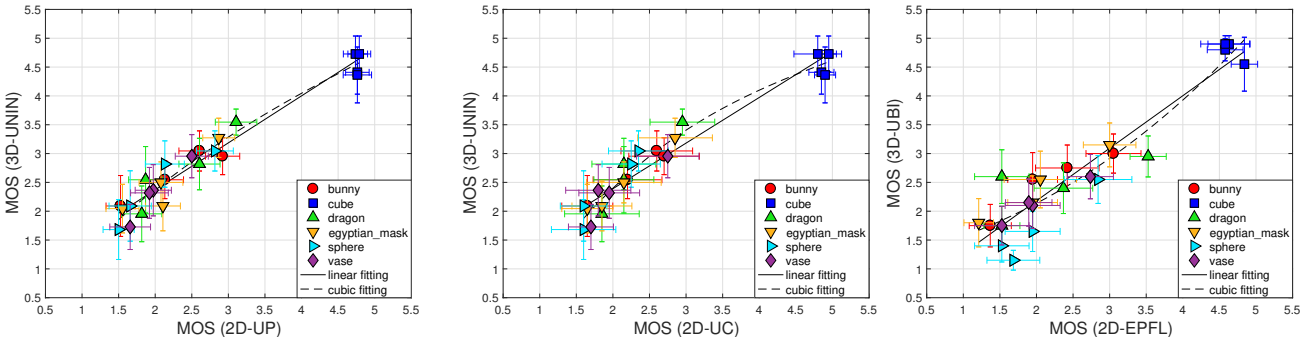
The subjective scores collected in this experiment were compared to ratings obtained from a previous experiment, which is described in.[5] Specifically, the same set of test contents was assessed using typical 2D displays in five test laboratories. The size of the displays used varied from 28" to 50", and the resolution from FHD to 4K. To compare our results, the performance indexes described in Section 4.2 are employed. In this study, we compute and present the performance indexes between every set of scores that was obtained for mesh visualization in 2D displays against every set of scores acquired using 3D displays, which leads to 15 possible combinations. Furthermore, for every pair that is formed, by alternating the set of scores that is considered as ground truth, we get 30 sets of correlation coefficients for the corresponding comparisons, which are reported in Table 6. In Figure 7, scatter plots between various pairs are indicatively presented, including every fitting function. Please note that in order to report the results, the following naming convention was used: *Rendering-University (laboratory)*

Based on our results, the correlation is relatively strong for every combination, which is confirmed by the high PCC and SROCC values, which are above 0.94 and 0.8, respectively. The RMSE and OR vary, but lie at low levels. The False Rating is 0%, while high percentages are observed for Correct Estimation and Correct Decision (i.e., above 95.8% and 79%, respectively). Another remark is that the False Tie percentages are consistently high when the UP is set as the ground truth. This is observed because the CIs in UP are very small due to the high number of participated subjects (i.e., 44).

