2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)

Exploiting user interactivity in quality assessment of point cloud imaging

Evangelos Alexiou and Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
firstName.lastName@epfl.ch

***Abstract*—Point clouds are a new modality for representation of plenoptic content and a popular alternative to create immersive media. Despite recent progress in capture, display, storage, delivery and processing, the problem of a reliable approach to subjectively and objectively assess the quality of point clouds is still largely open. In this study, we extend the state of the art in projection-based objective quality assessment of point cloud imaging by investigating the impact of the number of viewpoints employed to assess the visual quality of a content, while discarding information that does not belong to the object under assessment, such as background color. Additionally, we propose assigning weights to the projected views based on interactivity information, obtained during subjective evaluation experiments. In the experiment that was conducted, human observers assessed a carefully selected collection of typical contents, subject to geometry and color degradations due to compression. The point cloud models were rendered using cubes as primitive elements with adaptive sizes based on local neighborhoods. Our results show that employing a larger number of projected views does not necessarily lead to better predictions of visual quality, while user interactivity information can improve the performance.**

***Index Terms*—Point clouds, objective quality assessment, subjective evaluation, user interactivity**

I. Introduction

Point clouds denote a popular approach for volumetric content representations, which are expected to dominate immersive communications in the near future. Yet, a vast amount of data is required in order to store and deliver such contents and, thus, efficient compression algorithms and interoperable formats are required. Compression methods come at the cost of information loss that typically leads to degradation of the visual quality of a model, which by extension can affect the user experience. Thus, it is of critical importance to define adequate frameworks to accurately assess impact of such distortions. For this purpose, subjective or objective quality assessment methodologies and metrics are used. With the first considered as the ground truth data and the second as algorithms that predict it.

In case of point clouds, several studies on subjective evaluation have been reported in the literature. In particular, in [1], [2] and [3] raw point clouds without color information were assessed in typical 2D monitors and augmented reality using a head-mounted display, respectively. The visual quality of colorless models was also evaluated in [4], where the screened Poisson surface reconstruction algorithm was used as a rendering mechanism. In [5], the perceived quality of dynamic, colored point clouds that represented the avatars of human test subjects were encoded in real-time and assessed in a 3D tele-immersive system. In [6], a subjective experiment was conducted, where observers evaluated in a passive way the visual quality of colored stimuli that were encoded using an octree- and a graph-based scheme. The point clouds were rendered using cubes of adaptive sizes based on local resolutions. In [7], a subjective evaluation campaign was performed to assess the quality of voxelized point clouds whose geometry and color were encoded using various configurations of the codec described in [5], in an interactive platform. In [8], subjective experiments of volumetric videos encoded using the MPEG Point Cloud Compression TMC 2 were issued. The contents were rendered using splats of fixed size, determined heuristically to result in visualization of watertight models, while the participants assessed the stimuli in a passive way.

Subjective evaluations, although providing ground truth information for the visual quality of stimuli, are expensive and cumbersome. Thus, objective means to faithfully predict human judgements are required. For point cloud representation, objective quality metrics can be categorized based on the type of information they are able to assess, namely, geometry, color, or geometry-plus-color. The first two categories depend on individual errors that are assigned to pairs of associated points. The error is based on deviations of the geometric positions [9] or normal vectors [10] for geometry, and color values for color metrics, respectively. These two categories are limited due to the inability of mutually assessing structural and textural degradations. Conversely, the framework that was recently proposed in [7] is able to capture both geometry and color distortions as well as rendering artifacts, by making use of conventional 2D imaging algorithms on projected views of the displayed models. In this approach, the point clouds are voxelized at a fixed depth; that is, the coordinates are quantized to regular 3D grids of fixed size, and each point is enclosed by a cubic volumetric cell (voxel) of corresponding color, or blend of colors in case several points fall in the same cell. To render a model, the obtained voxels are orthographically projected on 2D pixel grids. Benchmarking results on the resulting

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research project Advanced Visual Representation and Coding in Augmented and Virtual Reality (FN 178854).

