Comparing the Quality of Highly Realistic Digital Humans in 3DoF and 6DoF: A Volumetric Video Case Study

Shishir Subramanyam* Jie Li Irene Viola Pablo Cesar

CWI, Amsterdam, The Netherlands

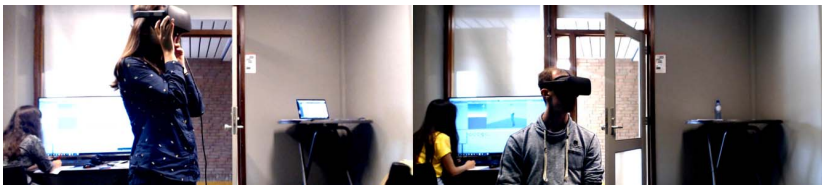


Figure 1: Users Evaluating Realistic Digital Humans in 6DoF (left) and 3DoF (right)

ABSTRACT

Virtual Reality (VR) and Augmented Reality (AR) applications have seen a drastic increase in commercial popularity. Different representations have been used to create 3D reconstructions for AR and VR. Point clouds are one such representation characterized by their simplicity and versatility, making them suitable for real time applications, such as reconstructing humans for social virtual reality. In this study, we evaluate how the visual quality of digital humans, represented using point clouds, is affected by compression distortions. We compare the performance of the upcoming point cloud compression standard against an octree-based anchor codec. Two different VR viewing conditions enabling 3- and 6 degrees of freedom are tested, to understand how interacting in the virtual space affects the perception of quality. To the best of our knowledge, this is the first work performing user quality evaluation of dynamic point clouds in VR; in addition, contributions of the paper include quantitative data and empirical findings. Results highlight how perceived visual quality is affected by the tested content, and how current data sets might not be sufficient to comprehensively evaluate compression solutions. Moreover, shortcomings in how point cloud encoding solutions handle visually-lossless compression are discussed.

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; —Interaction paradigms—Virtual reality;

1 INTRODUCTION

Recent advances in capturing, media processing, and 3D rendering technologies make VR/AR applications popular for mass consumption [34]. In this new media landscape, point clouds are becoming commonplace due to their simplicity and versatility. Still, the size of dense point clouds is significant (a frame of roughly 1M points takes around 19-20 MBytes), which need compression techniques before transmission. This paper provides an exhaustive quality comparison between different encoding configurations of digital humans, represented as point clouds. By investigating the differences in quality, we provide insights about how to optimise the delivery for both downloading and real-time communication. One key novelty of this paper is to study the quality based on realistic consumption conditions, in 3- and 6- Degrees of Freedom (DoF) scenarios.

*e-mail: {S.Subramanyam, Jie.Li, Irene.Viola, P.S.Cesar}@cwi.nl

Avatars are a core part of VR applications like social communication [28], sports training [21], or healthcare [20]. A major line of scientific work has focused on how to make such avatars more realistic, interactive, and autonomous [10, 24, 33]. In this paper, we focus instead on point clouds as a suitable representation for digital humans based on tele-portation principles [25]. In this case, the research problem is not so much how to render and animate them to make them look more realistic, but how to transport them optimally.

Given current advances in technology, real-time delivery of point clouds is becoming a realistic alternative; focusing the attention of the research community [23] and industry [32] in encoding and transmission. Still, given the massive number of points per representation, decisions need to be taken regarding the delivery (type of encoder, bit-rate) to ensure an acceptable quality of experience depending on the viewing conditions (3DoF, 6DoF). This is the core research question this paper answers.

Contributions of the paper are two-fold: 1) It provides a first evaluation of the quality of highly realistic digital humans represented as dynamic point clouds in immersive viewing conditions. Existing protocols [5, 7, 8, 40, 42] did not consider the dynamic of the point clouds, focused on one type of data set, and did not take into account VR viewing conditions; 2) It provides quantitative subjective results about the perceived quality of the contents, along with qualitative insights on what is important for users in interacting with digital humans in VR. Such results will help in better configuring the network conditions for the delivery of points clouds for real-time transmission, and have implications over ongoing research and standardisation work regarding the underlying compression technology.

Particularly, this paper extensively studies this current and relevant area of research by proposing 1) a new evaluation protocol, including the work to create dynamic point clouds for evaluation, and 2) quality of experience results. These results are based on an experiment with 52 participants, evaluating 72 stimuli based on eight dynamic point cloud sequences. Each point cloud sequence was compressed in four bit-rates, using two types of compression techniques. These 72 stimuli were evaluated in two viewing conditions (3DoF and 6DoF). The data gathered include rating scores, presence questionnaires, simulator sickness reports, and time spent watching the content. The results indicate that, while bit-rate savings can be obtained by choosing one compression solution over another, visually lossless compression has not been fully achieved by the algorithms under evaluation, even at rather large bit-rates. Moreover, the choice of content can have an impact on how users rate its quality, influencing the discriminating power of the selected protocol.

2642-5254/20/$31.00 ©2020 IEEE
DOI 10.1109/VR46266.2020.00-73

(a) *V-PCC* (b) *MPEG Anchor*

Figure 2: Point Cloud Digital Humans compressed using two point cloud codecs, V-PCC (left) and MPEG anchor (right), at the 4 selected bit-rates.

2 Related work

2.1 Quality assessment for point clouds

Capturing and displaying volumetric videos is becoming feasible [2, 30]. Point clouds are frequently used as a data format for volumetric video in augmented reality (AR) and virtual reality (VR) applications. Point clouds collate a large number of geo-referenced points to represent humans or objects in 3D. The color information can be provided with each point [40]. To visualize 3D content sufficiently, the number of points must be high, which results in large size and increases the difficulty to store and transmit the point clouds. To support low latency transmission in AR/VR applications within a limited bandwidth, compression is necessary. However, it remains challenging to measure and predict the acceptable quality of compressed point clouds.

There is a growing interest on subjective quality assessment of point clouds rendered on 2D displays. Zhang et al. [42] evaluated the quality degradation effect of resolution, shape and color on static point clouds. The results indicate that resolution is almost linearly correlated with the perceived quality, and color has less impact than shape on the perceived quality. Zerman et al. [40] compressed two dynamic human point clouds using a state-of-the-art algorithm [22], and assessed the effects of this algorithm and input point counts on the perceived quality. Their results showed that no direct correlation was found between human viewers' quality ratings and input point counts. In a recent study [11], a protocol to conduct subjective quality evaluations and benchmark objective quality metrics were proposed. The viewers passively assessed the quality of a set of static point clouds, as animations with pre-defined movement path. In a comprehensive work by Alexiou et al. [8], the entire set of emerging point cloud compression encoders developed in the MPEG committee were evaluated through a series of subjective quality assessment experiments. Nine static models, including both humans and objects, were used in the experiments. The experiments provided insights regarding the performance of the encoders and the types of degradation they introduce.