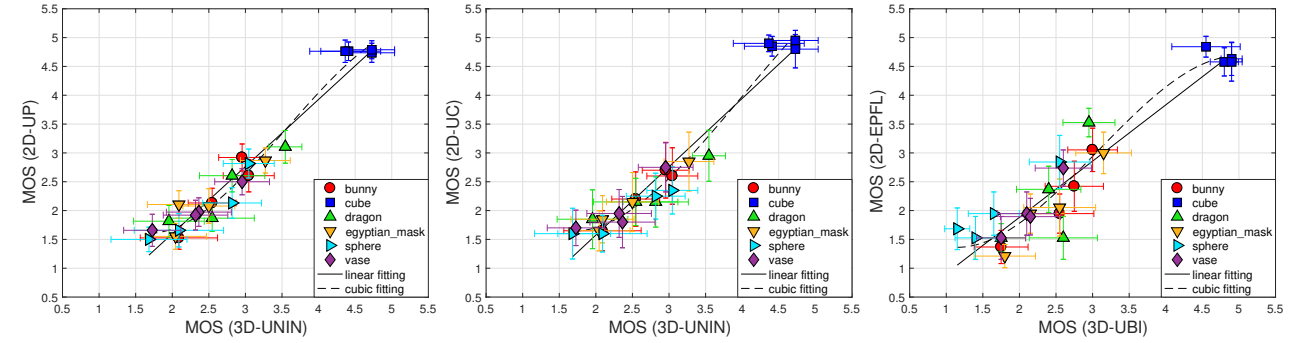
In Figure 7 (a) and (d), we compare the subjective scores obtained in UNIN using 3D rendering against the scores collected in UP using 2D rendering. When the scores of UNIN are set as the ground truth, the linear fitting function achieves a Y-intercept of 0.74 with an angle of 39.17°. By setting the UP scores as the ground truth, we have a Y-intercept of -0.73 with an angle of 49.36°, showing that the subjects in UP tended to rate lower and higher, the more and less severely degraded contents, respectively. Similar conclusions are draw from the Figures 7 (b) and (d). For both pairs, although different rating tendencies are observed, the performance indexes indicate that the discrimination of contents is consistent and highly correlated in different setups. In Figure 7 (c), by setting the EPFL scores as the ground truth, a Y-intercept of -0.07 and an angle of 44.28° is achieved, whilst in Figure 7 (e), by setting the UBI scores as the ground truth, the Y-intercept and the angle of the linear fitting function is 0.36 and 42.33°, respectively. Thus, no obvious tendencies are observed. In this case, the correlation coefficients show that in UBI there is a small percentage of Under Estimation of the visual quality of contents. Furthermore, the RMSE and OR values are among the highest between all combinations, revealing lower consistency and accuracy.

Table 6. Performance indexes to compare subjective scores for different mesh rendering approaches (Bold text represents the ground truth).

