User-Centered Adaptive Streaming of Dynamic Point Clouds for Virtual Reality Remote Communication

Subramanyam, S.

DOI
10.4233/uuid:0de0309e-0e1c-4919-ac55-a9b4dfe4fc6b

Publication date
2024

Document Version
Final published version

Citation (APA)
Subramanyam, S. (2024). *User-Centered Adaptive Streaming of Dynamic Point Clouds for Virtual Reality Remote Communication*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:0de0309e-0e1c-4919-ac55-a9b4dfe4fc6b

User-Centered Adaptive Streaming of Dynamic Point Clouds for Virtual Reality Remote Communication

USER-CENTERED ADAPTIVE STREAMING OF DYNAMIC POINT CLOUDS FOR VIRTUAL REALITY REMOTE COMMUNICATION

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T. H. J. J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 12 December 2024 om 12:30 uur

door

Shishir SUBRAMANYAM

Master of Science in Computer Science,
Deflt University of Technology, Delft, The Netherlands
geboren te Bangalore, India.

Dit proefschrift is goedgekeurd door de promoteren.

promotor: Prof. dr. P. Cesar
promotor: Prof. dr. A. Hanjalic
copromotor: Dr. I. Viola

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. T. H. J. J. van der Hagen,	Technische Universiteit Delft
Onafhankelijke leden:	
Prof. dr. S. Pont,	Technische Universiteit Delft
Prof. dr. F. Turck,	Universiteit Gent
Prof. dr. A. Raake,	Technische Universität Ilmenau
Dr. O. Niamut,	Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek

Keywords: Virtual Reality, Point Clouds, Adaptive Streaming, Quality Assessment

Printed by: Proefschriften Printen

ISBN 978-90-9039-461-9

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Contents

Summary

Remote communication applications have become a necessity in a globalised and connected world exemplified by the popularity of video conferencing applications. In recent years, virtual reality (VR) remote communication applications have emerged that aim to deliver a greater sense of co-presence and immersion in a shared virtual space where users are able to navigate freely while employing both verbal and non-verbal communication. Such applications require a volumetric user representation and point clouds have emerged as a popular format to represent real-time user reconstructions. However, volumetric point clouds are challenging to deliver over bandwidth-limited networks owing to the large volume of data required for dynamic streams. Adaptive streaming is the process of segmenting an object spatially and temporally in order to optimize the delivery of content by prioritizing the quality of spatial segments that are visible from a given viewport. In this thesis, we embed our research in the application scenario of VR remote communication with real-time point cloud user reconstructions and explore adaptive delivery optimizations. Previous work in the field mainly focused on using the entire point cloud object as the unit of bandwidth allocation in scenes containing multiple point clouds. Other recent work has relied on computationally complex surface estimation in order to spatially segment the point cloud that is unsuited to real-time applications.

This thesis focuses on investigating if a user-centred approach combined with a low-complexity adaptive streaming method can improve the quality of experience of interacting with point cloud user reconstructions in VR remote communication. We focus on optimizing the delivery of a remote user's reconstruction with spatial segmentation and surface orientation estimation in real-time combined with an auxiliary utility function to allocate the available bandwidth across available segments. The utility is defined based on the position and surface orientation of the point cloud segment and the position and orientation of the user's viewport.

We start our contributions in Chapter 2 by investigating subjective evaluation methodologies suitable for evaluating volumetric point cloud content. We first present the subjective evaluation methodology that we employed in the MPEG standardization activity for point cloud compression. This approach is based on creating video recordings of point clouds using a heuristic navigation path aimed at examining the content from a wide range of angles and distances that are then viewed on screens. We then propose a novel evaluation methodology for dynamic point clouds in immersive viewing environments with unrestricted user movement. This method is then used to compare the performance of point cloud codecs across a wide range of target bitrates across two viewing conditions. We found out that restricting user movements to three degrees of freedom had a small effect on the scores collected, however, most participants preferred to perform the evaluation with unrestricted movements. Most participants reported that they performed the evaluation based on the overall outline, the naturalness of movements and visual artefacts in the reconstruction.

In our second contribution in Chapter 3, we propose a low-complexity adaptive streaming approach based on spatially segmenting reconstructions based on camera visibility along with an auxiliary utility function to quantify the utility of segments based on the location and orientation of the user's viewport. We then compare five bandwidth allocation strategies to set the quality of each tile in the reconstruction along with a non-adaptive approach. We employed objective evaluation using image distortion metrics and a subjective evaluation using the protocol proposed in Chapter 2. We found that the best-performing bandwidth allocation strategy was able to achieve up to 57% bitrate saving while delivering the same objective quality and up to 65% bitrate savings while delivering the same subjective quality as compared to a non-adaptive approach using the same codec. We found that a utility-weighted approach to bandwidth allocation is able to deliver the highest quality gains as the transitions amongst adjacent tiles were not too drastic while maximizing the quality of high utility tiles.

Although the low-complexity adaptive streaming approach was validated using pre-recorded point clouds we need to deploy it in a live communication environment to ensure the quality gains are maintained. Therefore, in the last contribution in Chapter 4, we construct a fully functional VR remote communication application with live captured user reconstructions. We then propose a novel evaluation methodology for remote communication with a training task. We employ this method to compare the performance of our low-complexity adaptive streaming approach with a network adaptive approach and uncompressed streaming to serve as a baseline. We found that adaptive streaming leads to statistically significant gains in visual quality and quality of interaction. At higher bitrates, adaptive streaming is even able to match the quality delivered by uncompressed streaming. Adaptive streaming reduces CPU consumption while reducing the latency and increasing the playback framerate.

Reflecting on the achievements reported in this thesis, we conclude that subjective assessment needs to account for real user interaction patterns in order to assess delivery optimizations for volumetric media (Chapter 2). A low-complexity approach to adaptive streaming can result in significant gains to both objective and subjective quality (Chapter 3) while a utility-weighted distribution of bandwidth across tiles ensures smooth quality transitions and maximal quality gains. Adaptive streaming of point cloud user reconstructions in remote communication yields significant gains to visual quality and quality of interaction amongst remote participants (Chapter 4) while also improving playback performance and reducing resource consumption. However, there is a need for new standardized test methods for VR remote communication that can handle novel interaction techniques and immersive content. Delivery optimizations are essential for the widespread adoption of this technology on commodity hardware. With recent events like the Covid-19 global pandemic, there is a greater need for remote communication applications that offer a better sense of co-presence and mutual sensing of emotions. Moreover, such applications have the potential to reduce the carbon footprint of travel for the purpose of in-person communication.

1

Introduction

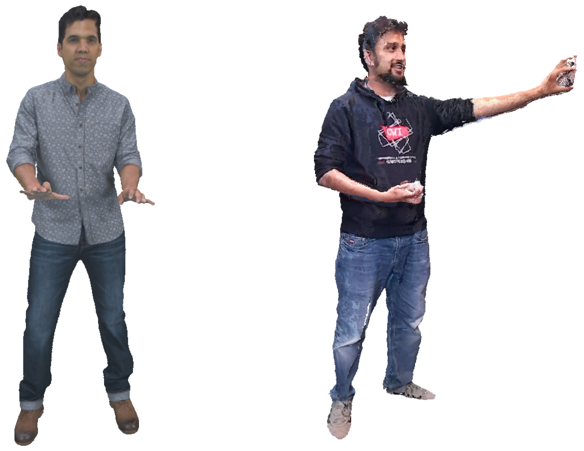
1.1. BACKGROUND AND PROBLEM MOTIVATION

Remote communication and collaboration has rapidly become a necessity in a globalized and connected world. Video conferencing applications have emerged as the most popular solution for remote communication. Notwithstanding their popularity, it is estimated that travel for the purpose of in-person communication is responsible for roughly eight percent of US energy consumption [1]. With recent events like the Covid-19 global pandemic there is an increased need for applications that can deliver a greater sense of co-presence and mutual sensing of emotions in remote communication. Current video conferencing solutions have clear limitations in this regard [2–5].

In recent years, advances in head-mounted displays (HMDs), scanning and 3D capturing devices have enabled a plethora of novel immersive virtual reality (VR) experiences. Volumetric reconstructions of real-world objects and scenes have facilitated 6 Degrees of Freedom (6DoF) VR applications where users are free to navigate within the 3D scene. Immersive Virtual Reality (VR) applications offer an increased sense of presence and immersion. These applications have emerged as a promising alternative for remote communication and telepresence [6–13]. They allow users to employ both verbal and non-verbal communication in a shared virtual space. Avatars are a core part of these VR applications like social communication[14], sports training[15], or healthcare[16]. A major line of scientific work has focused on how to make such avatars more realistic, interactive, and autonomous [17–19]. The avatars used to embody users in these applications can either be synthetic representations or real-time photorealistic 3D reconstructions typically using depth sensors. Previous work in the field has demonstrated that photorealistic user reconstructions improve immersion and communication [20, 21] as compared to synthetic avatars.