Only a limited number of point cloud quality assessment studies have been conducted in immersive environments. Mekuria et al. [23] evaluated the subjective quality of their codec performance in a realistic 3D tele-immersive system, in which users were represented as 3D avatars and/or 3D dynamic point clouds, and could navigate in the virtual space using mouse cursor in a desktop setting. Several aspects of quality, such as level of immersiveness, togetherness, realism, quality of motion, were considered. Alexiou and Ebrahimi [7] proposed the use of AR to subjectively evaluate the quality of colorless point cloud geometry. Tran et al. [37] suggested that, in case of evaluating video quality in an immersive setup, aspects such as cybersickness and presence should not be overlooked.

For the objective evaluation of point clouds, there are two main approaches. Considering the availability of point location and color information, either point-based or projection-based metrics can be used [36]. Current point-based approaches can assess either geometry- or color-only distortions. For geometry errors, three metrics were commonly used in studies [3, 6, 7, 35, 36], namely the point-to-point metrics, the point-to-plane metrics and the plane-to-plane metrics. These metrics computed using the root mean square (RMS) distance, mean square error (MSE) or Hausdorff distance. Moreover, the geometric Peak-Signal-to- Noise-Ratio (PSNR) is used for the point-to-point and point-to-plane metrics [35]. For color distortions, the total color degradation value is based either on the color MSE, or the PSNR, computed in either the RGB or the YCbCr color spaces [8]. The projection-based approaches map the rendered models onto planar surfaces, and conventional 2D imaging metrics are employed [12]. A comprehensive study [8] showed that the performance of the current objective metrics is not ideal, revealing the need for better solutions. Therefore, in this study, we do not include any above-mentioned objective metrics, and focus on subjective quality assessment.

2.2 Point cloud compression

A single point cloud frame is represented by an unordered collection of points sampled from the surface of an object. In a dynamic sequence of point clouds, there are no correspondences of points maintained across frames. Thus, detecting spatial and temporal redundancies is often difficult, making point cloud compression challenging. Octrees have been used extensively as a space partitioning structure to represent point cloud geometry. They are a 3D extension of the 2D quadtree used to encode video and images.

Research into point cloud compression can be broadly divided into two categories. The first is based on signal processing, Zhang et al. [41] proposed a method to compress point cloud attributes using a graph Fourier transform. They assume that an octree has been created and separately coded for geometry prior to coding attributes. De Queroz and Chou [27] used a region adaptive hierarchical transform to use the colors of nodes in lower levels of the octree to predict the colors of nodes in the next level. As these approaches require expensive computations of graph laplacians, they are not suitable for dynamic sequences in real-time applications. The second category of point cloud codecs are based on extending legacy solutions from image and video compression. Intra Frame coding in octrees can be achieved by entropy coding the occupancy codes, as shown in [23]. The authors then compress the color attributes by mapping them to a 2D grid and using legacy JPEG image compression.

In 2017, MPEG started a standardization activity to determine a new standard codec for point clouds, to be launched in 2020. They used the codec created by Mekuria et al. [23] as an anchor to evaluate proposals. To encode dynamic point cloud sequences MPEG provides two verification models [32], Geometry-PCC for point clouds with a sparse distribution, and Video-PCC for dense point clouds. V-PCC is based on leveraging existing 2D video codecs to compress point cloud geometry and attributes.

3 METHODOLOGY

3.1 Dataset Preparation

A dataset of dynamic point cloud sequences was used from the MPEG repository. All sequences were clipped to five seconds and sampled at 30 frames per second. This included point cloud sequences [13] [14] captured using photogrammetry (Longdress, Loot, Red and black, Soldier are shown in figure 3) and one sequence of a synthetic character sampled from an animated mesh (Queen). Four additional point cloud sequences; Manfred, Despoina, Sarge (shown in Figure 3) and Rachel were added for the evaluation. These sequences were created using motion captured animated mesh sequences.

Keyframes were selected at 30 frames per second and extracted along with the associated mesh materials. Particular care was put in ensuring the selected sequences have the characters facing the user and speaking in their general direction. Then, 1 million points were randomly sampled, independently per key frame to create a consistent groundtruth dataset. The points are sampled from the mesh surface with a probability proportional to the area of the underlying mesh face. This was done to ensure no direct point correspondences across point cloud frames, to mimic realistic acquisition and maintain consistency with the rest of the dataset. The point clouds sampled from meshes were used in Test T1 and the point clouds captured using photogrammetry were used in test T2. The X, Y, Z coordinates of each point is represented using an unsigned integer, as is required for the current version of the V-PCC software. Colors are encoded as 8bits per color in the RGB color space.

To encode the contents, we first use Release 7.0 of the VPCC MPEG codec. For test 1, the configuration files provided by MPEG for the *Queen* sequence are used for all the contents. We select the rate points 1, 3 and 5 from the provided preset V-PCC configurations and extend in to an additional final rate point using a Texture quantization parameter (QP) of 8, a geometry QP of 12 and an occupancy precision of 2. We re-label the rate points as R1, R2, R3 and R4, respectively. All sequences are encoded using the C2AI (Category 2 All Intra) config. For the photogrammetry sequences, we use the predefined dedicated configuration files for each sequence, at the same rate points. The VPCC compressed bitstream was used to set the bitrate targets for R1 to R4, separately for each sequence.

We then use the MPEG anchor codec [23] in an all intra configuration, and match the bit-rates per sequence and rate point (R1-R4) with a tolerance of 10%, as defined in the MPEG call for proposals. The codec was selected as it has a significantly lower encode and decode time and is suitable for real-time applications, as demonstrated by the authors. We use an octree depth from 7 to 10 for the rate points R1 to R4 respectively. The highest possible JPEG quantization parameter values were then chosen per sequence, while meeting the target bit rate set using VPCC.

3.2 Experiment setup

All point cloud sequences were rendered using the Unity game engine, by storing all the points of each frame in a vertex buffer, and then drawing procedural geometry on the GPU. The point clouds were rendered using a quadrilateral at each point location with a fixed offset of 0.08 units (this corresponds to a side length of approximately 2mm) around each point (placed at the centre) for all the sequences, to be consistent. In the case of bitrate R1 generated using the MPEG anchor, we increased the offset value to 0.16 by eye, as the resulting point clouds were too sparse (shown in Figure 2b). We maintain a fixed frame rate of 30fps throughout the experiment.

Participants were asked to wear an Oculus Rift Head Mounted Display to view each of the point cloud sequences. For the 3DoF condition, participants were asked to sit on a swivel chair placed at a fixed location in the room and navigate using head movements alone. For the 6DoF condition, participants were allowed to navigate freely within the room, as shown in Figure 1. Each sequence was 5 seconds long, after which the playback looped around. We set the background of the virtual room to mid-grey, to avoid distractions. The Oculus Guardian System was used to display in-application wall and floor markers if the participants got too close to the boundary. We used a workstation with 2 GeForce GTX 1080 Ti in SLI for the GPU and an Intel Core i9 Skylake-X 2.9GHz CPU.

3.3 Subjective methodology

To perform the experiments, the subjective methodology Absolute Category Rating with Hidden References (ACR-HR) was selected, according to ITU-T Recommendations P.910 [15]. Participants were asked to observe the video sequences depicting digital humans, and rate the corresponding visual quality on a scale from 1 to 5 (*1-Bad*, *2-Poor*, *3-Fair*, *4-Good*, and *5-Excellent*).