	PCC	SROCC	RMSE	OR	CE	OE	UE	CD	FR	FD	FT
2D-EPFL *vs* 3D-UBI	0.956	0.855	0.334	0.292	95.833%	0%	4.167%	84.420%	0%	4.348%	11.232%
3D-UBI *vs* 2D-EPFL	0.949	0.855	0.345	0.292	100%	0%	0%	81.884%	0%	6.522%	11.594%
2D-EPFL *vs* 3D-UC	0.971	0.904	0.270	0.167	100%	0%	0%	86.594 %	0%	0%	13.406%
3D-UC *vs* 2D-EPFL	0.974	0.904	0.244	0.167	100%	0%	0%	88.406 %	0%	4.348%	7.246%
2D-EPFL *vs* 3D-UNIN	0.978	0.925	0.238	0.250	95.833%	0%	4.167%	83.696 %	0%	0%	16.304%
3D-UNIN *vs* 2D-EPFL	0.966	0.925	0.236	0.042	100%	0%	0%	88.768 %	0%	2.536%	8.696%
2D-UBI *vs* 3D-UBI	0.981	0.915	0.199	0.042	100%	0%	0%	95.652 %	0%	1.087%	3.261%
3D-UBI *vs* 2D-UBI	0.971	0.915	0.262	0.250	100%	0%	0%	93.841 %	0%	2.536%	3.623%
2D-UBI *vs* 3D-UC	0.982	0.892	0.193	0.042	100%	0%	0%	96.377 %	0%	0.362%	3.261%
3D-UC *vs* 2D-UBI	0.968	0.892	0.268	0.167	100%	0%	0%	92.391 %	0%	0.725%	6.884%
2D-UBI *vs* 3D-UNIN	0.969	0.863	0.252	0.125	100%	0%	0%	97.101 %	0%	0.362%	2.536%
3D-UNIN *vs* 2D-UBI	0.941	0.863	0.309	0.292	100%	0%	0%	91.667 %	0%	2.174%	6.159%
2D-UC *vs* 3D-UBI	0.986	0.910	0.188	0.083	100%	0%	0%	91.667%	0%	1.449%	6.884%
3D-UBI *vs* 2D-UC	0.974	0.910	0.248	0.125	100%	0%	0%	90.580%	0%	4.348%	5.072%
2D-UC *vs* 3D-UC	0.983	0.896	0.205	0	100%	0%	0%	91.304%	0%	1.449%	7.246%
3D-UC *vs* 2D-UC	0.973	0.896	0.249	0.125	100%	0%	0%	92.029%	0%	1.812%	6.159%
2D-UC *vs* 3D-UNIN	0.987	0.954	0.181	0.083	100%	0%	0%	94.203%	0%	0.725%	5.072%
3D-UNIN *vs* 2D-UC	0.981	0.954	0.177	0	100%	0%	0%	93.478%	0%	2.536%	3.986%
2D-UNIN *vs* 3D-UBI	0.956	0.804	0.305	0.208	100%	0%	0%	87.681%	0%	1.812%	10.507%
3D-UBI *vs* 2D-UNIN	0.946	0.804	0.357	0.292	100%	0%	0%	90.217%	0%	3.623%	6.159%
2D-UNIN *vs* 3D-UC	0.970	0.874	0.250	0.125	100%	0%	0%	91.304%	0%	3.986%	4.710%
3D-UC *vs* 2D-UNIN	0.965	0.874	0.281	0.208	100%	0%	0%	90.942%	0%	3.623%	5.435%
2D-UNIN *vs* 3D-UNIN	0.973	0.896	0.237	0.125	100%	0%	0%	89.855%	0%	1.812%	8.333%
3D-UNIN *vs* 2D-UNIN	0.960	0.896	0.254	0.167	100%	0%	0%	88.768%	0%	3.261%	7.971%
2D-UP *vs* 3D-UBI	0.976	0.880	0.238	0.333	95.833%	0%	4.167%	80.797%	0%	1.449%	17.754%
3D-UBI *vs* 2D-UP	0.968	0.880	0.277	0.167	100%	0%	0%	83.696%	0%	13.768%	2.536%
2D-UP *vs* 3D-UC	0.978	0.903	0.227	0.250	100%	0%	0%	81.522%	0%	0.725%	17.754%
3D-UC *vs* 2D-UP	0.976	0.903	0.235	0.042	100%	0%	0%	85.507%	0%	11.232%	3.261%
2D-UP *vs* 3D-UNIN	0.985	0.950	0.188	0.292	95.833%	0%	4.167%	78.986%	0%	0%	21.014%
3D-UNIN *vs* 2D-UP	0.977	0.950	0.193	0	100%	0%	0%	88.406%	0%	9.420%	2.174%



(a) **3D-UNIN** vs 2D-UP (b) **3D-UNIN** vs 2D-UC (c) **3D-UBI** vs 2D-EPFL



(d) **2D-UP** vs 3D-UNIN (e) **2D-UC** vs 3D-UNIN (f) **2D-EPFL** vs 3D-UBI

Figure 7. No, Linear and Cubic fitting, to evaluate the correlation between different mesh rendering approaches (Bold text represents the ground truth).

