2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)

2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3503-2838-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR59233.2023.00021

QAVA-DPC: Eye-Tracking Based Quality Assessment and Visual Attention Dataset for Dynamic Point Cloud in 6 DoF

Xuemei Zhou*
Centrum Wiskunde & Informatica
TU Delft

Irene Viola†
Centrum Wiskunde & Informatica

Evangelos Alexiou‡
TNO Netherlands Organisation for
Applied Scientific Research

Jack Jansen§
Centrum Wiskunde & Informatica

Pablo Cesar¶
Centrum Wiskunde & Informatica
TU Delft

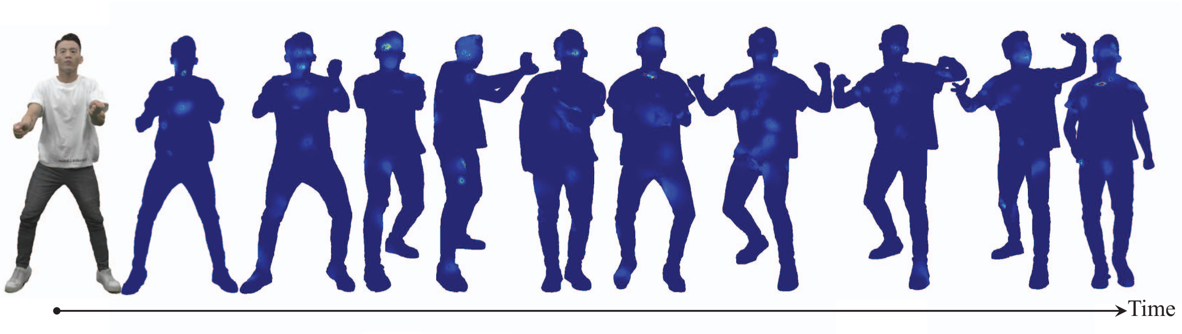


Figure 1: Fixation maps of *dancer* sequences with uniform temporal sampling every 30 frames.

Abstract

Perceptual quality assessment of Dynamic Point Cloud (DPC) contents plays an important role in various Virtual Reality (VR) applications that involve human beings as the end user, understanding and modeling perceptual quality assessment is greatly enriched by insights from visual attention. However, incorporating aspects of visual attention in DPC quality models is largely unexplored, as ground-truth visual attention data is scarcely available. This paper presents a dataset containing subjective opinion scores and visual attention maps of DPCs, collected in a VR environment using eye-tracking technology. The data was collected during a subjective quality assessment experiment, in which subjects were instructed to watch and rate DPCs at various degradation levels under 6 degrees-of-freedom inspection, using a head-mounted display. The dataset comprises 5 reference DPC contents, with each reference encoded at 3 distortion levels using 3 different codecs, amounting to a total of 9 degraded DPC contents. Moreover, it includes 1,000 gaze trials from 40 participants, resulting in 15,000 visual attention maps in total. The curated dataset can serve as authentic benchmark data for assessing the performance of objective DPC quality metrics. Additionally, it establishes a link between quality assessment and visual attention within the context of DPC. This work deepens our understanding of DPC quality and visual attention, driving progress in the realm of VR experiences and perception.

Index Terms: Volumetric video, Dynamic point cloud, Visual saliency, Visual attention, Subjective quality assessment, Objective quality metrics, Eye tracking, 6DoF

1 Introduction

Volumetric video has become available for representing real-world objects, due to the rapid development of capture devices, transmission technologies, and computational capabilities. Point cloud has emerged as one of the most popular formats for volumetric video representation. Specifically, Dynamic Point Cloud (DPC) can be used for automotive/robotic navigation [50], medical imaging [5], virtual video conferencing [20, 42], among others. A point cloud can be defined as a set of points in space represented in a 3D coordinate system. A DPC is essentially a sequence of individual point cloud frames played in succession. However, each point cloud frame requires a large number of points to faithfully represent the content and achieve a good Quality-of-Experience (QoE). Therefore, effective compression is essential before transmission, storage, rendering, and display. Quality degradation will be inevitably introduced during this end-to-end pipeline, which deteriorates the visual quality and affects the perception. Exploring the distortion characteristics of DPC and effectively measuring them in 6 Degrees-of-Freedom (DoF) is a challenge in both subjective and objective quality assessment [1].

Subjective quality assessment leads to ground-truth ratings for visual impairments that appear in a stimulus. Subjective quality assessment for DPC has been explored in desktop viewing conditions [46, 47] or in immersive environments with users consuming the contents through a Head-Mounted Display (HMD) under 6DoF [38, 43]. In the latter case, information about users' movement can be captured in addition to subjective quality ratings, to understand how users navigate and observe objects in VR space. A more accurate representation of the user's consumption is given by gaze data, which highlights the specific areas of content being viewed with focused attention. This information aids in the creation of visual attention maps. Incorporating visual attention into quality assessment has demonstrated potential improvement for predicting the visual quality of 2D/3D image/video [23, 49]. Nonetheless, visual attention for DPC is still in its infancy, thus hindering the utilization of its outcomes in aiding visual quality assessment.

A summary of existing subjective quality assessment and visual attention datasets for point clouds is shown in Table 1. Most of the studies in the literature involving DPCs are conducted with 2D

*E-mail: xuemei.zhou@cwi.nl
†E-mail: irene.viola@cwi.nl
‡E-mail: evangelos.alexiou@tno.nl. The work was partially conducted while the author was at CWI.
§E-mail: jack.jansen@cwi.nl
¶E-mail: p.s.cesar@cwi.nl

979-8-3503-2838-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ISMAR59233.2023.00021

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

Table 1: Publicly available subjective quality assessment and visual attention datasets for point clouds.

Dataset	Type	Degradation	Stimuli	Time	Display	Interaction	Opinion Score	Visual Attention
VsenseVVDB [46]	Dynamic	down-sampling, VPCC	32	6.6s	2D monitor	✗	✓	✗
VsenseVVDB2 [47]	Dynamic	Mesh: Draco+JPEG Point Clouds: GPCC, VPCC	28 136	10s	2D monitor	✗	✓	✗
Owlii [45]	Dynamic	Mesh: TFAN, FFmpeg Point Clouds: VPCC, FFmpeg	20	20s	2D monitor	✗	✓	✗
Subramanyam et al. [38]	Dynamic	CWI-PCL, VPCC	72	5s	HMD	✓	✓	✗
ViAtPCVR [3]	Static	Only reference	8	-	HMD	✓	✗	✓
QAVA-DPC(Ours)	Dynamic	VPCC, GPCC, CWI-PCL	50	10s	HMD	✓	✓	✓

monitors: the DPCs are pre-recorded and playback to the user using conventional video software [45–47]. However, the passive nature of display restricts user freedom, as DPCs can only be presented according to a predetermined trajectory. On the other hand, an immersive HMD-based display with 6DoF allows for a complete representation of the entire DPC, but typically involves a smaller number of DPCs (20), usually static or with shorter time duration (5 seconds), due to technical limitations that prevent a smooth playback in real time [38]. Due to such constraints, no visual attention dataset specifically designed for DPC has been released so far; existing research has primarily explored the attention of static point clouds [3], confining the scope to a few undistorted contents. There is currently a lack of studies that connect visual attention and visual quality specifically for DPC.