A series of pilot studies were conducted to determine the positioning of digital humans in the virtual space and the length of each sequence, to ensure the sequences were running smoothly within the limited computer RAM. Due to the huge size of the test material, it was not possible to evaluate all 8 point cloud contents in one single session, as long loading times would have brought fatigue to the participants and corrupted the results. Thus, we decided to split the evaluation into two separate tests: one focused on the evaluation of contents obtained from random sampling of meshes (**T1**: contents *Queen*, *Manfred*, *Despoina* and *Sarge*), and one focused on contents acquired through photogrammetry (**T2**: contents *Long dress*, *Soldier*, *Red and black*, and *Loot*). From each sequence, a subset of frames comprising 5 seconds was selected.

Before the test took place, 3 training sequences depicting examples of *1-Bad*, *5-Excellent* and *3-Fair* were shown to the users to help them familiarize with the viewing condition and test setup, and to guide their rating. The training sequences were created using one additional content not shown during the test, to prevent biased results. For test T1, content *Ana* was selected, whereas for test T2, content *Ulli Wagner* was chosen. Each content sequence was encoded using the point cloud compression algorithms under test.

For each test and viewing condition, 36 stimuli were evaluated. For each stimulus, the 5 second sequence was played at least once in full, and kept in loop until the participants gave their score. The order of the displayed stimuli was randomized per participant and per viewing condition, and the same content was never displayed twice in a row to avoid bias. Moreover, the presentation order of viewing conditions was randomized between participants, to prevent any confounding effect. Two dummy samples were added at the beginning of each viewing session to ease participants into the task, and the corresponding scores were subsequently discarded.

After each view condition, participants were requested to fill in the Igroup Presence Questionnaire (IPQ) [31] on a 1-7 discrete scale (1=fully disagree to 7=totally agree) and Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [18]. IPQ has three subscales, namely Spatial Presence (SP), Involvement (INV) and Experienced Realism (REAL), and one additional general item (G) not belonging to a subscale, which assesses the general "sense of being there", and has high loadings on all three factors, with an especially strong loading on SP [31]. SSQ was developed to measure cybersickness in computer simulation and was derived from a measure of motion sickness [18]. For both T1 and T2, after the two viewing conditions, participants were interviewed to 1) compare their experiences of assessing quality in 3DoF and 6DoF, and 2) reflect on the factors they considered when assessing the quality.

A total of 27 participants were recruited for T1 (12 males, 15 female, average age: 22,48 years old), whereas 25 participants were recruited for T2 (17 males, 8 females, average age: 28,39 years old). All participants were screened for color vision and visual acuity, using Isihara and Snellen charts, respectively, according to ITU-T Recommendations P.910 [15].

Figure 3: Sequences used for the test, from left to right: Manfred, Sarge, Despoina, Queen, Longdress, Loot, Red and black, Soldier

3.4 Data analysis

Outlier detection was performed separately for each test T1 and T2, according to ITU-T Recommendations P.913 [16]. The recommended threshold values $r_1 = 0.75$ and $r_2 = 0.8$ were used. One outlier was found in test T1, and the corresponding scores were discarded. No outliers were found in the scores collected for test T2.

After outlier detection, the Mean Opinion Score (MOS) was computed for each stimulus, independently per viewing condition. The associated 95% Confidence Intervals (CIs) were obtained assuming a Student's t-distribution. Additionally, the Differential MOS (DMOS) was obtained by applying HR removal, following the procedure described in ITU-T Recommendations P.913 [16].

Non-parametric statistical analysis was applied to understand whether statistical differences could be found among variables, using the MATLAB Statistics and Machine Learning Toolbox, along with the ARTool package in R [17].

4 Results

4.1 Subjective quality assessment

Figures 4 and 5 shows the results of the subjective quality assessment of the contents comprising test T1 and test T2, respectively, for both 3DoF and 6DoF viewing conditions. In particular, the MOS scores associated with the compressed contents are shown with solid lines, along with relative CIs, whereas the dashed lines represent the respective DMOS scores. The HR scores for each content are represented with a solid line to indicate the mean, and a shaded plot for the corresponding CIs.

To assess whether significant differences could be found between the two visual conditions under test, we ran a Wilcoxon signed-rank test on the scores obtained in the two DoF scenarios. The Wilcoxon test was chosen as the gathered data was not found to be normally distributed, according to the Shapiro-Wilk normality test ($W = 0.90$, $p < .001$ and $W = 0.91$, $p < .001$ for tests T1 and T2, respectively). Results of the Wilcoxon signed-rank test showed statistical significance for DoF for test T1 ($Z = 2.97$, $p = 0.0029$, $r = 0.07$), whereas for test T2, no significance was found ($Z = -1.96$, $p = 0.0502$, $r = 0.05$). Values seems to indicate an effect of the DoF in test T1; however the small r-value indicates that while the effect apparently exists, it is small.

It can be observed that codec V-PCC has generally a more favorable performance with respect to the MPEG anchor. This is especially evident for the contents acquired through photogrammetry (see Fig. 5), for which the gap among the two codecs is more pronounced. Wilcoxon signed-rank test confirmed statistical significance for the two codecs (T1: $Z = 9.87$, $p < .001$, T2: $Z = 20.18$, $p < .001$), albeit with different effect sizes between test T1 and T2 ($r = 0.24$ and $r = 0.50$, respectively).

A Friedman rank test performed on the scores revealed a significant effect of the content on the final scores, for both sets of contents (T1: $\chi^2 = 57.38$, $p < .001$, T2: $\chi^2 = 17.31$, $p < .001$). Table 1 shows the results of the post-hoc test conducted using Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/6$). Contents *Manfred*, *Sarge* and *Despoina* all show statistical significance with respect to content *Queen* ($p < .001$, $r > 0.20$ for all pairs). Statistical significance has also been observed between content *Manfred* and *Sarge*, albeit with a smaller effect size ($p < .001$, $r = 0.12$). For contents acquired through photogrammetry, statistical significance was found between contents *Long dress* and *Loot*, and *Loot* and *Red and black* ($p < .001$, $r > 0.20$ in both cases), as well as between contents *Long dress* and *Soldier*, *Loot* and *Soldier* ($p < .001$, $r > 0.10$), and *Red and black* and *Soldier* ($p = 0.0019$, $r = 0.10$). Results corroborate our previous statements on how contents *Long dress* and *Red and black* appeared to be given different scores with respect to contents *Loot* and *Soldier*.