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

views showed that this approach is superior to the state-of-the-art quality metrics. However, the number of projections to compute the objective scores was arbitrarily set to 6, while each view was treated as of equal importance. Furthermore, background pixels that do not belong to the displayed models were considered in the computations. Finally, this approach is not tested with different rendering algorithms other than voxelization with fixed depth.

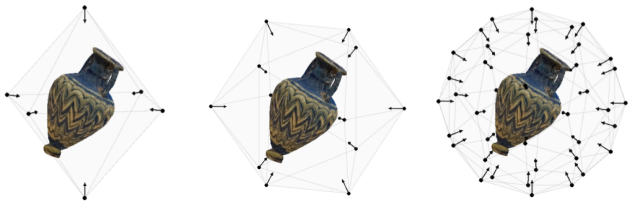
In this study, the aforementioned methodology is extended and importance weights are applied for the computation of the objective scores. In principle, different perspectives of a 3D model might be of different importance, as they could be more or less representative or informative regarding the presented content. Analogously, a non-uniform weighting function might be assigned on the views employed to predict the visual quality of a 3D model. To the best of our knowledge, weighted views have been considered only in [11] for objective evaluation of 3D meshes. The importance weights were obtained based on a surface visibility algorithm, typically used for viewpoint preference selection [12]. In our analysis, provided an interactive subjective evaluation scenario, we make the hypothesis that the importance of a view is related to the duration of inspection from participants during subjective assessment; thus, the projected images are weighted accordingly.

The contributions of this study can be summarized as follows: (a) Investigation on the impact of employing a different number of views in projection-based objective metrics. This is accomplished through a rigorous analysis by sub-sampling the view sphere in a regular way and exploring different configurations. (b) Introduction of a weighting function based on users' interactions. The displayed viewports are clustered to match the number of views under study, and larger weights are assigned to more frequently visited views. (c) Exploration of the generalization capabilities of projection-based metrics in a different rendering scheme, and performance improvement by removing background pixels from the computations.

II. Objective quality assessment framework

A. Generation of projected views

Given a distance, a 3D model can be inspected from an infinite number of points of the surrounding view sphere. Enabling a vast amount of viewpoints, though, is both impractical and unnecessary, as in a dense configuration two successive points provide very similar information. In our analysis, to address how many perspectives are sufficient and what is the impact of enabling additional views, a model is captured by K regularly-spaced viewpoints with the following camera layouts: (a) a single point that captures the frontal view of the content (i.e., $K = 1$), to examine whether a single image corresponding to the initial view of the model that was displayed to the subjects provides a good approximation of its visual quality; (b) the vertices of a surrounding octahedron (i.e., $K = 6$), which is idential to the setup of [7]; and (c) points lying on a surrounding geodesic sphere with coordinates determined by iterative subdivisions of a regular icosahedron up to 2 levels (i.e., $K = 12, 42, 162$). The latter is a commonly used arrangement in studies for view selection [11], [12], that provides a consistent approach to approximate uniformly distributed samples that are lying on the surface of a sphere. By iteratively subdividing the regular icosahedron, gradual granularity with progressive integration of new viewpoints on the previous set is achieved; this is important in order to identify whether additional views can improve the prediction of subjective visual quality. In Figure 1, indicative examples of the camera arrangements are illustrated.



(a) $K = 6$ (b) $K = 12$ (c) $K = 42$

Fig. 1: Camera layouts to capture views of the models.

Besides the number of viewpoints, additional influencing factors, such as the distance between the content and the cameras, the direction of the cameras, the lighting conditions, and the type of projection (e.g., orthographic, perspective) are fixed, in order to decrease the parameter space and simplify our analysis. In particular, the distance between the cameras and the model is fixed to match the one that was determined for the initial view presented to the subjects; thus, the model can be comfortably seen in its entirety from every point of the view sphere. The direction of every camera points at the center of the sphere (i.e., origin of the models), while the default lighting in the VTK library (i.e., headlight located at the current camera position) is used without any shading model. The stimuli are orthographically projected onto the renderer.