These volumetric reconstructions can be represented using dynamic meshes, time-varying meshes and 3D point clouds. Dynamic meshes record a single reconstruction with fixed connectivity amongst vertices and fixed textures on the mesh faces. The motion of the reconstruction over time is represented by the motion of the vertices usually recorded using motion capture. This format is typically used for prerecorded content with offline capture and specialized motion capture setups. Time-varying meshes record watertight meshes at every frame of the captured sequence with varying geometry, connectivity and textures [22]. This format is more suited for live capture however, meshes require significant pre-processing to triangulate faces and are sensitive to noise, making them challenging to employ in real-time applications.

Due to their relative simplicity for acquisition and rendering, 3D point clouds have been increasingly used to represent users in 6DoF social VR applications [9, 10]. Point clouds represent the geometry of the object using a collection of point coordinates; along with the geometry, associated attributes such as color and transparency can be stored on a point basis. Some examples of point clouds are shown in Figure 1.1. Point clouds incur no additional computational overhead for triangulation, making them suitable for real-time applications. However, dense, high precision photorealistic point clouds require a large volume of data (a single frame of roughly 1M points takes around 20 MBytes), making it challenging to transmit over bandwidth-limited networks [23, 24]. In recent years, there has been significant research interest in point cloud compression, including the introduction of new MPEG standard codecs for static and dynamic point clouds [25]



(a) *Offline dense voxelized point cloud [23]*

(b) *Live captured point cloud [26]*

Figure 1.1: Sample 3D human point cloud reconstructions

as well as low latency transmission mechanisms [9, 10].

As compared to traditional video, where users have a passive role, 6DoF social VR applications introduce additional challenges in the pipeline for delivering and representing content with low latency, in order to offer greater levels of interaction. Moreover, real-time interactions in a live communication system require low-latency delivery mechanisms and client buffer manager strategies. At the same time, user navigation can be exploited to optimize the delivery of volumetric data. As only parts of the content are visible at any given time, user-adaptive streaming solutions can be deployed to reduce the bandwidth allocation for parts of the content that are outside of the field of view, ensuring a better Quality of Experience (QoE) for the parts that are visible.

In order to reduce the wastage of bandwidth on content that is not consumed, we propose a user-adaptive streaming mechanism utilizing independently decodable spatial segments or tiles. This leads to the core hypothesis of this thesis:

User-centered adaptive streaming improves the quality of experience of interacting with point cloud user reconstructions in VR remote communication

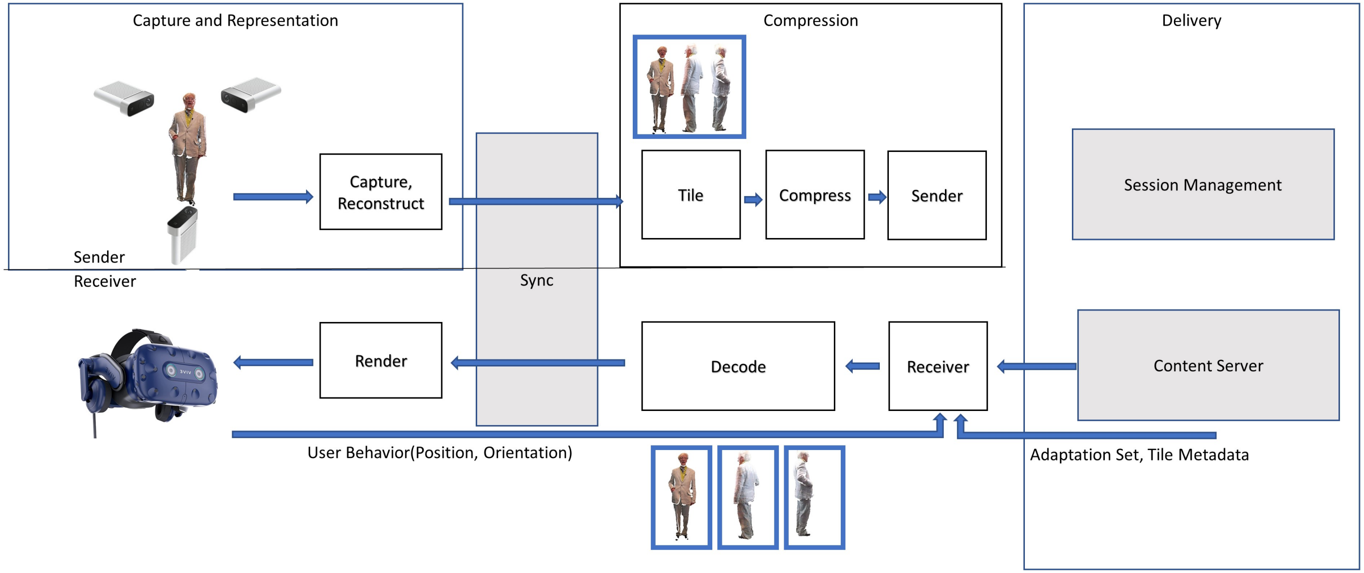


Figure 1.2: Overview of the VR remote communication pipeline

1.2. Basic Concepts

To understand the impact of user-centred adaptive streaming of point cloud user reconstructions we need to define the basic concepts involved in the delivery pipeline of point cloud based VR remote communication. To do this, we follow a reference architecture such as the pipeline shown in Figure1.2.

1.2.1. Capture and Representation

Point clouds of human subjects are typically captured using photogrammetry or direct depth capture. Photogrammetry is more suited for offline capture of dense photorealistic point cloud reconstructions using a large array of cameras that are then voxellized during postprocessing. An example of such a point cloud is shown in Figure 1.1 (a). In real-time systems, point clouds are captured using multiple consumer depth sensors to create a complete reconstruction as shown in Figure 1.1 (b). Each camera records an RGBD image that is converted to a 3D point cloud using the intrinsics of the individual sensors on a camera. Multiple cameras can be calibrated to a common coordinate system by computing an extrinsics matrix for each camera. The resulting point clouds can be fused together after synchronizing the cameras to generate a complete user reconstruction. After the reconstruction step, tiling can be applied to segment the point cloud into independent non-overlapping regions or tiles. Each tile contains some associated metadata such as the surface orientation, bounding box and centroid location. In the next step, the point cloud tiles are compressed into linearized data blocks to deliver them efficiently over bandwidth-limited networks at high framerates suitable for remote communication. In this thesis, we initially use offline point cloud datasets to develop and evaluate adaptive streaming. We then evaluate our approach in a VR remote communication pipeline with live captured point clouds.

1.2.2. Compression

Point cloud compression and streaming have received significant research attention in recent years with the launch of two new MPEG compression standards [25]. The V-PCC standard codec for dynamic point clouds projects point cloud geometry and attributes onto separate 2D patches that are then packed into video tracks along with an additional occupancy track. These video tracks are then encoded using legacy video codecs making this approach suitable for a relatively dense and uniform distribution of points. The G-PCC standard codec uses an octree space partitioning structure to code geometry and can be optionally combined with an additional surface reconstruction step using the TriSoup approach [27]. G-PCC also includes several modules for attribute coding, the lowest complexity coder uses the Region Adaptive Hierarchical Transform (RAHT) [28]. This codec is targeted at irregular sparse distributions of points making it suitable for live captured static point clouds. However, both codecs utilize high-complexity encoding, which makes them unsuitable for real-time communication. At the start of the MPEG standardization activity, an anchor codec proposed by Mekuria et al. [8] was introduced. This codec utilizes octree occupancy to code geometry and scans attributes to map them to a 2D grid to maximize correlations amongst co-located points and encodes them with legacy JPEG image compression. This approach offers low encode and low decode complexity making it suitable for real-time framerate-sensitive applications such as VR remote communication. In the work presented in this thesis, we use the MPEG anchor codec to evaluate adaptive streaming and we use the V-PCC standard codec to serve as a baseline.

1.2.3. DELIVERY

User adaptive streaming of 360° videos has received significant research and industrial interest in recent years. While legacy video codecs can be used to encode the planar representation of the 360° videos when the content is being played out, only a small fraction of the whole video is visible from the user's viewport at any given moment. In order to exploit user interactions and the corresponding visible portions of the sphere to optimize delivery, a popular approach is to differentiate the quality of different regions in the frame. This approach is included in the Motion Controlled Tile Sets (MCTS) coding included in the High-Efficiency Video Coding (HEVC) codec [29]. In this approach, the projected video is split into independently decodable, non-overlapping spatial regions called tiles. The tiles are encoded at multiple quality levels, and the client is able to select a quality for each tile in the frame based on visibility from the users' viewport. While additional bandwidth overhead is expected, due to a loss of compression efficiency and additional metadata [30], the approach allows for a more flexible delivery based on user interactions, which has led to it being successfully implemented and integrated into standards[31] [32].

In this thesis, we aim to extend the 360° video streaming approach to 3D dynamic point clouds with a user-centered approach based on navigation in 6DoF. For 360° videos spatial adaptation is done by enhancing the quality of visible portions of the image surrounding the viewer. In our approach, we implement spatial adaptation by segmenting each 3D object into tiles and estimating visibility from an external viewport based on proximity and surface orientation.