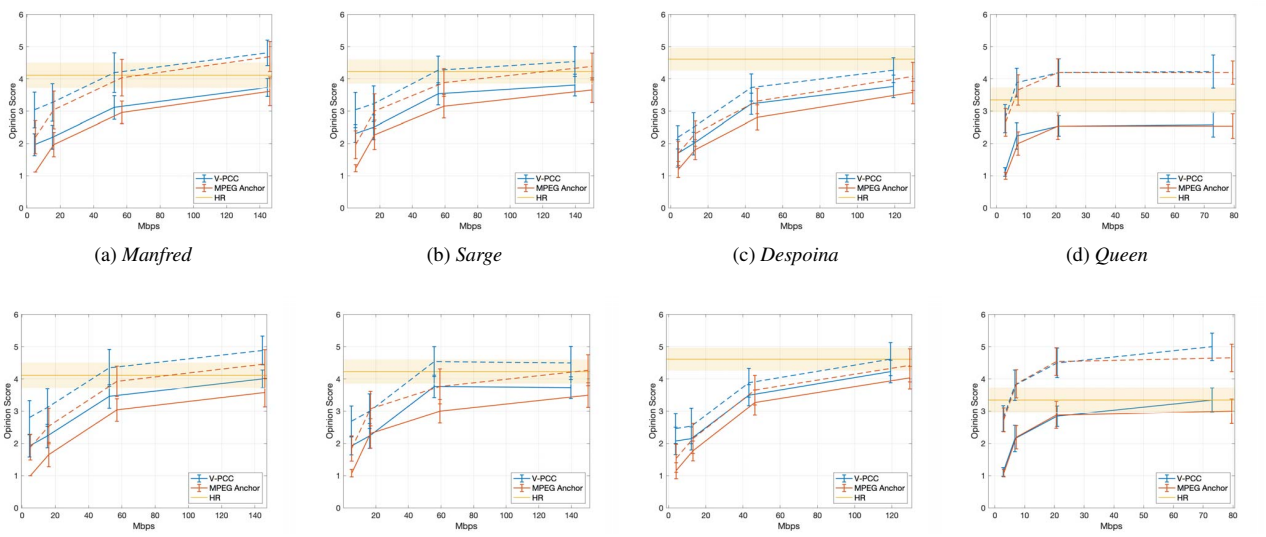
We also ran a Friedman rank test on the scores to assess whether the selected bit-rates were showing statistical significance. Results confirmed that the bit-rates have a significant effect for both tests (T1: $\chi^2 = 682.29$, $p < .001$, T2: $\chi^2 = 667.39$, $p < .001$). Post-hoc

Table 1: Pairwise post-hoc test on the contents for test T1 and T2, using Wilcoxon signed-rank test with Bonferroni correction.

		Z	p	r
T1	*Manfred - Sarge*	3.78	<.001	0.12
	Manfred - Despoina	2.09	0.036	0.07
	Manfred - Queen	7.48	<.001	0.25
	Sarge - Despoina	1.30	0.192	0.04
	Sarge - Queen	9.94	<.001	0.33
	Despoina - Queen	8.79	<.001	0.29
T2	*Long dress - Loot*	7.03	<.001	0.23
	Long dress - Red and black	1.08	0.279	0.05
	Long dress - Soldier	4.11	<.001	0.14
	Loot - Red and black	6.42	<.001	0.21
	Loot - Soldier	3.32	<.001	0.11
	Red and black - Soldier	3.10	0.002	0.10

Table 2: Pairwise post-hoc test on the bitrates for test T1 and T2, using Wilcoxon signed-rank test with Bonferroni correction.

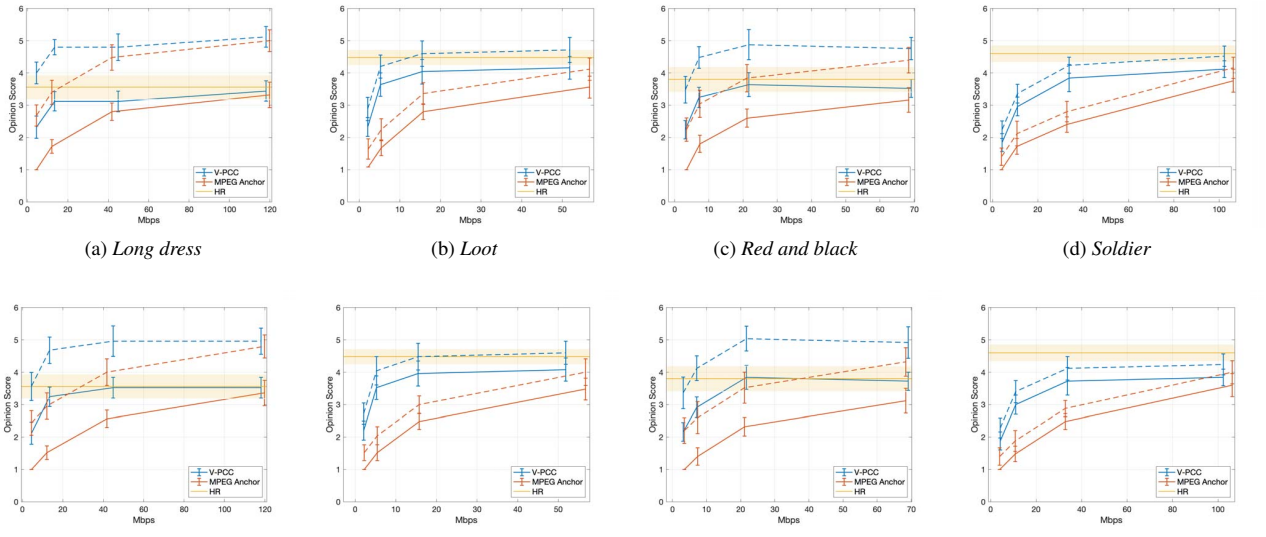
		Z	p	r
T1	R1 - R2	-14.21	<.001	0.50
	R1 - R3	-16.85	<.001	0.60
	R1 - R4	-17.08	<.001	0.60
	R2 - R3	-12.61	<.001	0.45
	R2 - R4	-14.45	<.001	0.51
	R3 - R4	-8.75	<.001	0.30
T2	R1 - R2	-14.20	<.001	0.50
	R1 - R3	-16.85	<.001	0.60
	R1 - R4	-17.08	<.001	0.60
	R2 - R3	-12.61	<.001	0.45
	R2 - R4	-14.45	<.001	0.51
	R3 - R4	-8.57	<.001	0.30



(a) *Manfred* (b) *Sarge* (c) *Despoina* (d) *Queen*

(e) *Manfred* (f) *Sarge* (g) *Despoina* (h) *Queen*

Figure 4: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps. HR scores are shown using a shaded yellow plot. Each column represents a content in test T2, whereas first row and second row depict results obtained using the viewing conditions 3DoF and 6DoF, respectively.



(a) *Long dress* (b) *Loot* (c) *Red and black* (d) *Soldier*

(e) *Long dress* (f) *Loot* (g) *Red and black* (h) *Soldier*

Figure 5: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps. HR scores are shown using a shaded yellow plot. Each column represents a content in test T2, whereas first row and second row depict results obtained using the viewing conditions 3DoF and 6DoF, respectively.

