Point cloud quality evaluation: Towards a definition for test conditions

Luis A. da Silva Cruz, Emil Dumic, Evangelos Alexiou, Joao Prazeres, Rafael Duarte, Manuela Pereira, Antonio Pinheiro, Touradj Ebrahimi

*Department of Electrical and Computer Engineering, University of Coimbra, Portugal
†Instituto de Telecomunicacoes, Portugal
‡Department of Electrical Engineering, University North, Croatia
¶University of Beira Interior, Portugal
¶§Multimedia Signal Processing Group, Ecole Polytechnique Federale de Lausanne, Switzerland

Abstract—Recently stakeholders in the area of multimedia representation and transmission have been looking at plenoptic technologies to improve immersive experience. Among these technologies, point clouds denote a volumetric information representation format with important applications in the entertainment, automotive and geographical mapping industries. There is some consensus that state-of-the-art solutions for efficient storage and communication of point clouds are far from satisfactory. This paper describes a study on point cloud quality evaluation, conducted in the context of JPEG Pleno to help define the test conditions of future compression proposals. A heterogeneous set of static point clouds in terms of number of points, geometric structure and represented scenarios were selected and compressed using octree-pruning and a projection-based method, with three different levels of degradation. The models were comprised of both geometrical and color information and were displayed using point sizes large enough to ensure observation of watertight surfaces. The stimuli under assessment were presented to the observers on 2D displays as animations, after defining suitable camera paths to enable visualization of the models in their entirety and realistic consumption. The experiments were carried out in three different laboratories and the subjective scores were used in a series of correlation studies to benchmark objective quality metrics and assess inter-laboratory consistency.

Index Terms—quality assessment, point cloud, quality metrics

I. INTRODUCTION

In recent years several 3D volume information representation formats, such as light fields, holography and points clouds (PC) have been attracting interest from industry and academia. Due to several important applications, like environment mapping in autonomous driving systems, urban landscape mapping, virtual and augmented reality (VR/AR), as well as the increasing availability of lower priced scanners, PCs are becoming a first choice 3D data representation format. However, PCs can be rather expensive in terms of storage and transmission costs, due to the large number of points used to represent the imaged scene and large number of attributes (e.g., color, normals, etc.) recorded per point. It is thus important to have efficient PC compression algorithms. The ISO/IEC JTC 1/SC 29/WG1 (JPEG) committee, through the activities of the JPEG Pleno project has created an ad-hoc group (AhG) focused on the development of solutions for the efficient representation of static PC. The AhG mandates include a task to “Run subjective tests and assess performance of objective metrics...”, which requires the definition of a suitable framework to conduct PC subjective quality evaluations in order to assemble a set of quality scores that can be used in (objective) quality estimators evaluation and tuning. To properly define the framework, research is needed on which PC rendering methods, what type of displays (2D, 3D, head-mounted), and what evaluation methodologies should be adopted.

In the literature, several studies on subjective quality assessment of PCs have been conducted, based on a wide range of different setups and protocols. In [1], a novel subjective quality evaluation methodology was presented that can be used to assess the difference between an original and a degraded 3D color model. The PC degradations were applied to geometry and color separately, using uniformly distributed noise. In [2] and [3] interactive approaches in a desktop setup and using a head-mounted display were proposed, respectively, in order to subjectively assess geometry-only point clouds subject to Gaussian noise and octree-based compression. In [4], a generic and real-time time-varying PC codec for 3D immersive video has been proposed, which is suitable for mixed reality applications where 3D PCs are acquired at a fast rate. Subjective quality of the codec performance was evaluated in a mixed reality system that combines natural PC data and computer graphics-based 3D contents. In [5], subjective assessment of PC denoising algorithms and evaluation of commonly used objective quality metrics was conducted. The contents were evaluated after applying the screened Poisson surface reconstruction algorithm [6]. The observers rated 2D video sequences of the models after rotations around the vertical and the horizontal axis. In [7] and [8], PCs rendered as surfaces, employing the same surface reconstruction technique [6], were visualized and assessed in 2D and 3D monitors, respectively,
following a passive evaluation. In [9], subjective and objective quality assessment of PCs degraded by compression artifacts was performed. The most popular objective quality metrics were benchmarked against human judgments obtained from the subjective experiment. The degradations considered were compression using Octree-based and graph-based methods. In [10], a subjective evaluation campaign along with benchmarking results of voxelized PCs subject to compression artifacts using the codec proposed in [4] was conducted, showing that geometry distortions are rated more severely with respect to color. In [11], subjective and objective quality assessment results were presented for volumetric video compression using the state-of-the-art MPEG Point Cloud Compression Test Model Category 2 (TMC2). Results show that, although it has a great effect on geometric quality metrics, the higher density input point count does not affect the perceived quality as long as the presentation is plausible for human viewers.