In the reference pipeline architecture, as shown in Figure 1.2, the sender prepares a manifest file that contains information on the number of tiles, tile metadata and the available quality levels with their associated estimated bandwidth requirements. The compressed data blocks for each tile and quality level are then uploaded to the server as media segments along with a manifest file that is periodically updated. This is then served to client receivers typically over HTTP. The receiver first requests the manifest file, then based on the location and orientation of the user's viewport and the available bandwidth selects an optimal quality for each tile. This quality selection is based on the trade-off between the required bitrate and the visual quality of the final reconstruction. The receiver then downloads the associated compressed blocks.

1.3. CHALLENGES / KNOWLEDGE GAPS

In order to design and evaluate tiled adaptive streaming of point clouds for VR remote communication, several challenges need to be addressed. This includes protocols for evaluating the quality of point cloud reconstructions in immersive environments, strategies to segment point clouds into independently decodable tiles and protocols to evaluate remote communication in VR.

1.3.1. Subjective Quality Assessment of Point Clouds

The JPEG and MPEG activities for the establishment of standards have raised interest in subjective evaluation of point cloud contents. Initially, most studies were focused on evaluating the quality of static contents on 2D displays. Zhang et al. [33] evaluated the impact of degradation such as down-sampling and noise generation. Their results indicate that human visual perception can tolerate color noise more than geometry or coordinate noise. Alexiou et al. [34] compared the Double Stimulus Impairment Scale and the Absolute Category Rating methodology for quality assessment of geometric degradations. Subjective evaluations using the Screened Poisson surface reconstruction for rendering purposes were conducted in [35], with results indicating different rating behaviours when compared to evaluations using raw point clouds. In [36], a comprehensive study of the rate-distortion performance of the entire set of encoders in the MPEG standard [25] across the rate points used in development [37] is presented, along with an evaluation of objective quality metrics. In [38], a subjective evaluation campaign of a subset of MPEG point cloud codecs was issued, across four independent laboratories with different testing equipment, revealing high correlation. Javaheri et al. [39] evaluated both subjectively and objectively the quality of point cloud contents under geometry artifacts occurring by different compression approaches. The same authors also provide an evaluation of point cloud rendering techniques and codecs [40] to compare point-based rendering methods (with and without recoloring) and surface reconstruction-based mesh rendering against the PCL [41], V-PCC and G-PCC codecs. The results indicate that surface reconstruction can mask some of the artifacts introduced by the codecs. Moreover, Dumic et al. [42] conducted a subjective evaluation on point cloud rendering and display devices, showing that users do not have a preference between 2D or 3D displays, while preferring inspection of raw point clouds against meshes created using Poisson surface reconstruction. Hooft et al. [43] compared subjective and objective evaluation results from adaptive streaming of multiple point clouds using algorithmically generated camera paths and the adaptation schemes described in [24], with obtained videos shown to users on a 2D screen.

Quality evaluation in immersive environments has recently gained popularity in the research community. Mekuria et al. [20] evaluated quality based on factors such as immersiveness, togetherness and realism between users in a social VR experience, which are represented using point cloud reconstructions as well as avatars. In [44], subjective quality assessment of point cloud contents was performed in Augmented Reality (AR) using HMDs, while in [45], the authors compare the collected quality scores with subjective ratings from experiments using a 2D monitor. The PointXR toolbox for subjective evaluation of static point clouds in VR was proposed in [46], and was employed to assess the performance of color encoders that have been integrated in G-PCC. Tran et al. [47] found that when evaluating quality in immersive environments, factors like cybersickness and presence should not be overlooked. However, the approach presented in [46] was based on loading a fixed point cloud representation on physical memory, making it unsuitable for evaluating adaptive streaming.

There is a need for an immersive evaluation environment and protocol to perform subjective quality assessment of dynamic point clouds with distortions and artefacts after compression.

1.3.2. TILING AND SEGMENTATION FOR VOLUMETRIC POINT CLOUDS

Segmentation is the division of the point cloud into clusters of points that are homogeneous with respect to a selected characteristic. In order to adaptively stream a point cloud object based on the orientation of the user's viewport, we need to cluster points based on the orientation of their underlying surface. There are two classes of methods to estimate normals to the surface for each point [41]. The most accurate method is to reconstruct the surface and create a watertight mesh from the point cloud frames. The normals to the surface can then directly be associated with each point. The second approach is based on inferring the underlying surface based on the point local neighborhood. Nguyen et al. [48] present a taxonomy of five classes of segmentation algorithms: edge based, region based, attribute based, model based, and graph based. Attribute-based segmentation methods can be used to tile point clouds, and account for surface orientation. These methods rely on estimating additional attributes for each point, e.g., normals, to obtain the surface orientation before clustering. Thus, these solutions are not suitable for a real-time system, as inferring the surface normal requires repeated eigendecomposition for the local neighbourhood of each point.

Initial works on adaptive streaming of point clouds utilized entire point cloud objects as the basic unit of bandwidth allocation in scenes containing multiple point cloud objects. Hosseini *et al.* [49, 50] propose DASH-PC for dynamic adaptive view aware point cloud streaming. They propose three spatial subsampling techniques to create multiple representations of point cloud objects in a scene. The density of each object representation is used by the client for bitrate allocation based on human visual acuity. Hooft *et al.* [24] propose PCC-DASH, a standards-compliant means for HTTP adaptive streaming. They present three heuristics based on the users viewport and distance to the object to allocate bitrate to different objects in the scene. Different ranking metrics and bitrate allocation heuristics had to be selected for different scenes and user navigation paths.

Another approach used in previous work is to split each point cloud object into tiles that are then used as the unit of bandwidth allocation. Park *et al.* [51] define a utility per tile based on the user's proximity, point cloud surface quality and display device resolution. To account for interactions, they propose a window-based design for the client buffer manager with greedy utility maximization. This type of rasterization and pixel occupancy based approach is not suitable as computing this at every frame is computationally expensive. He *et al.* [52] propose view-dependent streaming over hybrid networks. Each point cloud frame is projected onto the six faces of a bounding cube, with a color and a depth video created per face. The user can request videos that correspond to particular faces of the cube in high quality from the edge node of a bidirectional broadband network, reconstructing the point cloud from the downloaded depth and color videos at the receiver end. This approach requires a redundant extra reconstruction step at each receiver. Li *et al.* [53] propose a joint communication and computational resource allocation framework to stream and decode pre-recorded point clouds. They also propose a QoE metric to guide tile selection based on the users viewport, distance to tile and available quality levels. Lee *et al.* [54] propose GROOT a real-time streaming system to reduce decoding overhead by dividing the point cloud into cells defined by the leaf nodes of an octree represented in a parallel decodable tree. Han *et al.* [55] propose ViVo using a similar approach and employ machine learning models to predict viewport movement. Liu *et al.* [56] follow a similar approach, they include an uncompressed base layer and use fuzzy logic based quality selection. This type of approach using the leaf nodes of the octree as an enhancement layer is currently not suitable for real-time systems as it adds an extra surface orientation estimation step that introduces additional delays in the pipeline. Existing works either perform an objective quality evaluation using prerecorded navigation paths and image distortion metrics [24] or rely on recording videos using fixed navigation paths to perform subjective evaluation on 2D screens [43].

There is a need for developing low complexity point cloud tiling strategies suitable for adaptive streaming in real-time applications like VR remote communication.

1.3.3. Evaluating Photorealistic VR Remote Communication

Advances in low-latency streaming and volumetric point cloud delivery mechanisms have led to the emergence of novel teleimmersion systems that allow distributed remote users to communicate as themselves in a shared environment with realistic user reconstructions. Microsoft released the RoomAlive Toolkit for creating interactive Augmented Reality (AR) experiences [11, 57]. Mekuria et al. proposed a teleimmersive system that blends avatar representations and photo-realistic reconstructions of users in a shared virtual environment [20]. Cernigliaro et al. propose a point cloud multi-point control unit for optimizing holo-conferencing systems [58]. Gunkel et al. introduced VRComm [10], a web-based social VR communication system using photo-realistic user reconstructions that was evaluated using both simulations and subjective studies.

In order to optimize the delivery of volumetric content for end users in remote communication applications the quality of multiple objects and spatial segments within those objects needs to be set based on the needs of the user. The user's viewport location and orientation are considered along with the available bandwidth to adaptively set a quality level for visible surfaces based on proximity and visibility. To evaluate different adaptation strategies and streaming configurations the perceived quality of the resultant reconstruction needs to be evaluated with a user-centric approach.

Quality assessment for remote communication is usually conducted using subjective user studies that are either passive or active. Passive tests involve asking users to rate prerecorded clips of content. This approach to evaluation is more suited for standardized testing of codecs with offline content and has limited ecological validity in remote communication [59, 60]. Active tests involve multiple remote participants being placed in an interactive live communication system. The International Telecommunication Union published recommendations to define evaluation methods for quantifying the impact of terminal and communication link performance on point-to-point audiovisual communication [61]. The recommendation contains sample tasks such as name guessing, story comparison, picture comparison, object description and building blocks. Schmitt *et al.* [60] utilize the building blocks task to develop and evaluate personalized quality of experiment metrics for multiparty video conferencing at varying bitrates. Smith *et al.* [62] compare face-to-face communication with embodied and unembodied remote VR communication. They propose a task involving negotiating apartment layout and furniture placement based on blueprints. Li *et al.* [2] compare face-to-face, videoconferencing and Facebook spaces VR communication in the context of a photosharing task. They found that Facebook spaces is able to closely approximate face-to-face photosharing. In general, these methods have been used to compare VR remote communication with other technologies and with face to face communication.