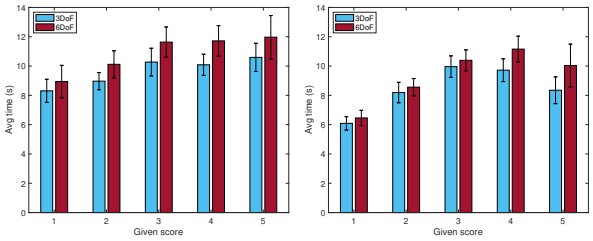
analysis using Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/6$), shown in Table 2 further confirmed that all pairwise comparisons were statistically significant, for both test T1 and T2 ($p < .001$, $r > 0.30$ for all pairs).

In order to further analyze the effect of DoF conditions, contents, codecs and bit-rates, and relative interactions, on the gathered scores, we fitted a full linear mixed-effects model on the data, accounting for randomness introduced by the participants. Due to the non-normality of our data, the aligned rank transform was applied prior to the fitting [39]. Since the transform is designed for a fully randomized test, it is not suitable for the scores collected during the test, as the HR addition makes the design matrix rank deficient. However, the transform can be applied to the differential scores used to obtain DMOS, as it follows a fully randomized design. Thus, it was decided to perform the analysis on the differential scores.

For test T1, analysis of deviance on the full mixed-effects model showed significance for main effects Content ($F = 48.14$, $df = 3$, $p < .001$), Codec ($F = 51.01$, $df = 1$, $p < .001$) and bit-rate

($F = 375.35$, $df = 3$, $p < .001$), but not for DoF ($F = 0.0003$, $df = 1$, $p = 0.988$). Moreover, significant interaction effects were found for *DoF - Content* ($F = 4.31$, $df = 3$, $p = 0.005$), *Content - bit-rate* ($F = 5.88$, $df = 9$, $p < .001$) and *Codec - bit-rate* ($F = 4.73$, $df = 3$, $p = 0.003$). Post-hoc interaction analysis with Holm p-value adjustment indicates that the difference between 3DoF and 6DoF has statistical significance at 5% level when comparing contents *Manfred* and *Queen* ($\chi^2 = 10.34$, $p = 0.008$), as well as *Inspector* and *Queen* ($\chi^2 = 8.35$, $p = 0.019$). In other words, the relative difference in scores between contents *Manfred* and *Queen* (and *Inspector* and *Queen*) was not found to be statistically equivalent in 3DoF with respect to 6DoF. This indicates that the DoF might have an effect on how contents are scored with respect to one another, for example by increasing or reducing their differences. Regarding the interaction effect between contents and bit-rates, post-hoc interaction analysis with Holm p-value correction showed statistical significance in differences between contents *Manfred* and *Queen* at bit-rates R2 and R4 ($\chi^2 = 29.52$, $p < .001$), between contents *Sarge* and *Despoina* at bit-rates R2 and R4 ($\chi^2 = 11.00$, $p = 0.028$), between *Sarge* and *Queen* at bit-rates R2 and R4 ($\chi^2 = 11.56$, $p = 0.022$), and between *Despoina* and *Queen* at bit-rates R1 and R2 ($\chi^2 = 13.75$, $p = 0.007$), R2-R3 ($\chi^2 = 13.59$, $p = 0.007$) and R2-R4 ($\chi^2 = 45.13$, $p < .001$). Results can be explained considering that the low HR scores given to content *Queen* meant a narrower range of ratings. Thus, bit-rate point R2, for example, presents relatively higher differential scores for *Queen* with respect to the rest of the contents, whereas for bit-rate point R4, due to the HR removal, all contents have similar ratings. This is reflected in the statistical analysis conducted on the scores. Finally, post-hoc interaction analysis with Holm p-value adjustment on differences between codecs and bit-rates shows that the difference among codecs is statistically significant at 5% level only between R1 and R2 ($\chi^2 = 10.51$, $p = 0.007$), R1 and R3 ($\chi^2 = 7.09$, $p = 0.031$), and R1 and R4 ($\chi^2 = 10.17$, $p = 0.007$). This indicates that the differences between codecs remain constant at all bit-rates, except for R1. This is in line with what observed in Fig. 4, which show similar trends for codec V-PCC with respect to the MPEG anchor, except for the lowest bit-rate point, for which V-PCC achieves better performance.

Results of analysis of deviance on the full mixed-effects model for test T2 showed significance for main effects Content ($F = 139.41$, $df = 3$, $p < .001$), Codec ($F = 692.24$, $df = 1$, $p < .001$) and bit-rate ($F = 485.11$, $df = 3$, $p < .001$), but not for DoF ($F = 2.57$, $df = 1$, $p = 0.115$), similarly to what was seen for test T1. Interactions were found significant at 5% level between Content and Codec ($F = 3.81$, $df = 3$, $p = 0.01$), Content and bit-rate ($F = 3.03$, $df = 9$, $p = 0.001$), and Codec and bit-rate ($F = 39.40$, $df = 3$, $p < .001$). The lack of significance in interactions involving DoF is in line with the results of the Wilcoxon signed-rank test, which showed no significance for DoF in test T2 ($Z = -1.96$, $p = 0.0502$, $r = 0.05$). Post-hoc interaction analysis with Holm p-value adjustment shows significance at 5% level for the differences among codecs for content *Long dress* with respect to content *Loot* ($\chi^2 = 10.09$, $p = 0.009$). This confirms what can be seen in Fig. 5: the gap among codecs is more prominent for content *Loot* with respect to *Long dress*, probably due to the reduced range associated with a low-rated HR. Post-hoc analysis on the interaction between contents and bit-rates indicates statistical significance at 5% level for differences among contents *Long dress* and *Soldier* when considering differences between bit-rates R1-R4 ($\chi^2 = 17.03$, $p = 0.001$) and R2-R4 ($\chi^2 = 11.81$, $p = 0.021$), and among contents *Red and black* and *Soldier* for differences between R1 and R4 ($\chi^2 = 11.80$, $p = 0.021$). Again, this can be explained considering that both *Long dress* and *Red and black* received remarkably lower scores, which resulted in a narrower rating range. Thus, differences among lowest and highest bit-rates are quite different between those two contents and *Soldier*, which benefited from a larger rating span. Lastly, post-hoc analysis on the interaction between codecs and bit-rates reveals statistical significance at 5% level for all pairwise comparison, except R1-R3 (R1-R2: $\chi^2 = 14.60$, $p < .001$, R1-R4: $\chi^2 = 46.58$, $p < .001$, R2-R3: $\chi^2 = 13.81$, $p < .001$, R2-R4: $\chi^2 = 113.34$, $p < .001$, R3-R4: $\chi^2 = 48.02$, $p < .001$). Indeed, in Fig. 5 it is quite evident that the curves for the two codecs follow different trends. In particular, codec V-PCC seems to saturate between R2 and R3, whereas a steeper slope is observed for the MPEG anchor.



(a) T1 (b) T2

Figure 6: Average time spent looking at the sequence (in seconds) and relative CIs, against score given to the sequence, for 3DoF (blue) and 6DoF (red), in test T1 (left) and T2 (right).