From each camera position, projections of the rendered models are acquired. The resolution of the captured images is 1100x1440, matching the resolution of the bitmaps that were displayed to the subjects during evaluation. The objective scores are computed based on these images. When assessing the state-of-the-art method [7], every pixel is involved in the computations. In the proposed framework, though, only the effective part of the rendered model is considered. For this reason, the images are captured in RGBA color space and, if needed, pixels that correspond to the background can be removed based on transparency values. An overall objective score is obtained by performing a (weighted) average of the scores associated to the projected views in each K-set.

B. Exploiting user interaction

Importance weights are assigned to each view of a model based on interactivity data that is recorded during a subjective evaluation experiment. In particular, the weights aim at reflecting the importance of a model's perspective in the final judgement of a subject. One way to compute such weights is based on the time of inspection of the corresponding views. In order to do so, it was decided to pre-filter the interactivity

information to reduce the noise. Let us define a track as a set of recorded interactions that corresponds to the inspection of a model by a subject. Firstly, a time threshold is applied on each track, in order to remove transitional views that were not carefully examined. In our case, the time threshold is set as one second. Secondly, interactivity data that corresponds to translations of the objects and, thus, different camera directions is excluded, as the translations are not considered in the camera layouts for the generation of views for objective scores. On the remaining data, each viewpoint of every track is mapped to the nearest view, given a camera arrangement. The total duration of inspection of a stimulus from one view can be obtained by aggregating the individual times of inspection from that particular view of the same stimulus across every subject. Similarly, the total duration of inspection of a content can be derived by combining the total duration of the stimuli that correspond to the content's variations (i.e., compressed versions). The weights of a stimulus or a content are computed as the ratio of the duration of inspection of the corresponding views, divided by the total time. In our case, weights per content are computed and applied on the projected views.

III. Validating experiment

A. Test content preparation

The content selection and preparation is identical to the methodology in [7]. Briefly, the dataset consists of 6 static contents clustered in two categories, namely, human bodies: *longdress_vox10_1300*, *loot_vox10_1200*, *redandblack_vox10_1550*, and inanimate objects: *amphoriskos*[1], *biplane*, *romanoillamp*. The majority of these models have been considered in recent activities of the JPEG and MPEG standardization committees. The reference models are compressed using a compression scheme that is described in [5] and is based on an octree structure to encode the geometry and the JPEG algorithm to encode the color values. Three octree depths (i.e., 8-, 9- and 10-bit) and three JPEG quality parameter values (i.e., QP = 10, 50, and 90) are applied to account for different levels of geometry and color quality, and all possible combinations are considered.

Before encoding, the reference models were pre-processed. Specifically, sub-sampling was performed whenever needed (i.e., *biplane*) in order to restrict the number of points of the contents in a small range of values. *Amphoriskos* was first converted to a polygon mesh to remove outliers and missing parts of the original model, and then was sampled. The resulting point clouds are voxelized using 3D grids of 10-bit depth. Finally, the models are scaled to fit in a minimum bounding box of size 1, and their center is placed in the position $(0, 0, 0)$. No rotations are applied.

B. Rendering scheme

To enable visualization of watertight models, the contents are rendered using primitive cubes of adaptive sizes based on local densities. In particular, the cube size for every point p is set analogously to the mean distance x of its 10 nearest neighbors. To avoid the magnification of sparse regions, or outlier points that deviate from surfaces (e.g., acquisition errors), we assume that x is a random variable following a Gaussian distribution $N(\mu_x, \sigma_x)$, and every point p with mean outside of a specified range, is classified as an outlier. In our case, this range is defined by the global mean $\mu = \bar{\mu}_x$ and standard deviation $\sigma = \bar{\sigma}_x$. For every point p, if $x \geq \mu + 3 \cdot \sigma$, or $x \leq \mu - 3 \cdot \sigma$, then p is considered as an outlier and x is set analogously to the global mean μ. Based on expert viewing, this approach was found to be rather efficient, with reduced complexity and visually pleasing results. Notice that this rendering methods is different from [7], where each point is represented as a projected voxel of fixed size onto a 2D plane.

(a) 8-bit octree (b) 9-bit octree (c) 10-bit octree

Fig. 2: Contents using our rendering methodology (QP = 90).