Visual attention of point cloud can benefit a myriad of vision tasks, such as segmentation, localization, and registration [13]. Improvement has been reported by using visual attention maps to weight quality maps for perceptual quality prediction [25]. By connecting visual attention and visual quality for DPC, the quality allocation between the salient region and the remaining area, saliency-aware compression and streaming, and saliency-aided objective quality metrics can be further investigated and optimized.

In this paper, we aim to create an eye-tracking-based Quality Assessment and Visual Attention dataset for DPCs (QAVA-DPC), which consists of diverse contents and encompasses various types of distortions. The associated visual attention maps can thereby enhance the understanding of human behavior within 6DoF environments, ultimately contributing to the optimization of QoE. Our contributions can be summarized as follows:

- We propose a new dataset, namely, QAVA-DPC, which contains 5 reference DPCs; each DPC is encoded by 3 codecs, with each codec configured at 3 distortion levels. Fixation maps are constructed, collected, and presented for both the reference and distorted sequences as heatmaps overlaid on top of the stimuli frames. To the best of our knowledge, this is the first time connecting visual attention and visual quality for DPC in VR.

- We release all raw data, containing the opinion scores and gaze samples collected in our study, alongside the software used to perform the experiment, and the scripts used to to export visual attention maps, at the following link: `https://github.com/cwi-dis/ISMAR_PointCloud_EyeTracking`.

2 Related Work

2.1 Eye Tracking Experiment for 3D contents

Owing to the human vision system's selectivity in responding to the most attractive features in the visual field, it's inappropriate to treat each voxel equally [23]. To explore visual attention for 3D contents, eye-tracking experiments remain the main way to understand human visual behavior. Sitzmann *et al.* [35] capture and analyze gaze and head orientation data of users exploring stereoscopic, static omnidirectional panoramas, for a total of 1,980 head and gaze trajectories for three different viewing conditions. They found the existence of a particular fixation bias, which can be used to adapt existing saliency predictors to immersive VR conditions. Nguyen *et al.* [27] introduce a large saliency dataset for 360-degree videos with a new methodology supported by psychology studies with HMD. They describe an open-source software implementing this methodology that can generate saliency maps from any head tracking data. Lavoué *et al.* [21] present a dataset that records the eye-movement data for rendered 3D shapes. During their experiment, 3D meshes are rendered using different materials and lighting conditions under different scenes, and the rendered videos of 3D meshes are shown on the screen for subjects to observe. Ding *et al.* [8] propose a novel 6DoF mesh saliency dataset that provides both the subject's 6DoF data and eye-movement data, and a 6DoF mesh saliency detection algorithm based on the uniqueness measure and the bias preference is developed. Alexiou *et al.* [3] conduct an eye-tracking experiment in an immersive 3D scene that offers 6DoF. A method to exploit the high-quality recorded gaze measurements is introduced based on per-session profiling, and a scheme to determine areas of fixations in a point cloud is proposed.

To the best of our knowledge, no dataset has been yet released for visual attention of DPCs, which is our main contribution.

2.2 DPC Quality Assessment

Whereas subjective and objective quality assessment of static point clouds has been explored in more detail in the literature [2], analogous research on DPCs is still a sophisticated and challenging problem, owing to numerous factors such as the evaluation methodology, rendering method, display equipment and so forth. Subjective quality scores, such as Mean Opinion Score (MOS) or Differential MOS (DMOS), are commonly used to quantify the subjective perception of visual artifacts. Zerman *et al.* [46] conduct a subjective experiment on two DPCs (VsenseVVDB) using MPEG VPCC compression [33]. The same author compares the mesh and point cloud representation formats (VsenseVVDB2) for a volumetric video compression scenario utilizing state-of-the-art compression techniques. The results show that meshes provide the best quality at high bitrates, while point clouds perform better for low-bitrate cases [47]. Hooft *et al.* investigate how and to what extent various aspects have more impact on the user's QoE, via extensive objective and subjective evaluation of volumetric 6DoF streaming [40]. Mekuria *et al.* evaluate the subjective quality of the CWI-PCL codec performance in a realistic 3D tele-immersive system in a virtual 3D room scenario, in which users are represented and interact as 3D avatars and/or 3D point clouds [26]. The results show that the degradation introduced by CWI-PCL is negligible. However, these experiments are all with a desktop setting in a passive manner. Viola *et al.* compare two different VR viewing conditions enabling 3/6 DoF, along with a desktop setting, to understand how interaction in the virtual space affects the perception of quality [43]. Results show no statistical difference between scores given in a desktop and VR setup; however, qualitative results highlighted the added value of interactive evaluation. One main limitation of the study lies in the length of the sequences used for the evaluation, as the authors use 150 frames for their study. Subramanyam *et al.* [39] evaluate the performance of several adaptive streaming solutions in an interactive VR experiment.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

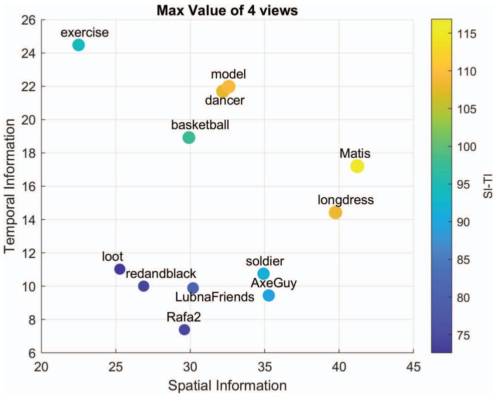


Figure 2: Distribution of SI and TI of 12 source DPCs from 3 datasets, the color value is computed by $\sqrt{(SI^2+TI^2)}$

.

In their setup, they compare the performance of MPEG VPCC with respect to CWI-PCL, using various adaptive streaming strategies.

In our study, we aim to complement existing literature by performing an experiment comparing the visual quality of several state-of-the-art compression techniques for DPC. We do so in an interactive manner, using an HMD-based VR rendering of 10s sequences from various datasets, which has not been done before in the literature in combination with eye-tracking.