4.2 Additional questionnaires and interaction data

4.2.1 IPQ & SSQ Questionnaires

For T1 and T2, the collected IPQ data under each subscale are all normally distributed as examined by the Shapiro-Wilk test ($p > 0.05$). A paired sample t-test was applied to check the differences between 3DoF and 6DoF in terms of SP, INV, REAL and G. For T1, there was a significant difference in SP between 3DoF (M=4.13, SD=0.92) and 6DoF (M=5.04, SD=0.67), t(26)=-4.44, $p < .001$, Cohen's d = 0.52 and also a significant difference in G between 3DoF (M=4.11, SD=1.28) and 6DoF (M=4.96, SD=1.13), t(26)=-2.60, $p < .01$, Cohen's d = 0.64. For T2, SP was also significantly different in 3DoF (M=4.16, SD=1.17) and 6DoF (M=4.83, SD=1.12), t(24)=-3.48, $p < .01$, Cohen's d = 0.45 and so was G between 3DoF (M=4.20, SD=1.61) and 6DoF (M=5.08, SD=1.19), t(24)=-3.56, $p < .01$, Cohen's d = 0.71. Other factors showed no significant differences between 3DoF and 6DoF in both T1 and T2.

With respect to SSQ, no significant differences ($p > 0.05$) were found between 3DoF and 6DoF in terms of cybersickness. We further tested whether there were order effects in experiencing cybersickness, where half of the participants started with 6DoF as the first condition and 3DoF as the second, and the remainder the inverse. No significant differences ($p > 0.05$) were found for any order effects in experiencing cybersickness.

4.2.2 Interaction time

Interaction time was found to be strongly correlated with MOS values in a study conducted on light field image quality assessment [38]. In particular, it was found that users tended to spend more time interacting with contents at high quality, whereas for low quality scores, less time was spent looking at the contents. In order to see whether similar trends could be observed in our data, we compared the average time spent watching the sequence in 3DoF and 6DoF, separately for each quality score given by the participants. Results are shown in Fig. 6. A positive trend can be observed between the given score and the average time spent looking at the sequence, with the exception of score 5, which for test T2 shows a negative trend with respect to the time. However, it should be considered that on average, a small percentage of scores equal to 5 were given in test T2 (10% of the total scores), thus, variations may be due to the difference in sample size. It is also worth noting that, on average,

participants spent more time looking at the sequences in 6DoF, with respect to the 3DoF case. Indeed, several participants pointed out that the lowest scores were the fastest to be given, whereas for higher quality, it was harder to decide on the rating.

4.2.3 Interviews

We asked the same interview questions for T1 and T2. So, we combined the interview transcripts of 52 participants (T1=27, T2=25). The categorized answers are presented as follows:

Factors considered when assessing quality. 56% of the participants mentioned that they assessed the quality based on three criteria: 1) overall outline and pattern distortion on body and on clothes, 2) natural gestures and movements of the digital humans, and 3) visual artifacts such as blockiness, blurriness, and extraneous floating artifacts. 48% of the participants mentioned the quality assessment criteria are content related, who agreed that it is easier to spot artifacts for the content with complex patterns (e.g., *Long dress*) and dominant colors (e.g., *Red and black*) than the content with uniformed colors (e.g., *Soldier* and *Sarge*). 46% of the participants considered facial expressions as an unignorable factor for quality assessment, which they believe is an important cue for social connectedness. For the extraneous floating artifacts (e.g., bubbles flickering outside the digital humans), 23% found it very annoying and lowered the overall quality for the content, but a few participants (8%) thought these artifacts do not influence their quality judgement.

Difficulties in assessment. 42% of the participants pointed out the difficulties in assessing the quality, especially for the high quality contents, which are not perfect and still have missing details like blurry faces or wrong fingers. 15% of the participants specifically pointed out that it is difficult to distinguish between quality level 3 to 5. 17% of the participants commented that it gradually became easier in rating the quality when they adapted to the contents. So, the second viewing condition was easier for them.

Comparison between 3DoF and 6DoF. 52% of the participants preferred 6DoF, because it allowed them to move closer to examine the details (e.g., shoes and fingers). They felt more realistic when walking in the virtual space. However, they also commented that 3DoF offered a fixed distance between them and digital humans, enabling a more stable and focused assessment. 21% of the participants preferred relaxation and passiveness in 3DoF, because they did not find much differences between 3DoF and 6DoF in terms of quality assessment, but they found 3DoF is less nauseous than 6DoF.

4.3 Analysis of results

Results vary considerably depending on the content under assessment. In particular, for test T1, content *Queen* is generally given lower ratings with respect to the other contents in the test. This is made evident by the MOS score given to the HR, which is equal to 3.35 for the 3DoF and 6DoF condition, indicating that even when uncompressed, the content was never considered as having a good quality. As a result, the MOS scores computed for the content have a limited range, spanning between 1.08 and 3.35 for the 6DoF case, and between 1 and 2.58 for the 3DoF (excluding HR). Such a narrow range is inadequate in expressing the quality variations among different compression parameters: for the 3DoF case in particular, paired t-test at 5% significance shows that bit-rate points R3 and R4 are statistically equivalent for both codecs, and for codec V-PCC R2 is considered statistically equivalent to R4, despite the latter being 10 times as large. Statistical analysis results confirmed that content *Queen* showed different rating patterns with respect to the other contents. The ratings given to the rest of the contents comprising T1 have a larger range, seemingly covering the entire rating space. Trends show that codec V-PCC is generally preferred to the MPEG anchor, especially at low bit-rates, whereas for the highest bit-rate point R4, the codecs are always statistically equivalent at 5% confidence level, in both 3DoF and 6DoF.

It is worth noting that the two codecs seldom reach the same quality level as the uncompressed HRs. In particular, for the 3DoF viewing condition, transparent quality (as in, the level of quality for which the distortions are "transparent" to the user, meaning that statistical equivalence with the HR has been observed) is only achieved by content *Manfred* at bit-rate R4, by both codecs. On the other hand, in the 6DoF scenario, V-PCC encoded contents at bit-rate R4 seem to always be statistical equivalent to the HR.

Rating variability among different contents is even more visible for contents acquired through photogrammetry. In particular, at high bit-rates contents *Long dress* and *Red and black* are consistently given lower scores with respect to contents *Loot* and *Soldier*, for both DoF conditions. In fact, as seen with content *Queen* above, both contents do not reach MOS levels higher than 4, even when considering the HR content. This indicates that the source content is never considered of excellent quality, even when no compression artifact is involved. This impacts the way scores are distributed across the rating space: left with a smaller rating range (as the higher rating values are never given), MOS results show that contents compressed at high bit-rates are considered statistically equivalent to the respective HR. This is particularly evident when considering the DMOS scores, which have an operational range between 4 and 5 for codec V-PCC, for a range of bit-rates spanning between 4 and 120 Mbps. On the other hand, for contents that were given higher scores (*Loot* and *Soldier*) and use the full rating space, results indicate what was already seen in test T1: no codec is able to reach transparent quality, meaning that scores given to the compressed content are always statistically different with respect to the HR.