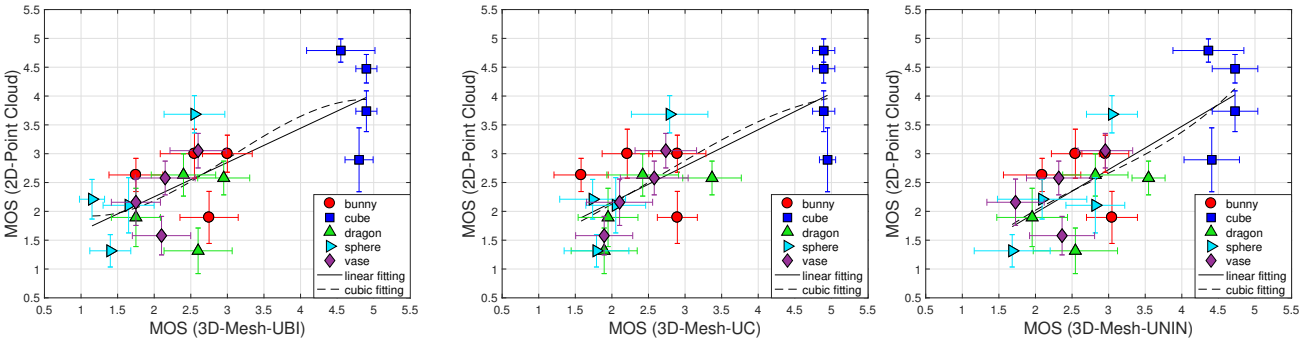
In general, though, based on our analysis we may conclude that subjective quality assessment of this visual modality does not strongly depend on the specifications of the display, as shown in Section 4.2, or the rendering methodology, as presented in this Section.

4.4 Comparison between Subjective Scores after PC and Mesh visualization

Finally, the subjective scores of this experiment were compared to ratings obtained in a previous experiment, where the visual quality of the same degraded point clouds (excluding *Egyptian_mask*) was assessed without enabling any intervening reconstruction algorithm before rendering.[11] The performance indexes described in Section 4.2 are used to compare subjective scores obtained from every test laboratory for mesh visualization in 3D monitors, and the subjective scores for point cloud visualization in 2D monitors. No fitting, linear and cubic fitting functions were tested, with the latter providing better fitting results that are used to report the performance indexes of Table 7.

Table 7. Performance indexes to compare subjective scores for different mesh rendering approaches (Bold text represents the ground truth).

	PCC	SROCC	RMSE	OR	CE	OE	UE	CD	FR	FD	FT
2D-PC *vs* 3D-Mesh-UBI	0.743	0.637	0.637	0.600	70 %	10%	20%	63.158%	0.526%	7.895%	28.421%
3D-Mesh-UBI *vs* 2D-PC	0.765	0.637	0.759	0.650	80%	10%	10%	66.842%	0%	8.947%	24.211%
2D-PC *vs* 3D-Mesh-UC	0.773	0.678	0.604	0.600	80%	10%	10%	60.526%	0%	11.579%	27.895%
3D-Mesh-UC *vs* 2D-PC	0.784	0.678	0.722	0.450	75%	10%	15%	70 %	0%	8.947%	21.053%
2D-PC *vs* 3D-Mesh-UNIN	0.760	0.680	0.619	0.750	85%	5%	10%	67.895%	0%	12.632%	19.474%
3D-Mesh-UNIN *vs* 2D-PC	0.773	0.680	0.609	0.500	95%	5%	0%	70.526%	0%	3.684%	25.789%



(a) **2D-PC** vs 3D-Mesh-UBI (b) **2D-PC** vs 3D-Mesh-UC (c) **2D-PC** vs 3D-Mesh-UNIN

(d) **3D-Mesh-UBI** vs 2D-PC (e) **3D-Mesh-UC** vs 2D-PC (f) **3D-Mesh-UNIN** vs 2D-PC

Figure 8. No, Linear and Cubic fitting, to evaluate the correlation between different mesh rendering approaches (Bold text represents the ground truth).

In Figure 8, we provide scatter plots indicating the correlation between the two experiments, including every fitting function. Based on the performance indexes, the correlation between these two tests is poor, which is in agreement with the results of our previous work.[5] The outcome of this analysis is that using a

surface reconstruction technique as a pre-rendering step to consume 3D objects leads to differently rated visible distortions in both 2D and 3D display setups, with respect to visualization of raw point cloud contents.