C. Test equipment and environment

The experiment was conducted in a test laboratory that follows the ITU-R Recommendation BT.500-13 [13]. The room is equipped with neon lamps of 6500 K color temperature, while the color of the walls and the curtains is mid gray. An Apple Cinema monitor of 27-inches and 2560x1440 resolution was used. The brightness was set to 120 cd/m^2 with a D65 white point profile. The conditions were adjusted and ambient light of 15 lux was measured next to the screen according to the ITU-R Recommendation BT.2022 [14]. The models were displayed in a renderer developed in the VTK library, allowing subjects to interact by rotation, translation and zooming, thus simulating realistic consumption of 3D models. The subjects' ratings were submitted by clicking through a graphical user interface developed in QT library. Special care was given to allow fast responsiveness in user's interactions and low waiting times in between content inspection.

D. Subjective evaluation methodology

The simultaneous Double-Stimulus Impairment Scale (DSIS) with 5-level grades is selected in this study. The reference and the degraded contents are displayed side-by-side, and the subjects are able to simultaneously interact with them before rating the visual quality of the latter, without any time limitation. The order of the stimuli is randomized and the side of the content under evaluation is randomly placed in the screen, per subject. A training phase preceded the

[1] https://sketchfab.com/

actual test, to allow the subjects to familiarize themselves with the renderer and obtain references for the types of artifacts. Considering that the interactivity information was logged for analysis purposes, it was considered that the occurrence of tiredness of the subjects would have an additional impact on the time they spend on every stimuli. Thus, the test was split in two sessions to avoid fatigue, limiting the expected total time to less than 10 minutes per session. The subjects, after giving every electronic device to the trainer, were instructed to avoid distractions and focus on their task to rate the distorted contents. In every test, for each of the 6 contents, 9 degradation levels are assessed, plus a hidden reference. Thus, each session is constituted by 30 stimuli. In total, 20 subjects participated in the experiment, with 10 males and 10 females (average age of 26.7 years).

E. Subjective quality metrics

To identify diverging scoring behavours from the subjects, the outlier detection scheme described in the ITU-R Recommendation BT.500-13 [13] is followed. As a result, no outlier was identified and, hence, the Mean Opinion Score (MOS) per stimuli is computed on a total of 20 scores. Additionally, the 95% confidence intervals (CIs) based on a Student's t-distribution are computed in order to identify the range in which the true mean lies. Finally, one-way ANOVA is issued to identify statistical differences among ratings of models that belong to different categories (i.e., inanimate objects, human bodies).

F. Objective quality metrics

The state-of-the-art objective quality metrics that assess geometry-only degradations can be grouped as: (i) point-to-point, (ii) point-to-plane [9], and (iii) plane-to-plane [10]. Each of these algorithms is based on the identification of pairs of associated points, and an error value for every point of the content under assessment is computed. This error is computed either based on the Euclidean distance (point-to-point), or the projected error along the normal vector of a reference point (point-to-plane), or the angular similarity (plane-to-plane). An objective score that reflects the quality level of the entire content under assessment is obtained by using either the Root Mean Square (RMS), or the Mean Squared Error (MSE), or the Hausdorff distance. In this study, both the original and the distorted contents are used as reference and both errors are computed; then, the maximum value is kept, which is referred as symmetric error. The point-to-point (po2point) and point-to-plane (po2plane) metrics are computed using ver. 0.12 of the software described in [15]. The plane-to-plane (pl2plane) metric is computed using the software released in [10]. The normal vectors of every content are estimated using 12 nearest neighbors, based on a plane-fitting algorithm [16].

The metrics that assess color-only information are based on standard formulas of 2D imaging between pairs of associated points. In this study, the PSNR is used after converting the default RGB color space to YCbCr, and the symmetric error is computed. The software described in [15] is employed to obtain the PSNR_Y, PSNR_U, and PSNR_V, and weights of 6, 1 and 1 are assigned to the luma and chroma channels, respectively [17].

For geometry-plus-color metrics, the 3D models are projected on image planes and the PSNR, SSIM, MS-SSIM [18], and VIFP [19] algorithms are applied on the captured views. To assess the state-of-the-art methodology, the same setup described in [7] is employed with 6 views, essentially projecting the model onto the faces of a surrounding cube. For the proposed framework, MATLAB scripts are modified for background removal[2].