2.3 Eye-Tracking based Objective Quality Assessment

Recent literature in eye-tracking-based visual saliency for immersive contents has mainly focused on task-free experiments to gather visual attention maps [29, 36]; no study has been conducted to link visual attention to visual quality assessment for volumetric contents. The literature suggests that visual attention might be beneficial for understanding the process of perception of visual quality for 2D images/videos; in fact, different metrics for Image Quality Assessment (IQA) have been extended with a computational model of visual attention [23], but the resulting gain on the metrics' performance is so far unclear. To better understand the added value of including visual attention in the design of objective metrics for 2D images, some works in the literature have taken advantage of recorded visual attention data. Lin *et al.* [24] perform two eye-tracking experiments: one with a free-looking task and one with a quality assessment task. They found a tendency that adding saliency to a metric yields a larger amount of gain in performance. The extent of the performance gain tends to depend on the specific objective metric and the image content. In addition, the gain is small for objective metrics that already show a high correlation with perceived quality for a given distortion type. Zhang *et al.* [49] propose a new methodology to eliminate the inherent bias due to the involvement of stimulus repetition. The refined methodology result in a new eye-tracking dataset with a large degree of stimulus variability. Based on ground-truth labeling, the statistical evaluation shows that visual attention of both the referenced and distorted scene is beneficial for IQA metrics, but the latter tends to further boost the effectiveness of the integration of attention in IQA metrics. Jin *et al.* [18] utilize the eye-tracker to create foveation-compressed VR datasets and evaluate both the foveated and non-foveated objective image/video quality assessment algorithms.

To better understand whether the findings regarding visual saliency and quality assessment on 2D images/videos can hold for volumetric contents, ad-hoc datasets that combine the two aspects are needed. That is the research gap we aim to fill with this paper.

3 QAVA-DPC Construction

3.1 Stimuli Selection

For the creation of the dataset, we selected 5 DPCs from 3 public datasets, namely VsenseVVDB2 [47], 8i [9] and Owlii [45]. To show the diversity of DPCs, we considered the Spatial Information (SI) and Temporal Information (TI) for each content [15]. We projected the source point cloud into 4 views, which are the left, right, front, and back view, of its bounding box to apply SI and TI separately, then obtain the maximum value among the 4 views over all the first 300 frames as the final SI/TI for one sequence. The distribution of all DPCs can be seen in Fig.2, we finally choose *dancer*, *exercise*, *long dress*, *rafa2*, and *soldier* as the contents in our dataset. The dispersed state in SI (horizontal axes)/TI (vertical axes) shows the diversity of our contents in the spatial/temporal domain.

3.2 Stimuli Processing

Before conducting the subjective experiment on DPC, specific procedures are necessary due to the codec implementations. These procedures, including pre-processing, encoding, and rendering, are aimed at minimizing additional influencing factors.

3.2.1 Pre-processing

The sequences mentioned above are selected from different datasets, which means the resolution, position and orientations vary. The DPCs should be life-size so to create a realistic tele-immersive scenario. To do so, we normalize the DPCs to a similar bounding box. The geometry precision of *dancer* and *exercise* is voxelized from 11 to 10. The source models are processed with rotation, translation, and scaling. Additionally, since the VPCC encoder fails to deal with decimals, the coordinates of DPCs are rounded before VPCC compression. CWI-PCL encoder has specific requirements for the resolution of DPCs, so before CWI-PCL compression the coordinates go through the scaling operation.

3.2.2 Encoding

Distorted versions are generated using the state-of-the-art MPEG PCC reference software Test Model Category 2 Version 18 ($TMC2$V-18.0) and Category 1&3 Version 14 (TMC_{13}V-14.0) from now on referred to as VPCC and GPCC [33]. We also adopt the CWI-PCL [26] codec as a comparison. GPCC is proposed mainly for the aim of compressing static point cloud, VPCC is developed for DPC compression, and CWI-PCL is mainly used to comply with real-time requirements. To compare them in a fair way, we set the GPCC encoder with Region-Aptive Hierarchical Transform (RAHT) to compress point-wise color attributes and Octree for geometry representation; the VPCC encoder with All Intra (AI) mode, which adapts intra-prediction for one frame; and the CWI-PCL intra frame, geometry coded with octree subdivision and color attributes encoded based on JPEG.

To define the configuration parameters for the encoders, the MPEG Common Test Conditions [37] are followed. To compare different codecs and different distortion levels, we select the distortion levels that can reveal a similar low-medium-high quality range among the 3 codecs. Specifically, for GPCC we select three distortion levels, namely R02, R04, and R05, by setting *positionQuantizationScale* and QP parameters. For VPCC, we select three distortion levels, namely R01, R03, and R05 by setting different geometry QP, attribute QP, and *occupancyPrecision* parameters. For CWI-PCL, we choose three combinations of octree depth with JPEG QP parameters to match a similar quality range, by looping over octree depth from 7 to 9 and JPEG QP from 25 to 95 (step size = 10). When testing on the dataset, the above parameter settings for the three codecs yielded subjectively similarly from the perspective of the quality range. Specific parameter settings are shown in Table 2. Each DPC has 3 compression codecs, and each codec has 3 distortion levels, for a total of 45 distorted DPCs.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

Table 2: Parameter sets for the selected encoders

Encoders	Distortion Level		
GPCC (Octree-RAHT)	R02 (0.125, 46)	R04 (0.5, 34)	R05 (0.75, 28)
VPCC (AI)	R01 (32,42, 4)	R03 (24, 32, 4)	R05 (16, 22, 2)
CWI-PCL	R01 (7, 25)	R02 (8, 95)	R03 (9, 95)

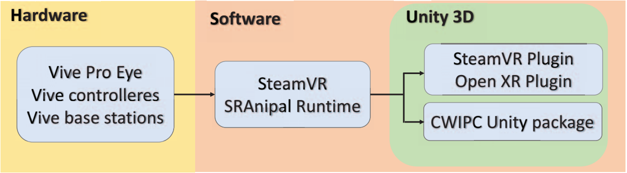


Figure 3: Schematic diagram with the hardware and software modules together with their inter-dependencies.

3.2.3 Rendering

Rendering is the process of producing a visual representation that can be consumed by users using an available display. In the case of point clouds, different rendering methods have a significant impact on perceived quality [17]. In our experiment, we choose to render the point clouds without any additional processing (e.g., surface reconstruction), directly using the point cloud data (point-based).

Our experiment software is developed in Unity (version 2021.3.10.f1), exploiting the SteamVR plugin (version 1.24.7) to connect with VR headsets and controllers. CWI Point Cloud (CWIPC) supported unity package (version 0.10.0) helps us import the DPCs and playback them inside Unity [30]. A high-level diagram indicating the hardware/software dependencies is provided in Fig.3. Notably, a large size of DPC file might take up too much memory and cause a system hang. So we first transform the DPC data to CWIPC-supported point cloud playback format to improve the software stability. To ensure smooth playback of DPC, we take advantage of the Unity Coroutine scheme to preload each DPC into memory before the user switches to next DPC. 5 DPCs with their corresponding operation are selected in our test. It should be noted for each sequence, we only choose the first 300 frames from the source model. The frame rate for playback is 30 frames per second, hence each video lasts for 10 seconds. We use HTC Vive Pro Eye devices with eye-tracking capabilities and Vive hand controllers for participants to interact in our experiment. To develop eye-tracking applications for the Vive Pro Eye we use the native HTC Vive SRanipal SDK. The sampling frequency (binocular) of the eye tracker is 120 HZ.