In order to evaluate communication, several questionnaires have been proposed in the literature. Toet *et al.* [63] propose the holistic mediated social communication (H-MSC) framework and associated questionnaire to evaluate the experience of spatial presence as well as social presence. The framework is general enough to be used for any mediated social communication system. Slater *et al.* [64] and Witmer *et al.* [65] have proposed two popular questionnaires aimed at measuring presence in virtual environments. Kangas *et al.* [66] present a pragmatic task-related questionnaire that they use to evaluate VR interaction techniques in a rigid object manipulation task. Li *et al.* [2] propose a social VR questionnaire that evaluates Quality of Interaction/Communication *(QoI)*, Presence/Immersion and Social Meaning. The existing ITU recommendations are insufficient to handle novel interaction techniques and immersive content inherent to VR communication.

There is a need to define protocols and to perform point cloud delivery optimization evaluations in realistic testing grounds to ensure ecological validity for remote communication. To this end, ITU-T has recently launched a new activity [67] to develop assessment methods for extended reality meetings.

1.4. Research questions

In order to address the knowledge gaps identified in the previous section and to evaluate the core hypothesis of this thesis regarding the impact of user-centred adaptive streaming of point clouds on VR remote communication, we define the following research questions. An overview of the research questions is shown in Figure 1.3.

1.4.1. Evaluating the perceived quality of dynamic point cloud sequences in immersive environments

In this thesis, we intend to design and study the impact of delivery optimization strategies on the quality of the resulting point cloud reconstruction. We focus on point cloud human reconstructions that are used in real-time VR remote communication. In order to drive optimizations in content delivery, it is essential to understand the perceived quality of dynamic point clouds in virtual environments and viewing conditions with ecological validity. This leads to our first research question:

R1: How can we measure the perceived quality of dynamic point cloud user reconstructions in immersive environments?

This research question is further divided into:

R1.1: What is the impact of immersive viewing environments on subjective quality assessment?

Previous work on subjective assessment of point clouds focused on static frames and did not perform the evaluation in VR [68–72]. In order to understand how interacting in the virtual space affects the perception of quality we need to compare different viewing conditions in VR. Specifically, we need to evaluate user preferences and perceived quality with restricted 3 degrees of freedom (3DoF) movement and unrestricted 6 degrees of freedom (6DoF) movement. We need to ensure that the immersive viewing environment does not introduce other impediments to the experience in terms of cyber sickness, user comfort and preferences.

R1.2: What factors do users rely on while performing subjective quality assessment of dynamic point clouds?

In order to optimize codec configurations and streaming methodologies, it is essential to identify the factors that affect the perceived quality of dynamic point cloud reconstructions. This is required to understand the underlying factors that influence the quality scores reported by users. This can help inform the design of future streaming architectures and optimal configurations of standardized codecs for prerecorded dynamic sequences of photorealistic point cloud user reconstructions.

1.4.2. INVESTIGATING THE SUITABILITY OF USER-CENTRED TILED ADAPTIVE STREAMING OF DYNAMIC POINT CLOUDS

In order to develop and validate user-centred adaptation strategies for streaming tiled point cloud sequences, we need to segment the point cloud surface into independently decodable tiles. We need to infer the underlying orientation of the point cloud surface in each of the tiles in order to optimize how the available bandwidth budget is spent and maximize the perceived quality of the reconstruction as viewed by the user from a given viewport. The client buffer manager needs to utilize the user's navigation patterns as well as the available bandwidth to optimize the quality of the reconstruction presented to the user. The utility of each tile needs to be identified based on the position and orientation of the point cloud surface in order to drive these optimizations. This leads to our second research question:

R2: How can we optimally allocate the available bandwidth across independently decodable spatial segments?

This research question is further divided into:

R2.1: Does user-centered tiled adaptive streaming offer significant gains to reconstruction quality and bandwidth savings? In order to determine if user-centered tiled adaptive streaming is suitable for VR remote communication, we need to compare and evaluate the objective and subjective quality gains achieved as compared to traditional network adaptive streaming with the same available bandwidth. We need to ensure that the gains are consistent across a range of target bitrates and point cloud contents without introducing additional artefacts and without inducing fatigue and motion sickness in a VR environment.

R2.2: What are the acceptable quality levels amongst adjacent tiles to maximize the quality of the final point cloud reconstruction?

We need to understand the acceptable quality transitions across adjacent tiles in the reconstruction presented to the user. Large quality differences can lead to obvious discontinuities in the appearance of the human reconstruction and small quality differences can lead to wastage of available bandwidth on surfaces that are either partially or completely occluded from the viewer. We need to optimize this trade-off consistently across content types, target bitrates and view distances.

1.4.3. Optimizing the delivery of point cloud user reconstructions in VR remote communication

VR remote communication applications offer a greater sense of co-presence and mutual sensing of emotions between remote users. Previous research on these applications has shown that realistic point cloud user reconstructions offer better immersion and communication as compared to synthetic user avatars. However, photorealistic point clouds require a large volume of data per frame and are challenging to transmit over bandwidth-limited networks. In order to optimize the limited bandwidth budget on the reconstruction viewed by the user and develop adaptive streaming strategies, we need to address our final research question and the associated sub-questions:

R3: How can we optimize the delivery of streams of dynamic point clouds in VR remote communication?

This research question is further divided into:

R3.1: How does tiled user-adaptive point cloud streaming impact the perceived quality of remote user reconstruction?

We need to understand the impact of tiled adaptive point cloud streaming on the perceived visual quality of the user reconstruction presented during remote communication. It is essential to identify if tiles of varying quality are able to deliver a higher QoE as compared to reconstruction with fixed quality across surfaces. In addition, it is also necessary to evaluate the impact of adaptive streaming on a user's sense of feeling understood, engaging in conversations, sensing the emotions of remote users and feeling comfortable in the virtual environment. In addition, we need to ensure that this approach does not inhibit a user's ability to interact and navigate in such applications by introducing motion sickness.

R3.2: What is the computational overhead of using tiled adaptive point cloud streaming?

We need to understand the computational overhead associated with adaptive streaming in terms of resource consumption. This can significantly impact the performance of the VR application in terms of latency and framerate. A significant reduction in playback performance can lead to adverse consequences such as cyber-sickness in VR. It is essential to ensure that adaptive streaming strategies are able to deliver a satisfactory experience in real-time framerate-sensitive applications like VR remote communication on commodity hardware.