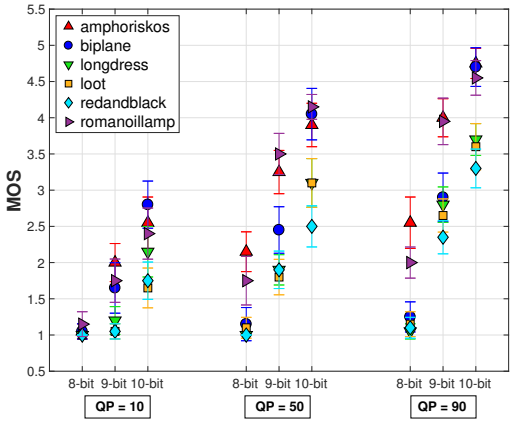


Fig. 3: Subjective scores against degradation levels.

G. Performance indexes

To benchmark the objective quality metrics against the ground truth subjective ratings, a predicted MOS is commonly obtained by applying a fitting function on the objective scores based on a regression model. The latter values are correlated with the subjective MOS. In our case, a monotonic cubic function is used for regression. Following the Recommendation ITU-T P.1401 [20], the Pearson linear correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), the root-mean-square error (RMSE), and the outlier ratio based on standard error (OR) are computed between the subjective and predicted MOS values, to assess the linearity, monotonicity, accuracy and consistency of the results.

IV. Results

A. Subjective scores

In Figure 3, the computed MOS along with the corresponding CIs are presented against the geometry and color degradation levels. As can be observed, geometry artifacts limit the visual quality of the compressed models, while color distortions are rated less severely. Moreover, different rating behaviors can be remarked for the two types of content, namely, human bodies and inanimate objects. In particular, distortions on the human bodies dataset are rated more critically, leading to lower scores. A one-way ANOVA applied on the ratings grouped per type of content, shows that the

[2]http://live.ece.utexas.edu/research/Quality/index_algorithms.htm

subjective scores are statistically significantly different (p-value $= 1.40e-29$), which is in alignment with [7].

B. Benchmarking of state-of-the-art

The subjective scores were found to be statistically different per type of content, thus, benchmarking analysis is applied on the contents that represent inanimate objects and human bodies, individually. In Table I our findings are reported. The best-performing metric for a particular index is indicated in bold. As can be seen, the MS-SSIM and plane-to-plane metrics found to be better in the inanimate objects and human bodies datasets, respectively. Statistical differences using Fisher z-scores according to the Recommendation ITU-T P.1401 [20] suggest that the MS-SSIM is statistically better than point-to-point and point-to-plane metrics based on the OR index, and point-to-plane with MSE, plane-to-plane metrics and VIFP, based on the RMSE index.

TABLE I: Performance indexes for state-of-the-art metrics.

	Inanimate objects				**Human bodies**			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
$\text{po2point}_{\text{MSE}}$	0.740	0.769	0.812	0.889	0.732	0.789	0.621	0.778
$\text{po2point}_{\text{Hausdorff}}$	0.735	0.758	0.819	0.889	0.732	0.781	0.621	0.778
$\text{po2plane}_{\text{MSE}}$	0.692	0.684	0.872	0.889	0.717	0.762	0.636	**0.741**
$\text{po2plane}_{\text{Hausdorff}}$	0.732	0.701	0.824	0.889	0.734	0.788	0.620	0.778
$\text{pl2plane}_{\text{RMS}}$	0.668	0.723	0.900	0.778	**0.782**	**0.813**	**0.568**	**0.741**
$\text{pl2plane}_{\text{MSE}}$	0.664	0.723	0.903	0.815	**0.782**	**0.813**	**0.568**	**0.741**
Color - PSNR_{YUV}	0.791	0.751	0.739	0.778	0.668	0.618	0.678	**0.741**
PSNR	0.739	0.672	0.814	0.704	0.740	0.771	0.613	0.815
SSIM	0.823	0.817	0.686	0.741	0.619	0.600	0.716	0.889
MS-SSIM	**0.884**	**0.855**	**0.566**	**0.630**	0.727	0.757	0.626	0.852
VIFP	0.693	0.645	0.871	0.778	0.662	0.566	0.683	0.778