In this work, we extend the previous efforts by including large-scale PCs in our experiment and proposing a passive evaluation scenario through animated videos with different navigation paths per model. The contents under assessment are encoded using two radically different encoding approaches, at different quality levels. The processed stimuli are displayed using a variable point size and camera distance per model and level of degradation, in order to result in perception of watertight models. Finally, the state-of-the-art objective quality metrics are benchmarked and their limitations are discussed.

II. SUBJECTIVE ASSESSMENT

In this section the design of the subjective test is described. In particular, the procedure to prepare the contents under assessment and the subjective evaluation methodology that was adopted are detailed.

A. Content Preparation

In this study, a dataset of 8 static PCs was used with texture and geometry information. In particular, 4 small-scale PCs, namely, _bumbameboi_, _longdress_, _romanoillamp_ and _shiva_, and 4 large-scale PCs, namely, _ucl_, _citiusp_, _ipanemacut_ and _ramos_ were selected. The _bumbameboi_, _citiusp_, _ipanemacut_, _ramos_ and _romanoillamp_ contents can be found in [12], the _longdress_ model can be found in [13], the _ucl_ in [14], while _shiva_ is part of [15]. In this paper, small-scale PCs can be defined as human sized objects or smaller, with up to about 1 million points. A frontal view of the reference PCs of every selected content are illustrated in Figure 1.

To degrade the original PCs, two different compression schemes were employed. The first method is annotated as octree pruning and it was implemented using the Point Cloud Library (PCL) [16]. Octree pruning was selected to represent artifacts occurred after regular removal of points, which are displaced by a known maximum error. Under this type of compression, the contents are enclosed in an octree structure. Modifying the size of the leaf nodes, which is referred as Level of Details (LoD), the resolution of the content is adjusted correspondingly. For instance, by increasing the LoD, the number of points of the compressed object naturally decreases. The second method is a projection-based encoder implemented in the 3DTK toolkit [17]. Different panorama resolutions and equi-rectangular projections provided by the software were used to encode the contents under assessment.

The models were compressed at three quality levels: high, medium and low. Each quality level was defined based on a percentage of remaining points after PCL compression, which was determined per content through expert viewing. In particular, starting from the first method, different octree depths were tested for every model. After defining the LoD that corresponds to the desired quality level, a target percentage of remaining points is identified. For the second method, 3DTK, a binary search tree was employed to determine the appropriate panorama size in order to achieve the required target percentage. The method starts with 32768x32768 pixels, which is a limitation imposed by the OpenCV software used by 3DTK. The number of points for every reference and distorted content, along with the octree depths that were used for octree pruning and the panorama resolution for the projection based compression, can be found in Table I. The 3DTK panorama resolution was determined to achieve similar number of points as PCL compression for every quality level. Thus, the resulting visual quality can be different than the target quality level.

Before encoding, the original PCs were scaled to fit in a bounding box of size 1 and translated at the origin (0,0,0). For both small-scale and large-scale contents, different point sizes per quality level and model were used; however, the same point size was used across one model. Moreover, different distances between the camera and the origin of every content were defined. Several adjustments for the combination of these parameters (i.e., point size and camera distance) were tested, and agreed after expert viewing, in order to result in perception of watertight models. Finally, to account for a fair comparison between the two encoding schemes, the same point size was used for the two codecs per quality level and content. The exact values that were used in our test, as specified in the CloudCompare [18] software, are reported in Table I.