Decisions on which codec to employ should be made depending on the use case. The MPEG anchor is more suitable for real-time system, due to its fast encoding time, and at high enough bitrates, differences with the other codec become less noticeable. V-PCC, on the other hand, might be more appropriate for on-demand streaming and storage, since it retains better quality for the same bitrate. For the majority of the contents under test, a bitstream size between 20 and 40Mbps seems to provide an acceptable quality. However, regarding the selection of the appropriate target bitrate, the decision should be made taking into account other factors, such as network conditions, available bandwith and scene complexity.

Statistical analysis showed a small effect of the chosen DoF condition on the gathered scores for test T1. In general, the two visualization scenarios led to similar trends in MOS values; however, several participants pointed out that, while 3DoF offered a more stable assessment, as the same point of view is used for all contents, 6DoF felt more realistic. Any decision between the two viewing conditions for quality assessment, thus, should be made considering the trade-off between immersive, personalized experience, and fairness of comparison between solutions.

5 Discussion

5.1 Datasets

Despite the rich literature in point cloud acquisition and compression, few point cloud datasets are publicly available. This is especially true when considering point cloud datasets depicting photo-realistic humans. One of the most popular and widely used full-body dataset, created by 8i Labs [13], consists of only 4 individual contents, whereas the HHI Fraunhofer dataset has 1 individual content [14]. In the context of point cloud compression, such scarcity of available data may lead to compression solutions being designed, optimized and tested while considering a considerably narrow range of input data, thus leading to algorithms that are overfitted to the specifics of the acquisition method used to obtain the contents. The consequences of such a scenario are reflected in our results. Whereas for the contents assessed in test T2 a large difference was observed between codec V-PCC and the MPEG anchor, for the contents in test T1 the gap was markedly lower, and indeed the significance of

the effect of the codec selection had a smaller effect size for test T1 with respect to test T2, as seen in section 4.1. Test T2 consisted of contents that had been used in multiple quality assessment experiments [8, 9, 11, 36], notably including the performance evaluation of the upcoming MPEG standard [32]. On the other hand, test T1 included contents that have not been used so far in assessment of point cloud compression solutions. The discrepancies in the results of the subjective quality assessment campaign indicate that performance gains may vary considerably when new contents are evaluated. A larger body of contents depicting digital humans, involving several acquisition technologies, is needed in order to properly design, train and evaluate new compression solutions in a robust way.

5.2 Personal preferences and bias

Subjective evaluation experiments are complicated by many aspects of human psychology and viewing conditions, such as participants' vision ability, translation of quality perception into ranking scores, adaptations and personal preferences for contents. Through carefully following the ITU-T Recommendations P.913 [16], we are able to control some of the aspects. For example, eliminate the scores given by the participants with vision problems; train participants to help them understand the quality levels; randomize the stimuli and viewing conditions to minimize the order effects. However, we noticed that personal preferences towards certain contents are difficult to control. Satgunam et al. [29]) found that their participants were divided into two preference groups: prefer sharper content versus smoother content. Similarly, Kortum and Sullivan [19] found that the "desirability" of participants had an impact on video quality responses, with a more desirable video clip being given a higher rating. In our experiments, content *Queen* is generally given lower ratings with respect to the other contents. In the interviews, many participants (27%) expressed dislike towards *Queen*, because of her lifeless look and static gestures; 40% showed their preference towards *Soldier*, due to his high-resolution facial features, unitoned clothes and natural movements. This observation suggests that quality assessment may need to be adjusted based on content and viewer preferences, and offering training with different contents.

5.3 Technological constraints and limitations

The two codecs used in this experiment introduce different distortions during compression. As the MPEG anchor codec uses the octree data structure to represent geometry, the number of points in the decoded cloud varies exponentially based on the tree depth. Thus, at lower bitrates, the decoded point clouds are quite sparse, and when the point size is increased to make them appear watertight, they have a block-y appearance. This codec design allows for future optimizations based on human perception of 3D objects in VR. The low delay encoding and decoding of this codec makes it suitable for real time applications such as social VR. On the other hand, the V-PCC codec leverages existing 2D video codecs to compress both geometry and color, which introduces noise in terms of extraneous objects, and general geometric artifacts such as misaligned seams. However, the approach yields better results at low bitrates, as demonstrated in our results. The codec is optimized for human perception of 2D video and this might not transfer to perception of 3D objects in VR. The mapping from 3D to 2D is critical to codec performance, thus the encoding phase has high complexity. Decoding has a lower delay, as it benefits from hardware acceleration of video decoders on GPUs, making this approach suitable for on demand streaming.

One of the main shortcoming of both compression solutions lays in their inability to reach visually-lossless quality, as demonstrated by our results. Achieving a visually pleasant result is of paramount importance for the market adoption of the technology; indeed, poor visual quality might lead consumers to tune off from the experience altogether [1]. Visual perception should be taken into account when designing compression solutions, especially at high bitrates, to ensure that in absence of strict bandwidth constraints, excellent quality can be achieved.

5.4 Protocols for subjective assessment in VR

Choosing the right methodology to follow in order to collect users' opinions is a delicate matter, as it can influence the statistical power of the collected score, and in some cases lead to difference in results. Single stimulus methodologies, in particular, lead to larger CIs with respect to double stimulus methodologies, and are more subject to be influenced by individual content preference [16]. An early study comparing single and double stimulus methodologies for the evaluation of colorless point cloud contents indicated that the latter was more consistent in recognizing the level of impairment, as relative differences facilitate the rating task [4]. However, the study pointed out that the single stimulus methodology shows more discrimination power for compression-like artifacts, albeit at the cost of wider CIs.

Double stimulus methodologies, while commonly used in video quality assessment and widely adopted in 2D-based quality assessment of point cloud contents [8, 11, 32], are tricky to adopt in VR technology, due to the difficulties in displaying both contents simultaneously in a perceptually satisfying way [26], while ensuring a fair comparison between the contents under evaluation. When dealing with interactive methodologies, in particular, synchronous display of any modification in viewport is usually enforced, to ensure that the two contents are always visible at the same condition [8, 38]. This is clearly challenging to implement in a 6DoF scenario, in which users are free to change their position in the VR space at any given time. Positioning the two contents side by side in the same virtual space would mean that, at any given time, they are seen from two different angles; the same problem would arise when temporal sequencing is employed. A toggle-based method like the one proposed in [26] is not applicable to moving sequences, as different frames would be seen between stimuli.

In our study, we saw that content preference had an impact on the ratings, as several contents were deemed of lower quality, as the scores given to the HR exemplify. Such bias resulted in a reduced rating range for the contents. Results of the interviews also pointed out that naturalness of gestures were an important criteria in assessing the visual quality. Such components would not be normally evaluated in a double stimulus scenario; however, they are important in understanding how human perception reacts to digital humans.