For the same stimuli, both reference and distorted versions are watertight by adjusting the point size to the average distance among its 10 nearest neighbors all over all points in the point cloud [38]. Within a DPC, we utilize the same point size for all frames. All the point clouds are rescaled to a similar size, around 1.8m in height, to mimic a realistic tele-immersive scenarios. The VR scene is illuminated by a virtual lamp on the ceiling centralized above the models. The lamp is set as an area light with intensity values of 2 in Unity to simulate ordinary lighting in a room.

3.3 Experimental Procedure

Since there is no specific recommendation for designing subjective quality assessment experiments for DPC in VR, we have made an effort to adhere to existing ITU recommendations on images/videos [12, 14, 16] to establish an appropriate assessment methodology for DPC. In our subjective study, we opted for Absolute Category Rating with Hidden Reference (ACR-HR). Each time only a single DPC was shown to the observer; test materials included impaired DPC with randomly inserted intact HR sequences, represented as any other test stimulus. To avoid the effects of contextual or memory comparisons, we randomly generated a playlist for each subject, and care was given to avoid displaying the same DPC model consecutively.

Before the experiment, the visual acuity and color vision of every subject was tested using Snellen [11] and Ishihara [6] charts. Each subject was informed in advance about the manner and purpose of the study as part of the informed consent procedure. At the beginning of the session, the inter-pupillary distance was measured and the headset was adjusted by the subject accordingly. Then, a training session was conducted to help familiarize the subjects with the system, including the controllers and the naming of each button to communicate more easily. One training sequence, namely *loot*, was used, which was not included in the dataset. The quality range of *loot* was similar to the quality range of the test videos, giving the subjects a sense of what they would see in the formal sessions. The subjects always started at the same location, which is 1.5 meters away from the center of the virtual room, but could move freely from there onward. A DPC was located in the center of the virtual room, and each DPC was randomly rotated between $[0^\circ, 360^\circ]$ to avoid bias. During the experiment, the subjects were instructed to view each model carefully in the VR environment, by moving freely during the playback of each DPC. The subjects were also required to stand still while doing the calibration and error profiling. A fixed distance was set between the subjects and the error profiling scene, which was a circle composed of 9 eye-ball markers.

After feeling comfortable with the set-up, the participants were informed about the task that is assigned to them: "we ask you to examine a set of human DPC models, each model will be looped three times, each loop is last for 10 seconds; after visualization, we will ask you to rate the quality of the stimuli you are looking at, and in the same time, we will record your gaze-related data". To determine the number of loops, we referred to related papers on video quality assessment and eye-tracking-based visual saliency detection [7,22,41,48]. Additionally, in [28], the effect of exposition time by repeating the same video from 1 to 4 loops was explored, concluding that more loops do not necessarily result in more unique fixation points for most videos. Hence we chose 3 loops instead of once or an unlimited number. There were two dummy objects at the beginning of each session to familiarize the subject with the testing procedure and the rating scales. For each subject, the test was split into two rounds, lasting for around 30 minutes each, with a mandatory 5-minute break in between. Before and after each round, participants were requested to fill in a Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [19]. For every model and subject, a round was split into four consecutive steps:

1 *Calibration* is the process by which the geometric characteristics of a participant's eyes are estimated as the basis for a fully customized and accurate gaze point calculation, which is implemented to optimize the eye tracking algorithm. Calibration was done at the beginning of the experiment, and only when calibration was successful users could enter into the DPC playback stage.

2 *Inspection of models* is the step where the participants are visualizing DPC, while their trajectory and gaze-related information are recorded.

3 *Quality evaluation of models* requires the subject to rate DPC. The rating button was marked with labels ranging from "Poor" to "Excellent" to facilitate anchoring the rating process, and subjects could use their controllers to select and submit a score without taking off the headset.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

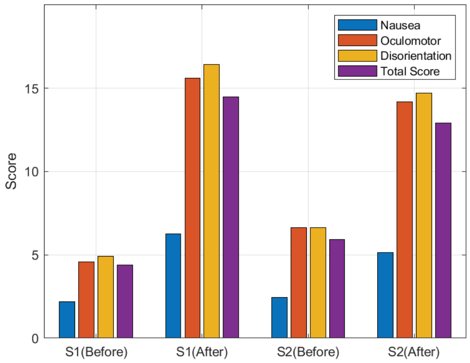


Figure 4: SSQ score for two test sessions

4 *Error profiling* is issued as the last step in order to estimate the accuracy of the gaze measurements due to calibration inaccuracies, or HMD displacements. A regular circle of 9 markers at pre-defined positions in the virtual scene was presented to the user. Based on the recorded gaze measurements, the average angular error was computed per marker online. This procedure allowed us to decide whether the gaze data obtained from a certain session was accurate or not.

A total of 40 participants took part in the subjective tests of this study, with a diverse composition that includes 1 non-binary individual, 19 males, and 20 females. The participants' ages ranged from 20 to 34, with an average age of 26.90 and a standard deviation of 3.51. Each participant observed half of the DPCs among all stimuli, leading to 20 opinion scores per sequence. In terms of occupation, the majority (80%) of the participants were students, ranging from bachelor to PhD levels. The remaining 20% were researchers, postdoctoral fellows, and one auditor. Regarding familiarity with VR devices, 7 participants had never experienced VR before the experiment, 26 participants had intermediate experience (using VR 1 to 3 times), and 7 of them were considered experts, having backgrounds as VR designers or researchers. Additionally, 22 out of 40 participants wore glasses during the experiment.

3.4 Data Processing

3.4.1 Processing of SSQ Data

SSQ comprises 16 symptoms which are further grouped into three different categories: Oculomotor, Nausea, and Disorientation; we also computed the total score. Fig. 4 suggests that simulator scores are increasing after performing the experiment. However, it can be seen that breaks help in reducing simulator sickness.

3.4.2 Processing of Opinion Scores

After removing the scores of the first two dummy objects, outlier detection was performed according to ITU-T Recommendations P.913 [16]. The recommended threshold values $r_1 = 0.75$ and $r_2 = 0.8$ were used. No outliers were found in our test. After outlier detection, the MOS was computed for each DPC. The associated 95% Confidence Intervals (CIs) were obtained assuming a Student's t-distribution. Additionally, the DMOS was obtained by applying HR removal, following the procedure described in ITU-T Recommendations P.913 [16].