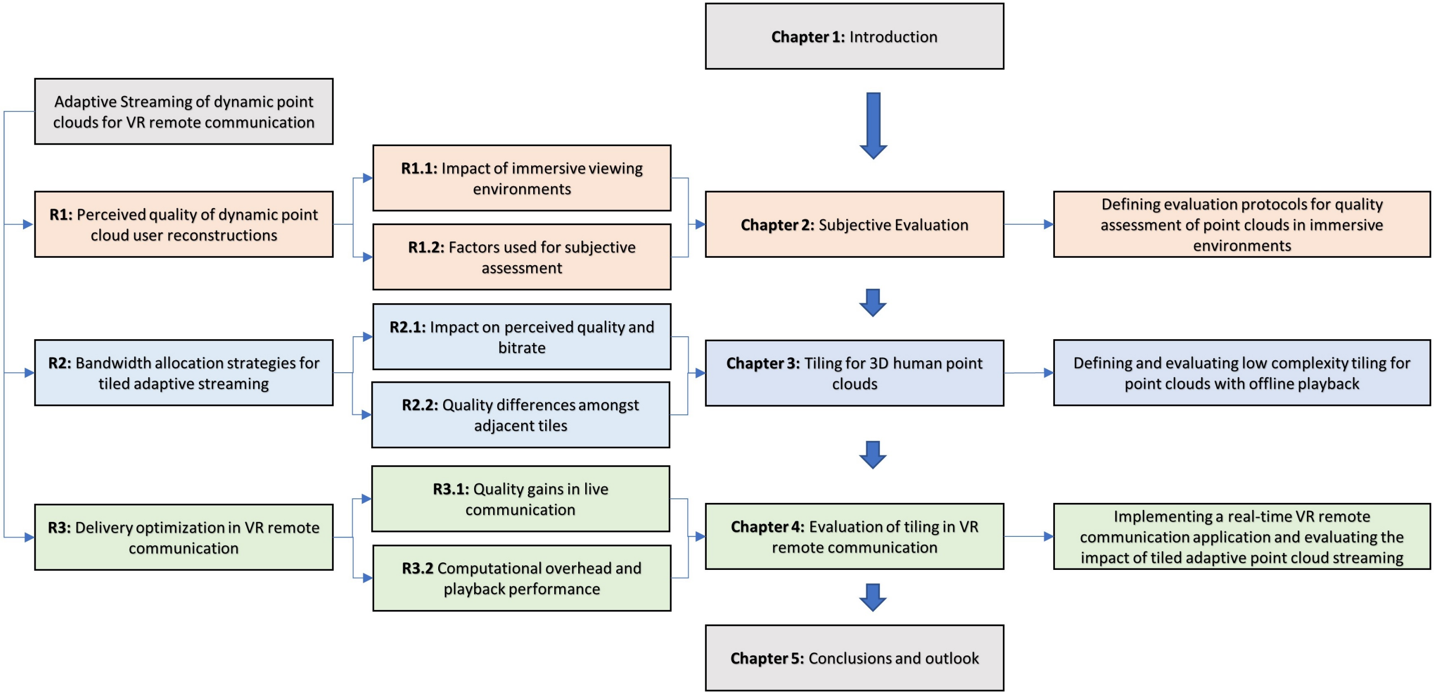

Chapter 1: Introduction
Adaptive Streaming of dynamic point clouds for VR remote communication
R1.1: Impact of immersive viewing environments
R1: Perceived quality of dynamic point cloud user reconstructions
Chapter 2: Subjective Evaluation
Defining evaluation protocols for quality assessment of point clouds in immersive environments
R1.2: Factors used for subjective assessment
R2.1: Impact on perceived quality and bitrate
R2: Bandwidth allocation strategies for tiled adaptive streaming
Chapter 3: Tiling for 3D human point clouds
Defining and evaluating low complexity tiling for point clouds with offline playback
R2.2: Quality differences amongst adjacent tiles
R3.1: Quality gains in live communication
R3: Delivery optimization in VR remote communication
Chapter 4: Evaluation of tiling in VR remote communication
Implementing a real-time VR remote communication application and evaluating the impact of tiled adaptive point cloud streaming
R3.2 Computational overhead and playback performance
Chapter 5: Conclusions and outlook

Figure 1.3: Research questions and thesis structure

1.5. Contributions and Thesis Outline

An overview of the thesis structure and the associated research questions needed to address our core hypothesis is shown in Figure 1.3. Below we provide a description of the contributions presented in each chapter to address the associated research questions.

1.5.1. On Evaluating the quality of dynamic point cloud reconstructions in immersive environments

Research question 1 investigates a subjective evaluation methodology for dynamic point clouds in 6DoF VR. In order to perform the instrumentation required to address this research question we provide the following contributions:

Contributions

The results presented in this chapter provide a comparison of the state-of-the-art V-PCC standard codec and the real-time MPEG Anchor codec across eight dynamic point cloud sequences and 4 rate points. The results indicate that neither codec is able to achieve transparent quality even at the highest bitrates. The codecs achieve statistically similar visual qualities at higher bitrates with significant variation amongst the different sequences tested. The contributions of this chapter can be summarized as:

1. Present the first approach to subjective evaluation of dynamic point clouds that was employed as part of the standardisation effort for point cloud compression

2. Provide a first evaluation of the quality of highly realistic digital humans represented as dynamic point clouds in immersive viewing conditions.

3. Provide quantitative subjective results about the perceived quality of the contents, along with qualitative insights on what is important for users in interacting with digital humans in VR

These contributions are presented in **Chapter 2** and are based on:

1. V. Baroncini, P. Cesar, E. Siahaan, I. Reimat, **S. Subramanyam**, "Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression." ISO/IEC JTC1/SC29/WG11 M41786 (2017).

2. **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," 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2020, pp. 127-136, doi: 10.1109/VR46266.2020.00031.

Methodology

In order to perform the evaluation, we developed a framework to render dynamic point cloud content in VR and define a quality evaluation protocol suitable for this environment. We employed the Absolute Category Rating with Hidden References (ACR-HR) based on ITU-T Recommendations P.910 [73] and conducted a user study to assess the quality of dynamic point clouds compressed using standard codecs for offline and real-time compression of point clouds. Existing protocols did not consider the dynamic nature of the point clouds, focused on one type of dataset and did not take into account VR viewing conditions. In addition, we also conducted a semi-structured interview to collect qualitative insights on the factors that are important to user's while performing the quality evaluation, a comparison of 3DoF and 6DoF viewing conditions and the difficulties faced by participants while performing the assessment. We also recorded a dataset of user navigation patterns in the scene defined by the position and orientation of their viewport.

1.5.2. ON USER-CENTERED TILING OF DYNAMIC POINT CLOUDS FOR ADAPTIVE STREAMING

Research question 2 investigates the validity of user-centred, real-time adaptive streaming through both objective and subjective quality evaluation. This can be used to optimize the delivery of a single dynamic point cloud object, through independently decodable tiles. To address this research question we provide the following contributions:

CONTRIBUTIONS

The results presented in this chapter demonstrate the validity of user-centred tiling of dynamic point clouds. We observe significant bitrate gains of upto 57% with respect to a network adaptive approach using the same point cloud codec based on objective evaluation. We demonstrate the significant impact of navigation paths on objective evaluation and provide a dataset of navigation patterns for future research. We observe significant bitrate gains of upto 65% to deliver comparable subjective quality using a novel tile selection strategy. The contributions of this chapter can be summarized as:

1. Propose a novel tile selection strategy to optimize the delivery of independently decodable point cloud tiles over bandwidth-limited networks
2. Demonstrate the validity of user-centred, real-time adaptive streaming to optimize the delivery of a single dynamic point cloud objects, through objective and subjective quality evaluation.
3. Provide a user-centred approach to collect a dataset of users' navigation paths viewing the point cloud with 6DoF, which can be used to evaluate adaptive streaming approaches.

These contributions are presented in **Chapter 3** and are based on:

1. **S. Subramanyam**, I. Viola, A. Hanjalic, and P. Cesar. 2020. User-centred Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 3669–3677. doi: 10.1145/3394171.3413535
2. **S. Subramanyam**, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic and P. Cesar. 2022 Subjective QoE Evaluation of User-Centered Adaptive Streaming of Dynamic Point Clouds. 2022 14th International Conference on Quality of Multimedia Experience (QoMEX)

METHODOLOGY

To perform the objective evaluation, we developed a framework to render tiled streams of point clouds and perform the adaptation in real-time. We define a real-time tiling approach and employ several bit rate allocation schemes from the literature along with a novel tile allocation scheme. We followed a user-centered approach to collect and employ a dataset of user navigation paths in 6DoF while viewing dynamic point cloud sequences. We then played back the sequences using the navigation paths and recorded screenshots at each viewport location. The quality was evaluated using image distortion metrics to measure the validity of tiled adaptive point cloud streaming. We then performed a subjective evaluation to validate the impact of tiled adaptive point cloud streaming on end-user QoE through a user study.

1.5.3. ON ADAPTIVE STREAMING OF POINT CLOUDS FOR VR REMOTE COMMUNICATION

Research question 3 investigates optimization strategies for delivering point cloud reconstructions in a real-time live remote communication environment. To address this research question we provide the following contributions:

CONTRIBUTIONS

In order to evaluate the impact of user-centred tiled adaptive streaming of point clouds we first describe a VR communication system with an end-to-end delivery pipeline and profile system performance under different streaming conditions. We perform an evaluation on the impact of adaptive streaming on quality of communication, task performance experience and perceived quality. The system design presented in this chapter is able to achieve similar visual quality with tiled adaptive streaming at 14 megabits per second as compared to uncompressed streaming requiring approximately 300 megabits per second. The results also demonstrate statistically significant gains to the quality of interaction and system resource consumption by employing tiled adaptive streaming. The contribution of this chapter can be summarized as:

1. Propose a novel protocol to evaluate the impact of delivery optimizations in VR remote communication

2. Provide an evaluation of user-adaptive point cloud streaming in terms of quality of communication, visual quality, task-related experience and system resource consumption

3. Present a software architecture for a real-time two-user VR remote communication application with live-captured point cloud user reconstructions

This contribution is presented in **Chapter 4** and is based on:

1. **S. Subramanyam**, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic and P. Cesar. Evaluating the Impact of Tiled User-Adaptive Real-Time Point Cloud Streaming on VR Remote Communication. Proceedings of the 30th ACM International Conference on Multimedia, October 10-14, 2022, Lisboa, Portugal

2. J. Jansen, **S. Subramanyam**, R. Bouqueau, G. Cernigliaro, M. Cabre, F. Perez, and P. Cesar. 2020. A pipeline for multiparty volumetric video conferencing: transmission of point clouds over low latency DASH. In Proceedings of the 11th ACM Multimedia Systems Conference (MMSys '20). Association for Computing Machinery, New York, NY, USA, 341–344. https://doi.org/10.1145/3339825.3393578

METHODOLOGY

We developed a two-user adaptive VR remote communication pipeline with live captured point cloud reconstructions. We propose a novel evaluation methodology using a training task and confederate users to serve as trainers. Participants were taught a neck exercise in VR and were asked to perform the exercise at the end. We employed questionnaires from the literature to gather data on subjective visual quality, quality of communication/ interaction as well as subjective task-related performance. In addition, we gathered data on system performance and resource consumption.