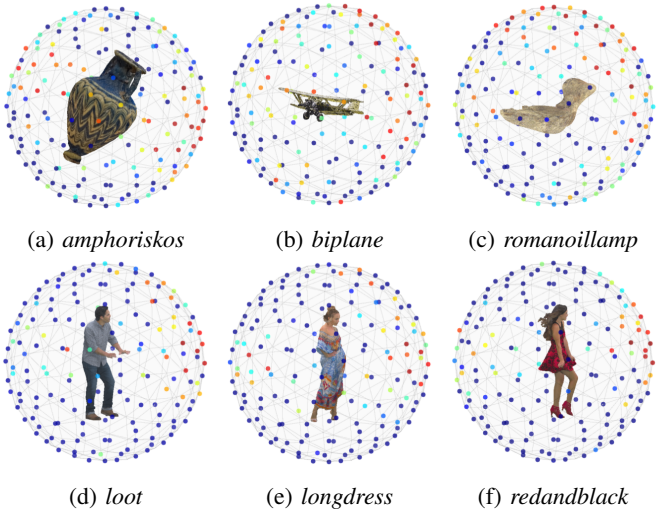
C. Benchmarking using the proposed framework

As described in Section II-A, different camera layouts are examined leading to a different number of K views, while the same analysis is repeated by incorporating importance weights based on the logged interactivity information. The performance indexes of the above test cases are reported in Tables II and III for inanimate objects and human bodies, respectively, with the first row indicating the results by excluding (avg), or including (w. avg) user interactivity.

Regarding the results for inanimate objects without using interactivity data, it is observed that as the number of views is increasing, the PCC index of the PSNR and MS-SSIM metrics remain approximately the same. The performance of the SSIM drops while the prediction power of the VIFP increases, with a minimum at $K = 12$ in both cases. For human bodies, the PCC index of SSIM, MS-SSIM and VIFP is decreasing by increasing number of viewpoints, while for PSNR it gradually improves, after reaching the minimum at $K = 6$.

By incorporating importance weights, it is evident that equal and consistently better results are obtained for inanimate objects and human bodies, respectively. Moreover, in both datasets, as the number of viewpoints is increasing, the PCC of the MS-SSIM metric remains high and stable.

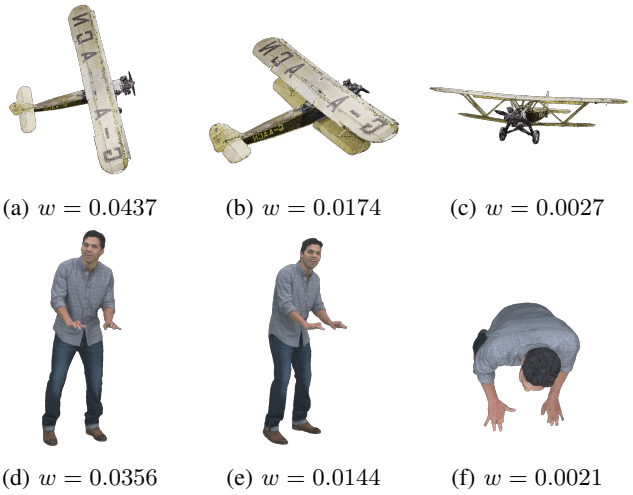
Overall, the MS-SSIM is the best-performing metric. By slight margins, the optimal layout for inanimate objects is observed when using $K = 6$ views including interactivity information, while for human bodies, the best performance is achieved by using only the frontal view. In the latter case, the performance notably worsens by excluding interactivity data, as the number of views is increasing.



(a) *amphoriskos* (b) *biplane* (c) *romanoillamp*

(d) *loot* (e) *longdress* (f) *redandblack*

Fig. 4: Dot markers on the view sphere correspond to camera positions for a 2-level subdivision of an icosahedron ($K = 162$). The color code represents the ranking of weights, ranging from dark blue (minimum) to dark red (maximum).



(a) $w = 0.0437$ (b) $w = 0.0174$ (c) $w = 0.0027$

(d) $w = 0.0356$ (e) $w = 0.0144$ (f) $w = 0.0021$

Fig. 5: Views of models with different importance weights.