3.4.3 Processing of Gaze Data

One subject walked into the body of two DPCs in the VR environment when observing, so the corresponding gaze data was not included. We ignored the initial 400 ms gaze data of each user to avoid unintentional gaze because of the unexpected appearance of the DPC. Then, only the valid gaze samples provided by the SRanipal SDK were selected. Each valid gaze sample was processed as follows:

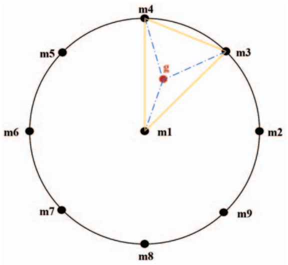
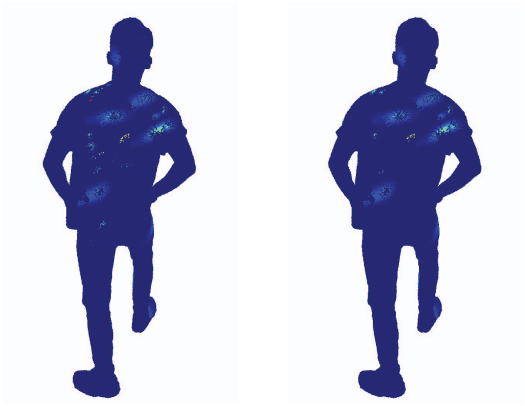


Figure 5: Estimation of the angular error for one gaze data



(a) Fixation points of *dancer*, front view (original fixations)

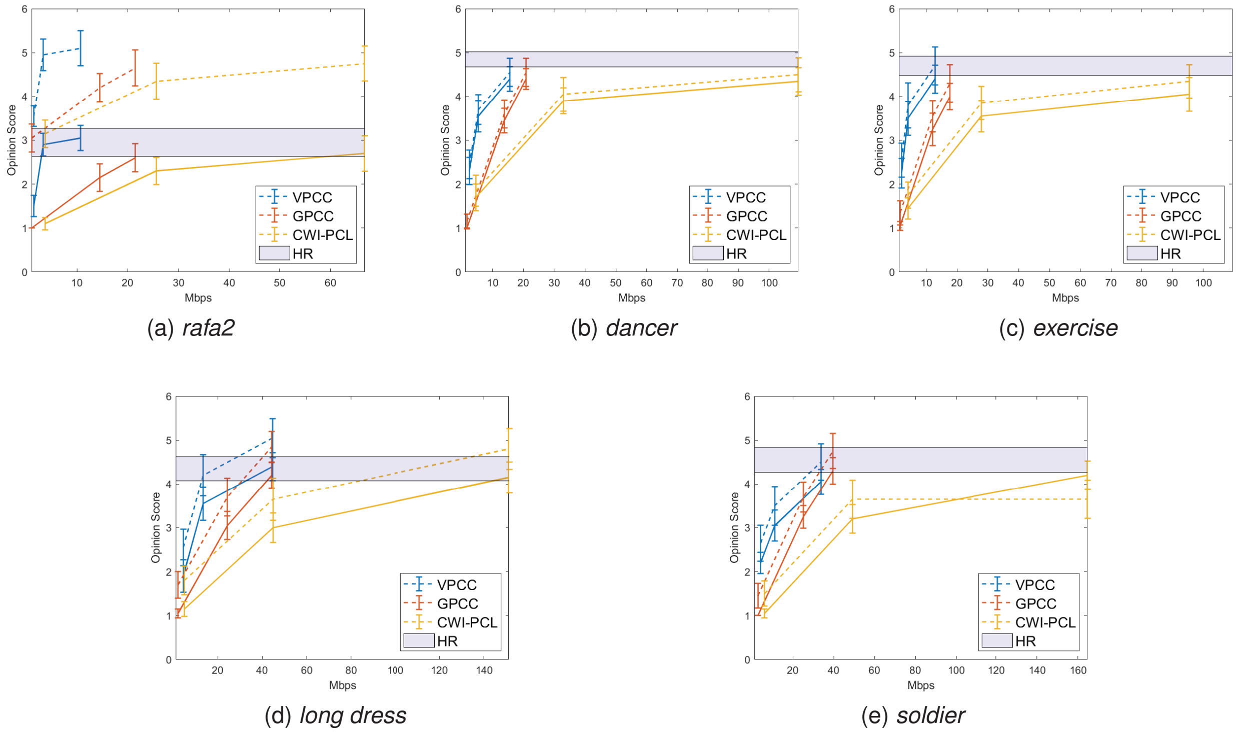
(b) Fixation points of *dancer*, front view (with filtering)

Figure 6: Fixation map comparison with/without filtering by DBSCAN.

1 *Verify the data validity of gaze data:* A barycentric interpolation with weights equal to corresponding angular errors obtained from the profiling was applied. A threshold of 7.5° was used to discard unintentional gaze. After displaying each target, 0.8 seconds will be waited before including the samples in actual calculations. This delay accounts for the initial moments in eye-tracking data during the actual experiment, which can be influenced by factors such as calibration stabilization, participant adaptation, and gaze analysis during fully engaged periods [34]. We used GazeMetrics [4], an open-source tool for measuring the data quality of HMD-based eye trackers, to compute the angular error. Finally, we applied a compensatory weighted average angular error to each gaze sample. This was repeated for every user gaze sample to maintain high-quality estimations while avoiding discarding useful data. Fig.5 illustrates the estimation of angular error for gaze data in 2D, g represents the intersection between the gaze ray and the plane formed by nine markers denoted as $m1$ to $m9$. These markers were positioned at a distance of 1.25 meters relative to the camera within the VR environment.

2 *Identifying fixation points of gaze data:* Taking into account the dynamic nature of our content, we chose the Dispersion-Threshold Identification (I-DT) [31] method. I-DT leverages the fact that fixation points, owing to their reduced velocity, tend to cluster in close proximity [44]. It identifies fixation points as groups of consecutive points within a particular dispersion, or maximum separation. The I-DT algorithm

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.



(a) *rafa2*

(b) *dancer*

(c) *exercise*

(d) *long dress*

(e) *soldier*

Figure 7: MOS (solid line) and DMOS (dashed line) against bit-rate, expressed in Mbps. HR scores are shown using a shaded purple plot.

requires two parameters, the dispersion threshold and the duration threshold. We set the dispersion threshold equal to 3° and the duration threshold equal to 100 ms, separately. Thus we took the average of these gaze points within the duration threshold as the fixation point.

3 *Mapping gaze data to DPC frames:* We proceeded by associating the filtered gaze data with the currently viewed frames and translating the gaze data (x,y,z) from world space into fixation points within that corresponding frame. As a result, we got all the gaze data in an endeavor to cover 300 frames in total. We adopted the truncated-cone-sector algorithm to assign weights to points in a given DPC frame [3].