Regarding the generation of animated videos that were consumed by the participants in order to rate the visual quality of the models under assessment, we followed different camera paths per type of content. In particular, for small-scale PCs, to account for an outer inspection of the content, the camera was rotated around the horizontal and, then, around the vertical axis of the center of each model in steps of 1°. For large-scale PCs, excluding _ucl_, the camera was rotated around the vertical axis of the center of the object in steps of 0.5°. In every case, the position of the camera was set so that every content could be viewed entirely. For _ucl_, a navigation path through the scenery was specified, which resulted in a more immersive scenario of consuming a large-scale content inner at different visualization distances. For this content, it was not possible to find one point in space around which the camera would rotate, in order to result in the perception of a watertight scene using fixed point size. A total of 720 still frames of 1920x1080 resolution were captured for every stimulus. The still images
were visually losslessly compressed (i.e., constant rate factor equal to 17) with an H.264/AVC encoder (using FFmpeg software), producing an animated video of 30 fps with a total duration of 24 seconds. The background color was set to black. Finally, the camera paths for inspection, as well as the acquisition of still images was performed using the CloudCompare [18] software.

### B. Evaluation Methodology

The subjective experiments were conducted in 3 different laboratories: University of Beira Interior (UBI), Covilhã, Portugal, University of Coimbra, 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 [19]. The equipment used per test laboratory is described in Table II. A passive subjective methodology was applied, with the subjects visualizing the generated video sequences in the MPV video player which was adjusted for our needs. The participants were able to provide their scores through a customized interface either during, or after the completion of the playback animation. This means, that the subjects were allowed to submit their judgment, potentially, before the end of the animated video.

The simultaneous Double-Stimulus Impairment Scale (DSIS) test method was adopted with 5-levels rating (1 - very annoying, 2 - annoying, 3 - slightly annoying, 4 - perceptible, but not annoying, 5 - imperceptible), including a hidden reference for sanity check. 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 processed 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 rest of the evaluations. Also, particular care was given not to present the same content type consecutively.

At the beginning of each individual evaluation, a training session took place, in order to familiarize the subjects with the artifacts under assessment. The shiva content was selected to represent a small-scale, and the ipanemacut model to represent a large-scale object; thus, they were excluded from the actual subjective tests. The training was performed using 8 animated video sequences. For both training contents, 3 different levels of degradation were presented together with the original video sequences, in order to illustrate the range of visible distortions.

An overall of 42 scores were obtained per evaluation session, considering that each subject assessed 6 test models degraded by 2 compression schemes at 3 distinct quality levels, plus hidden references. An outlier detection algorithm based on ITU-R Recommendation BT.500-13 [19] was applied to the collected scores from every test laboratory, separately. In every case, no outliers were found. Thus, the mean opinion scores (MOS) and the 95% Confidence Intervals (CIs), assuming a Student’s t-distribution were computed on every set of scores. For this calculation, we assumed that scores follow a normal distribution. In Table III, we report observers information, per test laboratory.

### III. RESULTS

In this section, the results of our analysis are provided. The MOS curves are demonstrated, indicating the performance of each codec from every test laboratory, followed by the inter-laboratory correlations. Finally, benchmarking results of the state-of-the-art objective quality metrics are reported.

#### A. Subjective Scores

The subjective scores against the quality levels are shown in Figure 2 for each tested codec and participating laboratory, independently. Specifically, the subjective scores obtained for

### Fig. 2. Subjective scores against quality levels per laboratory and compression method.

- The projection-based encoder (3DTK) and the octree-pruning (PCL), are presented in the left and right plots column, respectively. Based on observers’ opinion, the types of artifacts that are introduced by the octree-based are preferred over the types of artifacts that are caused by the projection-based encoder (as implemented by the 3DTK). This result is justified if we consider that by increasing the size of the points, distortions that occur by an encoding scheme that leads to regularly spaced sparser versions of the original contents, can be compensated. On the contrary, visible distortions, such as missing structures, or areas in a content that are presented due to occlusions, are notably more annoying and, thus, subjects rated them poorly.