In Figure 4, the importance weights associated with every view on the camera layout with $K = 162$ are presented for every model. For contents that represent inanimate objects, subjects tend to allocate more time on views that are more informative, as indicated in Figures 5a-5c. For instance, high weights are observed at viewpoints that are located on top of the *biplane* and the *romanoillamp* contents, while high weights are assigned around the equator of *amphoriskos* which is a rather symmetric model. For models that represent human bodies, users consistently spend more time in frontal views, as can be seen in Figures 5d-5f. This outcome is in accordance with [21], where subjects were explicitly asked to select the best view of a wide range of 3D models, in which a clear

TABLE II: Performance indexes for proposed framework on Inanimate objects data set.

		Frontal view (K = 1)				Octahedron (K = 6)				Icosahedron (K = 12)				1-subdiv. icosahedron (K = 42)				2-subdiv. icosahedron (K = 162)			
		PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
avg	PSNR	0.623	0.601	0.945	0.889	0.622	0.604	0.946	0.889	0.621	0.600	0.947	0.889	0.623	0.601	0.945	0.889	0.623	0.601	0.945	0.889
	SSIM	0.921	0.906	0.471	0.630	0.908	0.898	0.505	0.630	0.893	0.890	0.545	0.704	0.897	0.889	0.534	0.630	0.897	0.889	0.535	0.630
	MS-SSIM	**0.955**	**0.944**	**0.360**	**0.481**	0.957	**0.944**	0.350	0.481	0.953	0.944	0.366	**0.444**	0.953	0.945	0.365	**0.481**	0.953	**0.945**	0.365	**0.481**
	VIFP	0.914	0.903	0.491	0.556	0.931	0.924	0.442	0.593	0.927	0.925	0.452	0.481	0.930	0.922	0.444	0.481	0.930	0.925	0.444	0.481
w. avg	PSNR	0.623	0.601	0.945	0.889	0.625	0.600	0.943	0.889	0.635	0.607	0.933	0.852	0.620	0.597	0.948	0.889	0.627	0.606	0.942	0.889
	SSIM	0.921	0.906	0.471	0.630	0.917	0.904	0.482	0.593	0.889	0.889	0.552	0.667	0.910	0.899	0.501	0.630	0.893	0.889	0.543	0.593
	MS-SSIM	**0.955**	**0.944**	**0.360**	**0.481**	**0.958**	**0.944**	**0.346**	**0.444**	**0.956**	**0.949**	**0.354**	**0.444**	**0.955**	**0.947**	**0.359**	**0.481**	**0.954**	0.942	**0.364**	**0.481**
	MS-SSIM	0.914	0.903	0.491	0.556	0.925	0.927	0.460	0.519	0.931	0.923	0.442	0.519	0.926	0.919	0.455	0.519	0.927	0.927	0.454	0.556

TABLE III: Performance indexes for proposed framework on Human bodies data set.

		Frontal view (K = 1)				Octahedron (K = 6)				Icosahedron (K = 12)				1-subdiv. icosahedron (K = 42)				2-subdiv. icosahedron (K = 162)			
		PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
avg	PSNR	0.788	0.809	0.561	0.741	0.715	0.723	0.638	0.778	0.730	0.748	0.623	0.815	0.735	0.772	0.618	0.778	0.736	0.780	0.618	0.778
	SSIM	0.889	0.859	0.418	0.741	0.834	0.788	0.503	0.704	0.828	0.759	0.511	0.704	0.828	0.769	0.512	0.704	0.827	0.769	0.513	0.704
	MS-SSIM	**0.953**	**0.935**	**0.277**	**0.519**	0.937	0.927	0.319	**0.556**	0.930	0.920	0.334	0.593	0.930	0.915	0.336	0.556	0.929	0.915	0.337	0.556
	VIFP	0.938	0.927	0.317	0.667	0.925	0.921	0.347	0.519	0.924	0.919	0.348	0.556	0.922	0.906	0.352	0.556	0.921	0.906	0.354	0.556
w. avg	PSNR	0.788	0.809	0.561	0.741	0.774	0.799	0.577	0.741	0.753	0.792	0.600	0.778	0.784	0.815	0.566	0.778	0.770	0.805	0.581	0.778
	SSIM	0.889	0.859	0.418	0.741	0.880	0.853	0.433	0.704	0.877	0.850	0.437	0.704	0.883	0.857	0.427	0.704	0.880	0.857	0.433	0.704
	MS-SSIM	**0.953**	**0.935**	**0.277**	**0.519**	**0.950**	**0.935**	**0.286**	**0.556**	**0.946**	**0.935**	**0.297**	**0.519**	**0.950**	**0.933**	**0.286**	**0.519**	**0.949**	**0.936**	**0.287**	**0.519**
	VIFP	0.938	0.927	0.317	0.667	0.936	0.925	0.322	0.593	0.923	0.918	0.350	0.593	0.936	0.927	0.321	0.593	0.928	0.920	0.339	0.593