4 *Fusing multiple users' gaze data to DPC frames:* A fixation map is the aggregation of fixation points from all users viewing the same DPC frame at a given timestamp, which can mark the region of interest. In our experiment, unintentional observation could cause isolated fixation points on DPC frame after mapping. Thus, it was necessary to filter out these noisy fixation points. We choose the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [10] algorithm to filter out noisy fixation points. Based on the density of fixation points on the point cloud, the DBSCAN is configured to remove the noisiest fixation points in clusters with high density at the same time be able to retain certain core fixation points in clusters with less density [27]. Fig.6 illustrates the effect of filtering noisy fixations. DBSCAN requires two parameters: ε is the radius of the circle to be created around each data point to check the density and θ is the minimum number of points required inside that circle for that data point to be classified as a core point. θ should increase as the point size α of a point cloud becomes small, which means a high-density point cloud. The minimum number of points is computed as

$$\theta = \left[\frac{2^7}{1+20*\alpha} \right]. \tag{1}$$

ε is decided by k-distance graph [32]. We took the average of fixation maps generated by multiple users, which is defined as

$$VS_f = \frac{1}{N}\sum_{n=1}^{N}(VS_n). \tag{2}$$

where VS_f is the fixation map for each DPC frame, VS_n is the fixation map for each DPC frame by one subject, specifically, VS_n also takes the average number of times a frame is viewed by one subject. N denotes one specific frame that has been viewed by N subjects in total. After we got the averaged fixation map for one DPC frame, we applied the DBSCAN filtering operation to get the final fixation map.

4 Experimental Result

4.1 Analysis of Opinion Scores

Fig.7 shows the results of the subjective quality assessment of the contents in 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 MOS, and a shaded plot for the corresponding CIs.

While evaluating the point cloud codecs, we observe that, under similar bitrates, VPCC codec exhibits the best perceptual quality, GPCC the second, and CWI-PCL is the last codec for all 5 contents. This observation is consistent with [38,47]. From the perspective of contents, MOS and DMOS present similar trends, as the MOS for

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

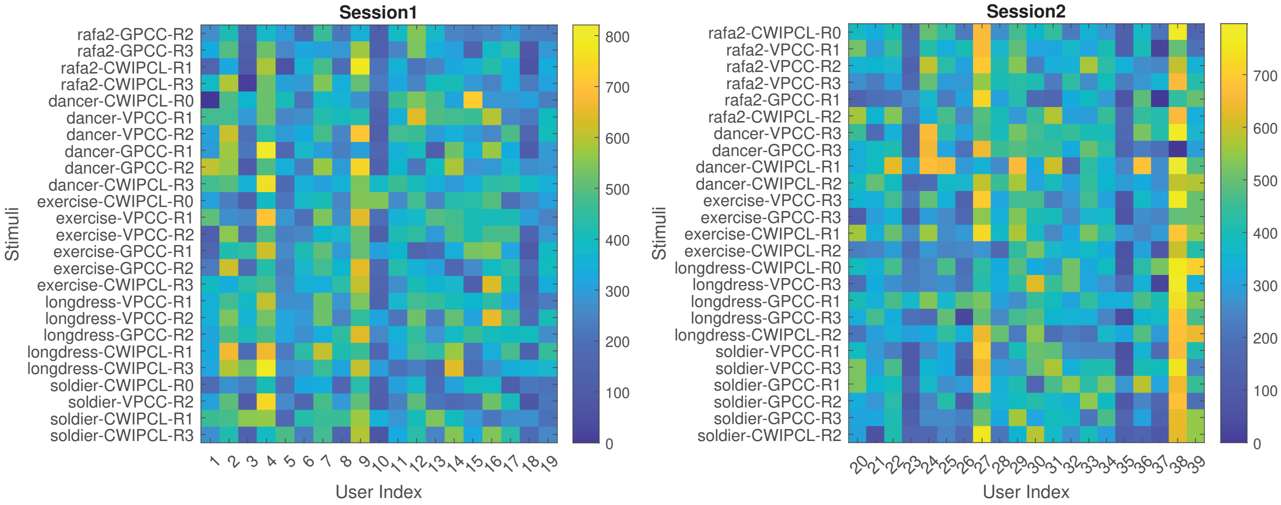


Figure 8: The fixations for each subject and for each content. Each row denotes the fixations on a specific content and each column denotes the fixations for each subject, respectively. $R1$ (low), $R2$ (medium), and $R3$ (high) indicate the bitrates of each codec, while $R0$ denotes the reference DPC.

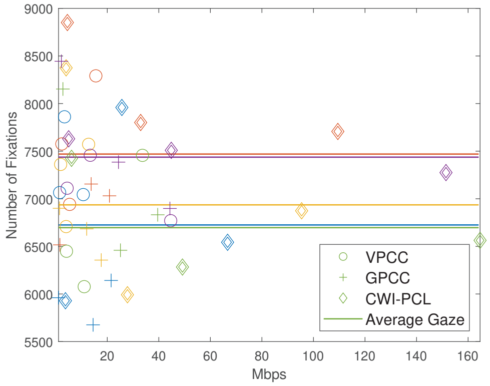


Figure 9: Fixations against bitrates, expressed in Mbps. The average number of fixations is expressed with a line. Each color denotes a content, specifically, *rafa2* is in blue, *dancer* is in red, *exercise* is in yellow, *long dress* is in purple and *soldier* is in green.

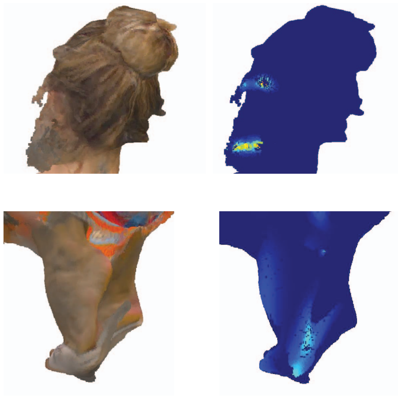


Figure 10: Fixation map on the hair and heal of *long dress*

the HR contents is between 4 and 5. The only exception to the trend is *rafa2*, for which the MOS for the reference content remains at around 3. This is likely related to the reconstruction error: compared with other contents captured in more professional studio settings, the reference version of *rafa2* does not offer a satisfactory quality. The calculated DMOS is between [3, 5], due to the fact that the reference content was rated so low.

4.2 Analysis of Gaze Data

To understand how and what users explore DPC in VR, we analyze the relationship between fixations and bitrates. Fig.8 represents the number of fixations of each subject on each content. It should be noted that the fixations are the filtered ones on individual DPC instead of the raw fixations of subjects. Fig.9 depicts the exact number of fixations across all subjects on different bitrates. Combining Fig.2, 8, 9, we have the following observations:

- Subjects are more interested in the high-motion DPC (i.e., with higher TI) compared with the low-motion one. For example, the average number of fixations on *dancer* and *long dress* is higher than *rafa2* and *soldier*, which have less TI on average.
- There is no indication that low-quality contents will receive less attention. In fact, we do not observe any particular trend regarding the number of fixations changing with varying quality levels. Visual attention for dynamic scenes should be considering both motion and quality.
- Certain subjects consistently exhibit a higher number of gaze fixations (e.g., user 27 and 38 in Fig. 8), possibly due to the individual differences of the participants or the accuracy of the device during the experiment.