1.5.4. CONCLUSIONS AND OUTLOOK

The conclusion of the thesis and the outlook for future research in the field are presented in **Chapter 5**. This chapter starts by returning to the research questions raised in chapter 1. We then summarize the discussion items and the publicly available resources that were produced in the course of conducting the research presented in this thesis. This includes datasets and demonstrations created using the delivery pipeline described in the previous chapter. The demonstrations showcase some of the use cases and applications that can benefit from the research presented in this thesis. Based on insights from the discussion and observed limitations of the research reported in the thesis, we point out possible directions for future work and conclude the thesis.

This chapter presents the following publicly available datasets:

1. I. Reimat, E. Alexiou, J. Jansen, I. Viola, **S. Subramanyam**, and P. Cesar. 2021. CWIPC-SXR: Point Cloud dynamic human dataset for Social XR. In Proceedings of the 12th ACM Multimedia Systems Conference (MMSys '21). Association for Computing Machinery, New York, NY, USA, 300–306. https://doi.org/10.1145/3458305.3478452

2. A. Chatzitofis, L. Saroglou, P. Boutis, P. Drakoulis, N. Zioulis, **S. Subramanyam**, B. Kevelham, C. Charbonnier, P. Cesar, D. Zarpalas, S. Kollias, P. Daras, "HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media," in IEEE Access, vol. 8, pp. 176241-176262, 2020, doi: 10.1109/ACCESS.2020.3026276.

This chapter also presents the following application scenarios:

1. I. Reimat, Y. Mei, E. Alexiou, J. Jansen, J. Li, **S. Subramanyam**, I. Viola, J. Oomen, P. Cesar, Mediascape XR: A Cultural Heritage Experience in Social VR, Proceedings of the 30th ACM International Conference on Multimedia, October 10-14, 2022, Lisboa, Portugal

2. A. Revilla, S. Zamarvide, I. Lacosta, F. Perez, J. Lajara, B. Kevelham, V. Juillard, B. Rochat, M. Drocco, N. Devaud, O. Barbeau, C. Charbonnier, P. de Lange, J. Li, Y. Mei, K. Ławicka, J. Jansen, N. Reimat, **S. Subramanyam**, P. Cesar, "A Collaborative

VR Murder Mystery using Photorealistic User Representations," 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2021, pp. 766-766, doi: 10.1109/VRW52623.2021.0266.

REFERENCES

[1] U. D. of Energy Advanced Research Projects Agency Energy (ARPA-E), *Facsimile appearance to create energy savings (faces),* (2017).

[2] J. Li, Y. Kong, T. Röggla, F. De Simone, S. Ananthanarayan, H. de Ridder, A. El Ali, and P. Cesar, *Measuring and understanding photo sharing experiences in social virtual reality,* in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 1–14.

[3] E. Gent, *Forget video conferencing—host your next meeting in vr,* (2020).

[4] D. S. McNamara and J. N. Bailenson, *Nonverbal overload: A theoretical argument for the causes of zoom fatigue,* Technology, Mind, and Behavior **2** (2021), 10.1037/tmb0000030, https://tmb.apaopen.org/pub/nonverbal-overload.

[5] A. Yassien, P. ElAgroudy, E. Makled, and S. Abdennadher, *A design space for social presence in vr,* in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Association for Computing Machinery, New York, NY, USA, 2020).

[6] Z. Yang, W. Wu, K. Nahrstedt, G. Kurillo, and R. Bajcsy, *Enabling multi-party 3d tele-immersive environments with <i>viewcast</i>,* ACM Trans. Multimedia Comput. Commun. Appl. **6** (2010), 10.1145/1671962.1671963.

[7] H. Fuchs, A. State, and J. Bazin, *Immersive 3d telepresence,* Computer **47**, 46 (2014).

[8] 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 (2016).

[9] J. Jansen, S. Subramanyam, R. Bouqueau, G. Cernigliaro, M. Martos Cabré, F. Pérez, and P. Cesar, *A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH,* in *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20 (ACM, New York, NY, USA, 2020).

[10] S. N. B. Gunkel, R. Hindriks, K. M. E. Assal, H. M. Stokking, S. Dijkstra-Soudarissanane, F. t. Haar, and O. Niamut, *Vrcomm: An end-to-end web system for real-time photorealistic social vr communication,* in *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 65–79.

[11] A. D. Wilson and H. Benko, *Projected augmented reality with the roomalive toolkit,* in *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, ISS '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 517–520.

[12] J. Lawrence, D. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong, *Project starline: A high-fidelity telepresence system,* ACM Trans. Graph. **40** (2021), 10.1145/3478513.3480490.

[13] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi, *Holoportation: Virtual 3d teleportation in real-time,* in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 741–754.

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

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

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

[17] J. Constine, *Facebook animates photo-realistic avatars to mimic VR users' faces,* (2018).

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

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

[20] R. Mekuria, P. Cesar, I. Doumanis, and A. Frisiello, *Objective and subjective quality assessment of geometry compression of reconstructed 3D humans in a 3D virtual room,* Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, 95991M (2015).

[21] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch, *The effect of avatar realism in immersive social virtual realities,* in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17 (Association for Computing Machinery, New York, NY, USA, 2017).

[22] A. Doumanoglou, D. S. Alexiadis, D. Zarpalas, and P. Daras, *Toward real-time and efficient compression of human time-varying meshes,* IEEE Transactions on Circuits and Systems for Video Technology **24**, 2099 (2014).

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

[24] J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, and H. Hellwagner, *Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression,* in *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19 (Association for Computing Machinery, New York, NY, USA, 2019) pp. 2405–2413.

[25] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, *et al.*, *Emerging MPEG Standards for Point Cloud Compression,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 133 (2019).

[26] 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*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 300–306.

[27] E. Pavez, P. A. Chou, R. L. de Queiroz, and A. Ortega, *Dynamic polygon cloud compression,* Microsoft Research Technical Report (2016).

[28] R. L. de Queiroz and P. A. Chou, *Compression of 3d point clouds using a region-adaptive hierarchical transform,* IEEE Transactions on Image Processing **25**, 3947 (2016).

[29] R. Ghaznavi-Youvalari, A. Zare, H. Fang, A. Aminlou, Q. Xie, M. M. Hannuksela, and M. Gabbouj, *Comparison of hevc coding schemes for tile-based viewport-adaptive streaming of omnidirectional video,* in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)* (2017) pp. 1–6.

[30] C. Concolato, J. Le Feuvre, F. Denoual, F. Mazé, E. Nassor, N. Ouedraogo, and J. Taquet, *Adaptive streaming of hevc tiled videos using mpeg-dash,* IEEE Transactions on Circuits and Systems for Video Technology **28**, 1981 (2018).

[31] O. A. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual, and S. Y. Lim, *Mpeg dash srd: Spatial relationship description,* in *Proceedings of the 7th International Conference on Multimedia Systems*, MMSys '16 (ACM, New York, NY, USA, 2016) pp. 5:1–5:8.

[32] MPEG, *Iso/iec 23000-20. omnidirectional media application format (omaf),* ISO/IEC JTC1/SC29 WG11 (2017).

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

[34] E. Alexiou and T. Ebrahimi, *On the performance of metrics to predict quality in point cloud representations,* in *Applications of Digital Image Processing XL*, Vol. 10396,

edited by A. G. Tescher, International Society for Optics and Photonics (SPIE, 2017) pp. 282 – 297.

[35] 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)* (2018) pp. 1–6.

[36] 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).

[37] MPEG3DG and Requirements, *Call for proposals for point cloud compression,* ISO/IEC JTC1/SC29 WG11 N16732, Geneva, CH (2017).

[38] S. Perry, H. P. Cong, L. A. da Silva Cruz, J. Prazeres, M. Pereira, A. Pinheiro, E. Dumic, E. Alexiou, and T. Ebrahimi, *Quality evaluation of static point clouds encoded using mpeg codecs,* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020) pp. 3428–3432.

[39] 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)* (2017) pp. 1–6.

[40] A. Javaheri, C. Brites, F. M. B. Pereira, and J. M. Ascenso, *Point Cloud Rendering after Coding: Impacts on Subjective and Objective Quality,* IEEE Transactions on Multimedia , 1 (2020).

[41] R. B. Rusu, *3d is here: Point cloud library,* Robotics and Automation (ICRA), 2011 IEEE International Conference (2011).

[42] E. Dumic, F. Battisti, M. Carli, and L. A. da Silva Cruz, *Point cloud visualization methods: a study on subjective preferences,* in *2020 28th European Signal Processing Conference (EUSIPCO)* (2021) pp. 595–599.

[43] 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)* (2020) pp. 1–6.

[44] 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)* (2017) pp. 1–6.

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

[46] E. Alexiou, N. Yang, and T. Ebrahimi, *PointXR: A Toolbox for Visualization and Subjective Evaluation of Point Clouds in Virtual Reality,* in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)* (2020) pp. 1–6.

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

[48] A. Nguyen and B. Le, *3d point cloud segmentation: A survey,* in *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (2013) pp. 225–230.

[49] M. Hosseini and C. Timmerer, *Dynamic Adaptive Point Cloud Streaming,* in *Proceedings of the 23rd Packet Video Workshop*, PV '18 (ACM, New York, NY, USA, 2018) pp. 25–30.