#### B. Comparison between Subjective Scores from different Labs

To confirm the validity of the subjective scores that were obtained from the participated test laboratories, we use a series of performance indexes. In particular, 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, to measure the linearity, monotonicity, accuracy and consistency of the results.
Recommendation ITU-T J.149 [21], to check for different
and Correct Decision (CD) percentages were computed using
Ranking (FR), False Differentiation (FD), False Tie (FT)
proposed in the Recommendation ITU-T P.1401 [20], to check
(UE) and Over Estimation (OE) percentages are used, as
the three laboratories results (Bold text represents the ground truth).

Fig. 3. No fitting, Linear and Cubic fitting, for correlation evaluation between
the three laboratories results (Bold text represents the ground truth).

Furthermore, the Correct Estimation (CE), Under Estimation
(UE) and Over Estimation (OE) percentages are used, as
proposed in the Recommendation ITU-T P.1401 [20], to check
for statistically equivalent MOS results. Finally, the False
Ranking (FR), False Differentiation (FD), False Tie (FT)
and Correct Decision (CD) percentages were computed using
Recommendation ITU-T J.149 [21], to check for different
conclusions on data points pairs.

In Figure 3, scatter plots with the MOS obtained in the
different laboratories against every other participated labora-
tory are presented, along with every fitting function that was
applied (i.e., no-fitting, linear fitting, monotonic cubic fitting).
As can be seen, the correlation is high for every regression
model. Thus, in Table IV, the performance indexes without
applying any fitting are reported, to account for the worst case
scenario. Based on these scores, a correct estimation of 100% is
achieved among each pair combination. The correct decision is
very high with a percentage of approximately 90%, while the
false differentiation and false tie remain quite low, with false
ranking being equal to 0%. Naturally, slightly better results
are observed by applying the other two models.

C. Correlation between Subjective and Objective Scores

The subjective scores were correlated with state-of-the-art
objective metrics for quality assessment of PCs. The point-to-
point (p2point) and point-to-plane (p2plane) metrics were used
to estimate geometric distortions [22], using the Mean Squared
Error (MSE) and the Hausdorff distance as the geometric
error measure. The geometry PSNR ratio is also computed,

### TABLE IV

<table>
<thead>
<tr>
<th>Objective Metric</th>
<th>PCC</th>
<th>SRDCC</th>
<th>RMSE</th>
<th>OR (%)</th>
<th>CE (%)</th>
<th>UE (%)</th>
<th>OE (%)</th>
<th>CD (%)</th>
<th>FR (%)</th>
<th>FD (%)</th>
<th>FT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p2point_MSE</td>
<td>0.580</td>
<td>0.728</td>
<td>1.188</td>
<td>0.889</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2plane_MSE</td>
<td>0.504</td>
<td>0.640</td>
<td>1.169</td>
<td>0.889</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2point_Haus</td>
<td>0.698</td>
<td>0.790</td>
<td>0.969</td>
<td>0.667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2plane_Haus</td>
<td>0.644</td>
<td>0.800</td>
<td>1.034</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR_p2point_MSE</td>
<td>0.412</td>
<td>0.421</td>
<td>1.161</td>
<td>0.778</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR_p2plane_MSE</td>
<td>0.439</td>
<td>0.341</td>
<td>1.215</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR_p2point_Haus</td>
<td>0.646</td>
<td>0.607</td>
<td>1.032</td>
<td>0.833</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR_p2plane_Haus</td>
<td>0.667</td>
<td>0.592</td>
<td>1.092</td>
<td>0.778</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE_YUV</td>
<td>0.757</td>
<td>0.760</td>
<td><strong>0.883</strong></td>
<td>0.639</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2plane_MSE</td>
<td>0.310</td>
<td>0.196</td>
<td>1.286</td>
<td>0.889</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSNR</td>
<td>0.550</td>
<td>0.640</td>
<td>1.169</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSIM</td>
<td>0.384</td>
<td>0.434</td>
<td>1.249</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.571</td>
<td>0.619</td>
<td>1.110</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIFP</td>
<td>0.753</td>
<td>0.767</td>
<td>0.890</td>
<td>0.694</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE V