preference for frontal views in human bodies and faces is reported. This may explain why for human bodies dataset the frontal view found to be the best configuration, while for the inanimate objects a higher number of perspectives is needed.

V. Conclusions

In this paper, the state-of-the-art in projection-based objective quality assessment of point clouds is extended by investigating the impact of applying different camera layouts to capture views of the models, as well as exploiting user interactivity data. Our results suggest that, independently of the type of content, even one view could be enough to achieve high performance. It is also shown that the interactivity information from subjects assessing the contents can be beneficial, as the prediction power of the objective quality metrics is improved, especially in the case of models representing human bodies.

References

[1] 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)*, May 2017, pp. 1–3.

[2] ——, "On the performance of metrics to predict quality in point cloud representations," in *Proceedings of SPIE*, ser. Applications of Digital Image Processing XL, vol. 103961H, Sep 2017.

[3] 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)*, Oct 2017, pp. 1–6.

[4] E. Alexiou, T. Ebrahimi, M. V. Bernardo, M. Pereira, A. Pinheiro, L. A. Da Silva Cruz, C. Duarte, L. G. Dmitrovic, E. Dumic, D. Matkovics, and A. Skodras, "Point Cloud Subjective Evaluation Methodology based on 2D Rendering," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6.

[5] R. Mekuria, K. Blom, and P. Cesar, "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, April 2017.

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

[7] 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 *Proceedings of SPIE*, ser. Applications of Digital Image Processing XLI, vol. 107520I, Sep 2018.

[8] E. Zerman, P. Gao, C. Ozcinar, and A. Smolic, "Subjective and Objective Quality Assessment for Volumetric Video Compression," in *IS&T Electronic Imaging, Image Quality and System Performance XVI*, 2019.

[9] 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)*, Sep 2017, pp. 3460–3464.

[10] E. Alexiou and T. Ebrahimi, "Point Cloud Quality Assessment Metric Based on Angular Similarity," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, July 2018, pp. 1–6.

[11] G. Lavoué, M. C. Larabi, and L. Váša, "On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 8, pp. 1987–1999, Aug 2016.

[12] X. Bonaventura, M. Feixas, M. Sbert, L. Chuang, and C. Wallraven, "A Survey of Viewpoint Selection Methods for Polygonal Models," *Entropy*, vol. 20, no. 5, 2018.

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

[14] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," International Telecommunications Union, Aug 2012.

[15] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Updates and Integration of Evaluation Metric Software for PCC," ISO/IEC JTC1/SC29/WG11 input document MPEG2017/M40522, Hobart, Australia, April 2017.

[16] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface Reconstruction from Unorganized Points," in *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '92. ACM, July 1992, pp. 71–78.

[17] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards-Including High Efficiency Video Coding (HEVC)," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 22, no. 12, pp. 1669–1684, 2012.

[18] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, Nov 2003, pp. 1398–1402 Vol.2.

[19] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb 2006.

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

[21] H. Dutagaci, C. P. Cheung, and A. Godil, "A Benchmark for Best View Selection of 3D Objects," in *Proceedings of the ACM Workshop on 3D Object Retrieval*, ser. 3DOR '10. New York, NY, USA: ACM, 2010, pp. 45–50.