[50] M. Hosseini,
Adaptive rate allocation for view-aware point-cloud streaming, (2017), 10.13140/RG.2.2.23436.26244.

[51] J. Park, P. A. Chou, and J. Hwang, *Rate-Utility Optimized Streaming of Volumetric Media for Augmented Reality,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 149 (2019).

[52] L. He, W. Zhu, K. Zhang, and Y. Xu, *View-dependent streaming of dynamic point cloud over hybrid networks,* in *Advances in Multimedia Information Processing – PCM 2018* (Springer International Publishing, Cham, 2018) pp. 50–58.

[53] J. Li, C. Zhang, Z. Liu, W. Sun, and Q. Li, *Joint communication and computational resource allocation for qoe-driven point cloud video streaming,* (2020).

[54] K. Lee, J. Yi, Y. Lee, S. Choi, and Y. M. Kim, *Groot: A real-time streaming system of high-fidelity volumetric videos,* in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (Association for Computing Machinery, New York, NY, USA, 2020).

[55] B. Han, Y. Liu, and F. Qian, *Vivo: Visibility-aware mobile volumetric video streaming,* in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (Association for Computing Machinery, New York, NY, USA, 2020).

[56] Z. Liu, J. Li, X. Chen, C. Wu, S. Ishihara, Y. Ji, and J. Li, *Fuzzy logic-based adaptive point cloud video streaming,* IEEE Open Journal of the Computer Society **1**, 121 (2020).

[57] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, *Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units,* in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14 (Association for Computing Machinery, New York, NY, USA, 2014) p. 637–644.

[58] G. Cernigliaro, M. Martos, M. Montagud, A. Ansari, and S. Fernandez, *PC-MCU: Point Cloud Multipoint Control Unit for Multi-User Holoconferencing Systems,* in *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* (Association for Computing Machinery, New York, NY, USA, 2020) p. 47–53.

[59] ITU-T P.1301, *P.1301 : Subjective quality evaluation of audio and audiovisual multiparty telemeetings,* International Telecommunication Union (2017).

[60] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar, *Towards individual qoe for multiparty videoconferencing,* IEEE Transactions on Multimedia **20**, 1781 (2018).

[61] ITU-T P.920, *P.920 : Interactive test methods for audiovisual communications,* International Telecommunication Union (2000).

[62] H. J. Smith and M. Neff, *Communication behavior in embodied virtual reality,* in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (Association for Computing Machinery, New York, NY, USA, 2018) p. 1–12.

[63] A. Toet, T. Mioch, S. N. Gunkel, O. Niamut, and J. B. van Erp, *Holistic framework for quality assessment of mediated social communication,* (2021).

[64] M. Slater, M. Usoh, and A. Steed, *Depth of presence in virtual environments,* Presence: Teleoper. Virtual Environ. **3**, 130–144 (1994).

[65] B. G. Witmer and M. J. Singer, *Measuring presence in virtual environments: A presence questionnaire,* Presence: Teleoper. Virtual Environ. **7**, 225–240 (1998).

[66] J. Kangas, S. K. Kumar, H. Mehtonen, J. Järnstedt, and R. Raisamo, *Trade-off between task accuracy, task completion time and naturalness for direct object manipulation in virtual reality,* Multimodal Technologies and Interaction **6** (2022), 10.3390/mti6010006.

[67] ITU-T P.QXM, *QoE assessment of eXtended Reality (XR) meetings,* International Telecommunication Union (2022).

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

[69] 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)* (IEEE, 2017) pp. 1–6.

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

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

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

[73] ITU-T P.910, *Subjective video quality assessment methods for multimedia applications,* International Telecommunication Union (2008).

2

EVALUATING THE QUALITY OF DYNAMIC POINT CLOUD HUMAN RECONSTRUCTIONS IN IMMERSIVE ENVIRONMENTS

This chapter explores the subjective quality evaluation of dynamic point cloud reconstructions of humans. We describe existing subjective evaluation methodologies and present the approach employed in the evaluation of a new point cloud compression standard by MPEG. We then propose a protocol to perform the assessment in a realistic immersive environment. We present the first work performing subjective evaluation of dynamic point clouds in VR. The contributions of this chapter include quantitative subjective results comparing the performance of a state-of-the-art point cloud compression standard and a low complexity real-time point cloud codec across a wide range of target bitrates spanning 3 to 140 Mbps. We also provide qualitative insights on the factors important to users for the quality evaluation task. This work is meant to provide an understanding of the perceived quality of dynamic point cloud human reconstructions in immersive virtual environments that can be used to drive adaptive delivery optimizations for VR remote communication.

This chapter is based on the following :

1. ***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," 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2020, pp. 127-136, doi: 10.1109/VR46266.2020.00031.*
2. *V. Baroncini, P. Cesar, E. Siahaan, I. Reimat, **S. Subramanyam**, "Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression." ISO/IEC JTC1/SC29/WG11 M41786, MPEG: 120th Meeting, Macao, Oct. 2017*

2.1. Introduction

Recent advances in capturing, media processing, and 3D rendering technologies make VR/AR applications popular for mass consumption [1]. Avatars are a core part of VR applications like social communication [2], sports training [3], or healthcare [4]. A major line of scientific work has focused on how to make such avatars more realistic, interactive, and autonomous [5–7]. In this thesis, we focus on point clouds as a suitable representation for digital humans based on teleportation principles [8]. In this chapter, we establish a protocol and methodology to evaluate the quality of point cloud human reconstructions. This protocol will then be used in subsequent chapters to investigate point cloud delivery optimizations in the context of VR remote communication.

In this new media landscape, point clouds are becoming commonplace due to their simplicity and versatility. Still, the size of dense point clouds is significant (a frame of roughly 1M points takes around 19-20 MBytes), which needs compression techniques before transmission. Given current advances in technology, real-time delivery of point clouds is becoming a realistic alternative; focusing the attention of the research community [9] and industry [10] in encoding and transmission. Still, given the massive number of points per representation, decisions need to be taken regarding the delivery (type of encoder, bit-rate) to ensure an acceptable quality of experience depending on the viewing conditions (3DoF, 6DoF). In this chapter, we provide an exhaustive quality comparison between different encoding configurations of digital humans, represented as point clouds. By investigating the differences in quality, we provide insights into optimising the delivery for both playback and real-time communication.

The contributions of this chapter are three-fold:

- We present the subjective evaluation methodology used during the MPEG standardization activity for point cloud compression.
- We propose a protocol for evaluating the quality of highly realistic digital humans represented as dynamic point clouds in immersive viewing environments. We then present the results of a user study employing this protocol to assess the performance of two point cloud codecs.
- We provide quantitative subjective results about the perceived quality of the contents, along with qualitative insights on what is important for users in interacting with digital humans in VR and the factors that drive the quality score.

These contributions help us address the research question *R1.1 (What is the impact of immersive viewing environments on subjective quality assessment?)*. We propose and employ a novel evaluation protocol based on realistic viewing conditions in two viewing conditions, in 3- and 6- Degrees of Freedom (DoF). We perform the quality evaluation in these viewing conditions through a user study. We also address the research question *R1.2 (What factors do users rely on while performing subjective quality assessment of dynamic point clouds?)*. We interview participants to provide qualitative results in order to identify the underlying factors driving the quality score.

The results are based on an experiment with 52 participants, evaluating 72 stimuli based on eight dynamic point cloud sequences. Each point cloud sequence was compressed in four bit-rates, using two types of compression techniques. These 72 stimuli were evaluated in two viewing conditions (3DoF and 6DoF). The data gathered include rating scores, presence questionnaires, simulator sickness reports, and time spent watching the content. The results indicate that, while bit-rate savings can be obtained by choosing one compression solution over another, visually lossless compression has not been fully achieved by the algorithms under evaluation, even at rather large bit-rates. Moreover, the choice of content can have an impact on how users rate its quality, influencing the discriminating power of the selected protocol. These results will help in configuring pipeline components for the delivery of point clouds for real-time transmission and have implications for ongoing research and standardisation work regarding the underlying compression technology.

In the remainder of this chapter, we first discuss the related work, highlighting the contributions beyond the state of the art. We then present the methodology of the experiment, detailing the dataset used and the protocol. This is followed by the results accompanied by an exhaustive statistical analysis. We then discuss the larger findings from the experiment and describe the implications for future research followed by the conclusion.

2.2. RELATED WORK

Capturing and displaying volumetric videos is becoming feasible [11, 12]. Point clouds are frequently used as a data format for volumetric video in augmented reality (AR) and virtual reality (VR) applications. Point clouds collate a large number of geo-referenced points to represent humans or objects in 3D. The color information can be provided with each point [13]. To visualize 3D content sufficiently, the number of points must be high, which results in large size and increases the difficulty to store and transmit the point clouds. To support low latency transmission in AR/VR applications within a limited bandwidth, compression is necessary. However, it remains challenging to measure and predict the acceptable quality of compressed point clouds.