We also explore where the subjects are looking at the DPC in VR, and how the quality degradation will impact the visual attention in a dynamic scene. Subjects pay attention to unrealistic rendering artifacts, such as high-heeled shoes and hair of *long dress*. Fig.10 depicts the fixation map on these two areas. Certain frames miss the heelpiece; certain frames exhibit unnatural hair rendering. Fig.11 shows the fixation map of both the reference and all distorted *long dress* point cloud frames. We can see subjects are interested in the face and the area with high motion. For all 5 contents, subjects tend to focus on the faces and the front view of the DPC, despite the random rotation of the DPC itself. No differences are observed for the salient area in different distortion levels. The heat values on the face are consistent across all the distortion levels; the heat value in high-motion areas floats with the motions; the heat value on the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

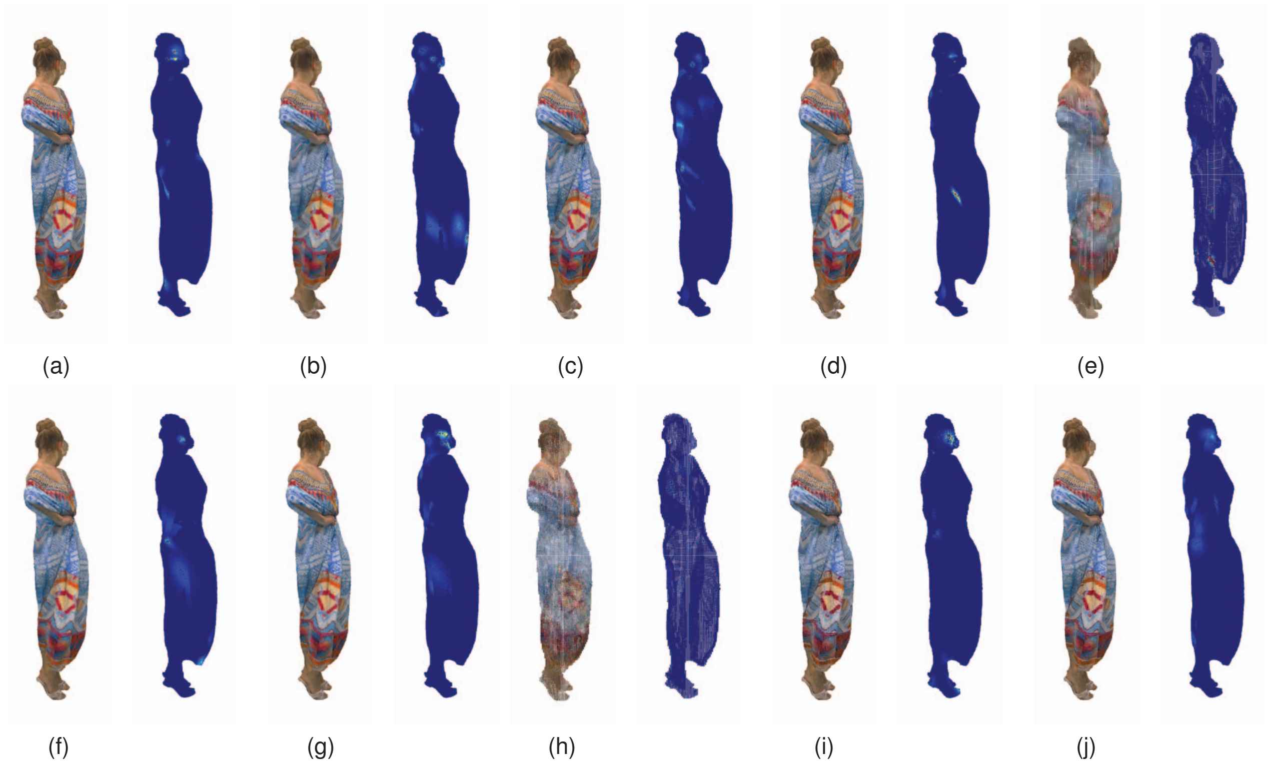


Figure 11: The referenced and distorted versions of point cloud *long dress* (frame 128) with corresponding visual attention maps based on the proposed processing protocol. Figures 11a is the reference version. Figures 11b-11j are the distorted version of *long dress* from low bitrate to high bitrate. Specifically, 11b-11d: VPCC, 11e-11g: GPCC, 11h-11j: CWI-PCL.

remaining area has no pattern. This randomness may come from unintentional fixation or the random rotation of DPCs.

5 Discussion

5.1 Dataset applications and prospective extensions

QAVA-DPC, encompassing MOS/DMOS, users' gaze data, and our meticulously processed visual attention maps, holds significant potential as a foundational reference for the following aspects: 1) Since it includes the raw data alongside the visual attention maps, the dataset can be used by researchers and practitioners to develop and test novel algorithms for the post-processing of gaze data and the creation of visual attention maps; 2) The dataset can be used for the development of objective quality metrics and visual attention prediction models for DPC without needing to conduct resource-intensive user studies; 3) Existing point-based objective quality metrics can be refined and tailored for DPC, to explore how to incorporate visual attention, and what is the added value; 4) Visual attention maps can be used as a comparison to static point clouds, to find the intrinsic differences between visual attention in dynamic and static contents in terms of perceptual quality assessment.

5.2 Task-dependent visual attention

As our experiment was devoted to evaluating the visual quality of the DPCs, the attention of our participants might have been focused on parts of the contents that assisted them in the task: for example, areas with patterns on which distortions would easily be spotted. That does not necessarily mean that the same area would be a salient region had the test been administered with a different task or task-free. Further experiments are needed to understand how visual attention changes based on the context of the task.

5.3 Influence of reference quality

Our results highlight how the same content, when mixed among different sets of contents, can receive different ratings: thus, the quality ratings of one content should always be considered in the context of the content set in which they are placed. Our results also show the importance of selecting the right set of contents for a subjective experiment. Reference quality should be considered in order to avoid biasing the subjects towards one or another content despite the SI-TI information. How to automatically perform such prediction of reference quality is, however, an open research issue [1].

6 Conclusion

In this study, we collect a dataset that includes gaze and quality scores for DPCs inspected in 6DoF. As part of the dataset, we additionally release the software used for the subjective experiment, the raw data that were collected, and the post-processing scripts used to compute visual attention maps. The dataset includes 50 DPCs, using 5 contents, and 9 distorted and 1 reference versions per content. A total of 15,000 visual attention maps for each DPC frame are finally provided. The main task of the presented study was to evaluate the quality of DPCs. In the future, we aim at expanding the dataset to include a task-free eye-tracking experiment, to make a comparison of these two visual attention maps and explore how the task impacts visual attention in virtual reality.

Acknowledgments

This work was supported through the NWO WISE grant and the European Commission Horizon Europe program, under the grant agreement 101070109, *TRANSMIXR* https://transmixr.eu/. Funded by the European Union.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

References

[1] E. Alexiou, Y. Nehmé, E. Zerman, I. Viola, G. Lavoué, A. Ak, A. Smolic, P. Le Callet, and P. Cesar. Subjective and objective quality assessment for volumetric video. In *Immersive Video Technologies*, pp. 501–552. Elsevier, 2023.