<table>
<thead>
<tr>
<th>Objective Metric</th>
<th>PCC</th>
<th>SRDCC</th>
<th>RMSE</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>p2point_MSE</td>
<td>0.580</td>
<td>0.728</td>
<td>1.188</td>
<td>0.889</td>
</tr>
<tr>
<td>p2plane_MSE</td>
<td>0.504</td>
<td>0.640</td>
<td>1.169</td>
<td>0.889</td>
</tr>
<tr>
<td>p2point_Haus</td>
<td>0.698</td>
<td>0.790</td>
<td>0.969</td>
<td>0.667</td>
</tr>
<tr>
<td>p2plane_Haus</td>
<td>0.644</td>
<td>0.800</td>
<td>1.034</td>
<td>0.861</td>
</tr>
<tr>
<td>PSNR_p2point_MSE</td>
<td>0.412</td>
<td>0.421</td>
<td>1.161</td>
<td>0.778</td>
</tr>
<tr>
<td>PSNR_p2plane_MSE</td>
<td>0.439</td>
<td>0.341</td>
<td>1.215</td>
<td>0.861</td>
</tr>
<tr>
<td>PSNR_p2point_Haus</td>
<td>0.646</td>
<td>0.607</td>
<td>1.032</td>
<td>0.833</td>
</tr>
<tr>
<td>PSNR_p2plane_Haus</td>
<td>0.667</td>
<td>0.592</td>
<td>1.092</td>
<td>0.778</td>
</tr>
<tr>
<td>MSE_YUV</td>
<td>0.757</td>
<td>0.760</td>
<td><strong>0.883</strong></td>
<td>0.639</td>
</tr>
<tr>
<td>p2plane_MSE</td>
<td>0.310</td>
<td>0.196</td>
<td>1.286</td>
<td>0.889</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.550</td>
<td>0.640</td>
<td>1.169</td>
<td>0.861</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.384</td>
<td>0.434</td>
<td>1.249</td>
<td>0.861</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.571</td>
<td>0.619</td>
<td>1.110</td>
<td>0.861</td>
</tr>
<tr>
<td>VIFP</td>
<td>0.753</td>
<td>0.767</td>
<td>0.890</td>
<td>0.694</td>
</tr>
</tbody>
</table>
as defined in [22]: ratio of the max distance of nearest neighbors divided by the squared geometric error value (MSE or Hausdorff). The MSE on the YUV representation was also employed to estimate the color degradations. In particular, the default RGB was initially converted to the YUV colorspace, following the ITU-R Recommendation BT.709-3 [23]. Then, a weight of 6 for luma, and 1 per chroma channel were applied [24], in order to compute a single value between a pair of associated points. Moreover, the plane-to-plane (p2plane) metric [25] is employed. For every content, the normal vectors were estimated on 10 nearest neighbors using PCL. Finally, the projection-based metrics described in [10] were applied on the total number of frames that formed the animated videos which were assessed by the subjects during the experiment.

It should be noted that in order to address excessive memory allocation issues occurred during computations of objective quality metrics and normal estimation algorithms for two large-scale PCs (i.e., ucl and ramos), the latter were segmented in several smaller patches. The outputs of the objective metrics were modified appropriately and pools of individual distortions were obtained to accurately compute the global distortions.

To compare the objective scores against the subjective ground truth, the performance indexes proposed in the Recommendation ITU-T P.1401 [20] are employed. Specifically, the PCC, SROCC, RMSE and OR were issued on pairs of MOS and predicted MOS, to measure the performance of each metric. The predicted MOS for every objective metric, was obtained after applying the monotonic cubic fitting function on the objective scores. Based on our results presented in Table V, the best-performing metrics found to be the MSE_YUV, which captures only color distortions, and the VIFP, which is a projection-based metric, for different ground truth datasets.

IV. CONCLUSIONS

In this study, a subjective methodology for point cloud compression quality evaluation is presented. Our methodology successfully provides effective results for two different codecs, showing that point removal in a regular way is less annoying to human subjects. Moreover the proposed methodology suggests an alternative framework for passive evaluation that reveals high correlation with objective measures. This can be explained because in this experiment we used a variety of point clouds regarding its size and point number, with both geometry and color information. This is opposed to the previous experiments in which point clouds have similar number of points, or only geometry information. The reliability of the method can be confirmed by the comparison between subjective scores between the involved test laboratories which remains high as well as similar correlation between subjective and objective scores in every laboratory.

REFERENCES