5. CONCLUSIONS

In this work was conducted a subjective evaluation of octree-based compression artifacts of PCs, rendered as mesh objects in 3D displays, a stereo with passive technology, a stereo with active technology and a auto-stereoscopic. The experiment was performed on three independent laboratories and results reveal high correlation among test labs, although different displays were used. The results using 3D displays reveal a very high correlation with previous results using 2D visualization of the same content. However, the results reveal a poor correlation with a subjective test where data was shown as a set of points, without any surface reconstruction. Hence, the surface reconstruction algorithm influences the perception of quality.

Finally, a comparison of the subjective scores of every laboratory with the state-of-the-art PC objective metrics shows that the visual quality cannot be sufficiently predicted for every type of content.

ACKNOWLEDGMENTS

This work has been conducted in the framework of the project PTDC/EEI-PRO/2849/ 2014 - POCI-01-0145-FEDER-016693 funded by the Portuguese FCT-Fundação para a Ciência e Tecnologia and co-funded by FEDER–PT2020 partnership agreement, and the project FAVIETOL and PCOMPQ of Instituto de Telecomunicações, FCT UID/EEA/50008/2013. The research was also supported by the Swiss National Foundation for Scientific Research (FN 200021_178854) project Advanced Visual Representation and Coding in Augmented and Virtual Reality.

REFERENCES

[1] WG1, "JPEG Pleno Point Clouds – Use Cases and Requirements." Stuart Perry (Editor), Doc. N80018, 80th JPEG Meeting, Berlin, Germany (July 2018).

[2] Alexiou, E. and Ebrahimi, T., "On subjective and objective quality evaluation of point cloud geometry," in [*2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*], 1–3 (May 2017).

[3] Rusu, R. B. and Cousins, S., "3D is here: Point Cloud Library (PCL)," in [*2011 IEEE International Conference on Robotics and Automation*], 1–4 (May 2011).

[4] Kazhdan, M. and Hoppe, H., "Screened Poisson Surface Reconstruction," *ACM Trans. Graph.* **32**, 29:1–29:13 (July 2013).

[5] Alexiou, E., Bernardo, M. V., da Silva Cruz, L. A., Dmitrovic, L. G., Duarte, R., Dumic, E., Ebrahimi, T., Matkovic, D., Pereira, M., Pinheiro, A., and Skodras, A., "Point cloud subjective evaluation methodology based on 2d rendering," in [*2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*], (May 2018).

[6] Zhang, J., Huang, W., Zhu, X., and Hwang, J. N., "A subjective quality evaluation for 3D point cloud models," in [*2014 International Conference on Audio, Language and Image Processing*], 827–831 (July 2014).

[7] Mekuria, R., Blom, K., and Cesar, P., "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video," *IEEE Transactions on Circuits and Systems for Video Technology* **27**, 828–842 (April 2017).

[8] Mekuria, R., Laserre, S., and Tulvan, C., "Performance assessment of point cloud compression," in [*2017 IEEE Visual Communications and Image Processing (VCIP)*], 1–4 (Dec. 2017).

[9] Javaheri, A., Brites, C., Pereira, F., and Ascenso, J., "Subjective and objective quality evaluation of 3D point cloud denoising algorithms," in [*2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*], 1–6 (July 2017).

[10] Javaheri, A., Brites, C., Pereira, F., and Ascenso, J., "Subjective and objective quality evaluation of compressed point clouds," in [*2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*], 1–6 (Oct. 2017).

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

[12] Alexiou, E., Upenik, E., and Ebrahimi, T., "Towards subjective quality assessment of point cloud imaging in augmented reality," in [*2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*], 1–6 (Oct. 2017).

[13] MPEG 3DG and Req., "Call for proposals for Point Cloud Compression V2." ISO/IEC MPEG2017/N16763 (April 2017).

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

[15] Tian, D., Ochimizu, H., Feng, C., Cohen, R., and Vetro, A., "Evaluation Metrics for Point Cloud Compression." ISO/IEC MPEG2016/M39316 (Jan. 2017).

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

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