6 Conclusion

We compare the performance of the point cloud compression standard V-PCC against an octree-based anchor codec (MPEG anchor). Participants were invited to assess the quality of digital humans represented as dynamic point clouds, in both 3DoF and 6DoF conditions. The results indicate that codec V-PCC has a more favorable performance than the MPEG anchor, especially at low bit-rates. For the highest bit-rate, the two codecs are often statistically equivalent. Results indicate that the content under test has a significant influence on how the scores are distributed; thus, new data sets are needed in order to comprehensively evaluate compression distortions. Moreover, current encoding solutions, while efficient at low bitrates, are unable to provide visually lossless results, even when large volumes of data are available, revealing significant shortcomings in point cloud compression. We also point out that commonly-used double stimulus methodologies for quality evaluation often reduce the rating task to a difference recognition, while insights on the quality of the original contents are missed.

Acknowledgments

This work is funded by the European Commission H2020 program, under the grant agreement 762111, *VRTogether*, http://vrtogether.eu/

REFERENCES

[1] OTT: Beyond Entertainment Consumer Survey Report. https://www.conviva.com/research/ott-beyond-entertainment/.

[2] D. S. Alexiadis, D. Zarpalas, and P. Daras. Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras. *IEEE Transactions on Multimedia*, 15(2):339–358, 2012.

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

[4] E. Alexiou and T. Ebrahimi. On the performance of metrics to predict quality in point cloud representations. In *Applications of Digital Image Processing XL*, vol. 10396, p. 103961H. International Society for Optics and Photonics, 2017.

[5] E. Alexiou and T. Ebrahimi. Impact of visualisation strategy for subjective quality assessment of point clouds. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2018.

[6] E. Alexiou and T. Ebrahimi. Point cloud quality assessment metric based on angular similarity. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2018.

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

[8] E. Alexiou, I. Viola, T. M. Borges, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi. A comprehensive study of the rate-distortion performance in mpeg point cloud compression. *APSIPA Transactions on Signal and Information Processing*, 8:27, 2019. doi: 10.1017/ATSIP.2019.20

[9] E. Alexiou, P. Xu, and T. Ebrahimi. Towards modelling of visual saliency in point clouds for immersive applications. In *26th IEEE International Conference on Image Processing (ICIP)*, 2019.

[10] J. Constine. Facebook animates photo-realistic avatars to mimic VR users' faces, 2018.

[11] L. A. da Silva Cruz, E. Dumić, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheiro, and T. Ebrahimi. Point cloud quality evaluation: Towards a definition for test conditions. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2019.

[12] R. L. de Queiroz and P. A. Chou. Motion-compensated compression of dynamic voxelized point clouds. *IEEE Transactions on Image Processing*, 26(8):3886–3895, 2017.

[13] E. d'Eon, B. Harrison, T. Myers, and P. A. Chou. 8i Voxelized Full Bodies - A Voxelized Point Cloud Dataset, ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006, Geneva. January 2017.

[14] T. Ebner, I. Feldmann, O. Schreer, P. Kauff, and T. v. Unger. HHI Point cloud dataset of a boxing trainer, ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document MPEG2018/m42921, Ljubljana. July 2018.

[15] ITU-T P.910. Subjective video quality assessment methods for multimedia applications. International Telecommunication Union, April 2008.

[16] ITU-T P.913. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. International Telecommunication Union, March 2016.

[17] M. Kay and J. Wobbrock. mjskay/artool: Artool 0.10.6, Feb. 2019. doi: 10.5281/zenodo.2556415

[18] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[19] P. Kortum and M. Sullivan. The effect of content desirability on subjective video quality ratings. *Human factors*, 52(1):105–118, 2010.

[20] S. Y. Liaw, G. A. C. Carpio, Y. Lau, S. C. Tan, W. S. Lim, and P. S. Goh. Multiuser virtual worlds in healthcare education: A systematic review. *Nurse education today*, 65:136–149, 2018.

[21] J.-L. Lugrin, M. Landeck, and M. E. Latoschik. Avatar embodiment realism and virtual fitness training. In *2015 IEEE Virtual Reality (VR)*, pp. 225–226. IEEE, 2015.

[22] K. Mammou. PCC test model category 2 v0. *ISO/IEC JTC1/SC29/WG11 N17248*, 1, 2017.

[23] 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*, 27(4):828–842, 2017.

[24] S. Narang, A. Best, A. Feng, S.-h. Kang, D. Manocha, and A. Shapiro. Motion recognition of self and others on realistic 3D avatars. *Computer Animation and Virtual Worlds*, 28(3-4):e1762, 2017.

[25] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 741–754. ACM, 2016.

[26] A.-F. Perrin, C. Bist, R. Cozot, and T. Ebrahimi. Measuring quality of omnidirectional high dynamic range content. In *Applications of Digital Image Processing XL*, vol. 10396, p. 1039613. International Society for Optics and Photonics, 2017.

[27] R. D. Queiroz and P. A. Chou. Compression of 3D Point Clouds Using a Region-Adaptive Hierarchical Transform. *IEEE Transactions on Image Processing 25*, June 2016.

[28] D. Roth, K. Waldow, M. E. Latoschik, A. Fuhrmann, and G. Bente. Socially immersive avatar-based communication. In *2017 IEEE Virtual Reality (VR)*, pp. 259–260. IEEE, 2017.

[29] P. N. Satgunam, R. L. Woods, P. M. Bronstad, and E. Peli. Factors affecting enhanced video quality preferences. *IEEE Transactions on Image Processing*, 22(12):5146–5157, 2013.

[30] O. Schreer, I. Feldmann, T. Ebner, S. Renault, C. Weissig, D. Tatzelt, and P. Kauff. Advanced volumetric capture and processing. *SMPTE Motion Imaging Journal*, 128(5):18–24, 2019.

[31] T. W. Schubert. The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15(2):69–71, 2003.

[32] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, et al. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2018.

[33] M. Seymour, K. Riemer, and J. Kay. Actors, avatars and agents: potentials and implications of natural face technology for the creation of realistic visual presence. *Journal of the Association for Information Systems*, 19(10):953–981, 2018.

[34] M. Slater and M. V. Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:74, 2016.

[35] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro. Geometric distortion metrics for point cloud compression. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3460–3464. IEEE, 2017.

[36] E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi. A novel methodology for quality assessment of voxelized point clouds. In *Applications of Digital Image Processing XLI*, vol. 10752, p. 107520I. International Society for Optics and Photonics, 2018.

[37] H. TT Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang. A subjective study on user perception aspects in virtual reality. *Applied Sciences*, 9(16):3384, 2019.

[38] I. Viola and T. Ebrahimi. A new framework for interactive quality assessment with application to light field coding. In *Applications of Digital Image Processing XL*, vol. 10396, p. 103961F. International Society for Optics and Photonics, 2017.

[39] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146. ACM, 2011.

[40] E. Zerman, P. Gao, C. Ozcinar, and A. Smolic. Subjective and objective quality assessment for volumetric video compression. In *Fast track article for IST International Symposium on Electronic Imaging 2019: Image Quality and System Performance XVI proceedings*, 2019.

[41] C. Zhang, D. Florencio, and C. Loop. Point cloud attribute compression with graph transform. *Image Processing (ICIP), 2014 IEEE*

International Conference on, October 2014.

[42] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang. A subjective quality evaluation for 3D point cloud models. In *2014 International Conference on Audio, Language and Image Processing*, pp. 827–831. IEEE, 2014.