[2] 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:e27, 2019. doi: 10.1017/ATSIP.2019.20

[3] E. Alexiou, P. Xu, and T. Ebrahimi. Towards modelling of visual saliency in point clouds for immersive applications. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4325–4329. IEEE, 2019.

[4] I. B. Adhanom, S. C. Lee, E. Folmer, and P. MacNeilage. Gazemetrics: An open-source tool for measuring the data quality of hmd-based eye trackers. In *ACM symposium on eye tracking research and applications*, pp. 1–5, 2020.

[5] Q. Cheng, P. Sun, C. Yang, Y. Yang, and P. X. Liu. A morphing-based 3d point cloud reconstruction framework for medical image processing. *Computer methods and programs in biomedicine*, 193:105495, 2020.

[6] J. Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924.

[7] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet. A dataset of head and eye movements for 360 videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 432–437, 2018.

[8] X. Ding and Z. Chen. Towards mesh saliency detection in 6 degrees of freedom. *arXiv preprint arXiv:2005.13127*, 2020.

[9] E. d'Eon, B. Harrison, T. Myers, and P. A. Chou. Jpeg pleno database: 8i voxelized full bodies (8ivfb v2)–a dynamic voxelized point cloud dataset, 2019.

[10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, vol. 96, pp. 226–231, 1996.

[11] F. L. Ferris III, A. Kassoff, G. H. Bresnick, and I. Bailey. New visual acuity charts for clinical research. *American journal of ophthalmology*, 94(1):91–96, 1982.

[12] J. Gutierrez, P. Perez, M. Orduna, A. Singla, C. Cortes, P. Mazumdar, I. Viola, K. Brunnström, F. Battisti, N. Cieplińska, et al. Subjective evaluation of visual quality and simulator sickness of short 360° videos: Itu-t rec. p. 919. *IEEE transactions on multimedia*, 24:3087–3100, 2021.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

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

[15] ITU-T P.910. Subjective video quality assessment methods for multimedia applications. International Telecommunication Union, Apr. 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, Mar. 2016.

[17] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso. Point cloud rendering after coding: Impacts on subjective and objective quality. *IEEE Transactions on Multimedia*, 23:4049–4064, 2021. doi: 10.1109/TMM.2020.3037481

[18] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C. Bovik. Subjective and objective quality assessment of 2d and 3d foveated video compression in virtual reality. *IEEE Transactions on Image Processing*, 30:5905–5919, 2021.

[19] 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.

[20] S. F. Langa, M. Montagud, G. Cernigliaro, and D. R. Rivera. Multiparty holomeetings: Toward a new era of low-cost volumetric holographic meetings in virtual reality. *Ieee Access*, 10:81856–81876, 2022.

[21] G. Lavoué, F. Cordier, H. Seo, and M.-C. Larabi. Visual attention for rendered 3d shapes. *Computer Graphics Forum*, 37(2):191–203, 2018. doi: 10.1111/cgf.13353

[22] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7):547–558, 2010.

[23] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of visual communication and image representation*, 22(4):297–312, 2011.

[24] H. Liu and I. Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011. doi: 10.1109/TCSVT.2011.2133770

[25] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE transactions on Image Processing*, 14(11):1928–1942, 2005.

[26] 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. doi: 10.1109/TCSVT.2016.2543039

[27] A. Nguyen and Z. Yan. A saliency dataset for 360-degree videos. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pp. 279–284, 2019.

[28] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *2018 Tenth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2018.

[29] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 205–210, 2017.

[30] I. Reimat, E. Alexiou, J. Jansen, I. Viola, S. Subramanyam, and P. Cesar. Cwipc-sxr: Point cloud dynamic human dataset for social xr. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pp. 300–306, 2021.

[31] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 71–78, 2000.

[32] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[33] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko. Emerging mpeg standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2019. doi: 10.1109/JETCAS.2018.2885981

[34] L. Sidenmark, M. N. Lystbæk, and H. Gellersen. Ge-simulator: An open-source tool for simulating real-time errors for hmd-based eye trackers. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, ETRA '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3588015.3588417

[35] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018.

[36] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. doi: 10.1109/TVCG.2018.2793599

[37] S.Schwarz. Common Test Conditions for PCC. ISO/IEC JTC1/SC29/WG11 Doc. N18665, Jul. 2019.

[38] S. Subramanyam, J. Li, I. Viola, and P. Cesar. Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 127–136. IEEE, 2020.

[39] S. Subramanyam, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic, and P. Cesar. Subjective qoe evaluation of user-centered adaptive streaming of dynamic point clouds. In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2022.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.

[40] J. van der Hooft, M. T. Vega, C. Timmerer, A. C. Begen, F. De Turck, and R. Schatz. Objective and subjective qoe evaluation for adaptive point cloud streaming. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020. doi: 10.1109/QoMEX48832.2020.9123081

[41] A. van Kasteren, K. Brunnström, J. Hedlund, and C. Snijders. Quality of experience of 360 video–subjective and eye-tracking assessment of encoding and freezing distortions. *Multimedia tools and applications*, 81(7):9771–9802, 2022.

[42] I. Viola, J. Jansen, S. Subramanyam, I. Reimat, and P. Cesar. Vr2gather: A collaborative social vr system for adaptive multi-party real-time communication. *IEEE MultiMedia*, 2023.

[43] I. Viola, S. Subramanyam, J. Li, and P. Cesar. On the impact of vr assessment on the quality of experience of highly realistic digital humans. *arXiv preprint arXiv:2201.07701*, 2022.

[44] H. Widdel. Operational problems in analysing eye movements. In *Advances in psychology*, vol. 22, pp. 21–29. Elsevier, 1984.

[45] Y. Xu, Y. Lu, and Z. Wen. Owlii dynamic human textured mesh sequence dataset. In *ISO/IEC JTC1/SC29/WG1 1 input document m41658*, 2017.

[46] E. Zerman, P. Gao, C. Ozcinar, and A. Smolic. Subjective and objective quality assessment for volumetric video compression. *Electronic Imaging*, 2019(10):323–1, 2019.

[47] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic. Textured mesh vs coloured point cloud: A subjective study for volumetric video compression. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020. doi: 10.1109/QoMEX48832.2020.9123137

[48] W. Zhang and H. Liu. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, 26(3):1275–1288, 2017. doi: 10.1109/TIP.2017.2651410

[49] W. Zhang and H. Liu. Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications. *IEEE Transactions on Image Processing*, 26(5):2424–2437, 2017.

[50] Y. Zhang, L. Wang, and Y. Dai. Plot: a 3d point cloud object detection network for autonomous driving. *Robotica*, pp. 1–17, 2023.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on June 23,2024 at 21:03:50 UTC from IEEE Xplore. Restrictions apply.