There is a growing interest on subjective quality assessment of point clouds rendered on 2D displays. Zhang et al. [14] evaluated the quality degradation effect of resolution, shape and color on static point clouds. The results indicate that resolution is almost linearly correlated with the perceived quality, and color has less impact than shape on the perceived quality. Zerman et al. [13] compressed two dynamic human point clouds using a state-of-the-art algorithm [15], and assessed the effects of this algorithm and input point counts on the perceived quality. Their results showed that no direct correlation was found between human viewers' quality ratings and input point counts. In a recent study [16], a protocol to conduct subjective quality evaluations and benchmark objective quality metrics were proposed. The viewers passively assessed the quality of a set of static point clouds, as animations with pre-defined movement path. In a comprehensive work by Alexiou et al. [17], the entire set of emerging point cloud compression encoders developed in the MPEG committee were evaluated through a series of subjective quality assessment experiments. Nine static models, including both humans and objects, were used in the experiments. The experiments provided insights regarding the performance of the encoders and the types of degradation they introduce.

Only a limited number of point cloud quality assessment studies have been conducted in immersive environments. Mekuria et al. [9] evaluated the subjective quality of their codec performance in a realistic 3D tele-immersive system, in which users were represented as 3D avatars and/or 3D dynamic point clouds, and could navigate in the virtual space using mouse cursor in a desktop setting. Several aspects of quality, such as level of immersiveness, togetherness, realism, quality of motion, were considered. Alexiou and Ebrahimi [18] proposed the use of AR to subjectively evaluate the quality of colorless point cloud geometry. Tran et al. [19] suggested that, in case of evaluating video quality in an immersive setup, aspects such as cybersickness and presence should not be overlooked.

2.2.1. Point cloud compression

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

Research into point cloud compression can be broadly divided into two categories. The first is based on signal processing, Zhang et al. [20] proposed a method to compress point cloud attributes using a graph Fourier transform. They assume that an octree has been created and separately coded for geometry prior to coding attributes. De Queroz and Chou [21] used a region adaptive hierarchical transform to use the colors of nodes in lower levels of the octree to predict the colors of nodes in the next level. As these approaches require expensive computations of graph laplacians, they are not suitable for dynamic sequences in real-time applications. The second category of point cloud codecs are based on extending legacy solutions from image and video compression. Intra Frame coding in octrees can be achieved by entropy coding the occupancy codes, as shown in [9]. The authors then compress the color attributes by mapping them to a 2D grid and using legacy JPEG image compression.

In 2017, MPEG started a standardization activity to determine a new standard codec for point clouds, to be launched in 2020. They used the codec created by Mekuria et al. [9] as an anchor to evaluate proposals. To encode dynamic point cloud sequences MPEG provides two verification models [10], Geometry-PCC for point clouds with a sparse distribution, and Video-PCC for dense point clouds. V-PCC is based on leveraging existing 2D video codecs to compress point cloud geometry and attributes.

2.3. MPEG Point Cloud Quality Assessment

MPEG launched the point cloud compression standardization activity in 2017 with the goal of building an open standard for efficiently representing 3D point clouds. The activity began with a call for proposals (CfP) for three categories of point clouds; static surfaces, dynamic sequences and dynamically acquired LiDAR scans. The codec proposed by Mekuria et al. [22] was selected as an Anchor codec and several companies responded to the CfP with proposals that aimed to outperform the Anchor. The CfP included a novel dataset of dynamic human reconstructions [23] consisting of four 10-second sequences.

In order to select compression technologies that could serve as candidates for future standards, proposals were evaluated using both objective quality evaluation and a novel subjective evaluation approach [24].

We performed the subjective evaluation at our premises in collaboration with other experts. We selected three static frames and three dynamic sequences from the dataset to restrict the number of stimuli that needed to be tested and recruited 22 participants. To prepare the dataset we recorded 10-second video clips of static and dynamic point cloud contents using a rendering tool that allowed rotations and translations along the three Cartesian axes. The video clips were recorded using fixed navigation paths designed to view the content from a wide range of angles and distances. In this manner, uncompressed video clips were recorded for decoded point clouds from each proposal at four target bit-rates in the YUV 444 progressive 10 bit video files.

We selected the absolute category rating test method recommended in ITU-T P.910 [25] due to the large number of stimuli that had to be evaluated. Participants were asked to view the video clips one at a time in groups of two or three at a fixed distance in front of a 4K consumer TV set. The ambient light was maintained at around 30 nits to prevent distractions while allowing participants to fill out scoring sheets. The results of this evaluation were published in [24] where one proposal clearly had the highest performance across all dynamic content. This proposal later became the MPEG V-PCC standard codec that is expected to become an indispensable component for representing, delivering and visualizing 3D point cloud objects [26]. The objective evaluation results were also found to be highly correlated with these results.

From our experience participating in this standardization activity, we learned that the double stimulus methodology is unsuitable for this modality due to difficulties in rendering two dynamic point cloud sequences simultaneously and ensuring a fair and consistent comparison across stimuli. In addition, the navigation paths used do not reflect actual user behaviour while interacting with human point cloud reconstructions. These paths were found to have a large effect on the collected scores and are not realistic for target applications such as VR remote communication. In the next section, we present our subsequent quality evaluation experiment conducted in an immersive viewing environment with no restrictions on user movement.

2.4. METHODOLOGY

2.4.1. DATASET PREPARATION

A dataset of dynamic point cloud sequences was used from the MPEG repository. All sequences were clipped to five seconds and sampled at 30 frames per second. This included point cloud sequences [23, 27] captured using photogrammetry (Longdress, Loot, Red and black, Soldier are shown in figure 2.1) and one sequence of a synthetic character sampled from an animated mesh (Queen). Four additional point cloud sequences; Manfred, Despoina, Sarge (shown in Figure 2.1) and Rachel were added for the evaluation. These sequences were created using motion captured animated mesh sequences.

Keyframes were selected at 30 frames per second and extracted along with the associated mesh materials. Particular care was put in ensuring the selected sequences have the characters facing the user and speaking in their general direction. Then, 1 million points

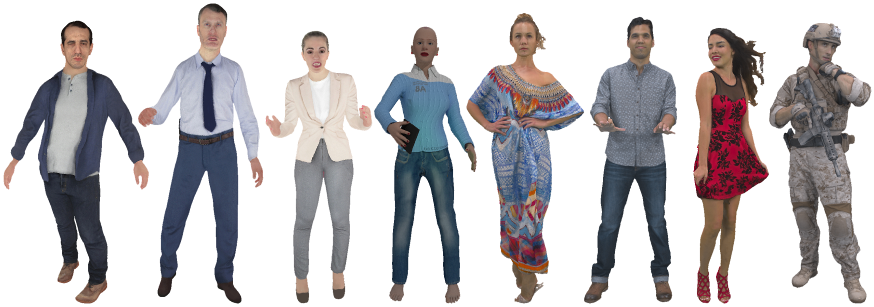


Figure 2.1: Sequences used for the test, from left to right: Manfred, Sarge, Despoina, Queen, Longdress, Loot, Red and black, Soldier

were randomly sampled, independently per key frame to create a consistent groundtruth dataset. The points are sampled from the mesh surface with a probability proportional to the area of the underlying mesh face. This was done to ensure no direct point correspondences across point cloud frames, to mimic realistic acquisition and maintain consistency with the rest of the dataset. The point clouds sampled from meshes were used in Test T1 and the point clouds captured using photogrammetry were used in test T2. The X, Y, Z coordinates of each point is represented using an unsigned integer, as is required for the current version of the V-PCC software. Colors are encoded as 8bits per color in the RGB color space.

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

We then use the MPEG anchor codec [9] in an all intra configuration, and match the bit-rates per sequence and rate point (R1-R4) with a tolerance of 10%, as defined in the MPEG call for proposals. The codec was selected as it has a significantly lower encode and decode time and is suitable for real-time applications, as demonstrated by the authors. We use an octree depth from 7 to 10 for the rate points R1 to R4 respectively. The highest possible JPEG quantization parameter values were then chosen per sequence, while meeting the target bit rate set using V-PCC.

2.4.2. Experiment setup

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

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

2.4.3. SUBJECTIVE METHODOLOGY

To perform the experiments, the subjective methodology Absolute Category Rating with Hidden References (ACR-HR) was selected, according to ITU-T Recommendations P.910 [25]. Participants were asked to observe video sequences depicting digital humans one at a time, and rate the corresponding visual quality on a scale from 1 to 5 (*1-Bad*, *2-Poor*, *3-Fair*, *4-Good*, and *5-Excellent*). Each rating was performed independently and reference uncompressed versions of each digital human were included in the test and presented as any other stimulus.

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

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

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

After each view condition, participants were requested to fill in the Igroup Presence Questionnaire (IPQ) [28] on a 1-7 discrete scale (1=fully disagree to 7=totally agree) and Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [29]. IPQ has three subscales, namely Spatial Presence (SP), Involvement (INV) and Experienced Realism (REAL), and one additional general item (G) not belonging to a subscale, which assesses the general "sense of being there", and has high loadings on all three factors, with an especially strong loading on SP [28]. SSQ was developed to measure cybersickness in computer simulation and was derived from a measure of motion sickness [29]. For both T1 and T2, after the two viewing conditions, participants were interviewed to 1) compare their experiences of assessing quality in 3DoF and 6DoF, and 2) reflect on the factors they considered when assessing the quality.

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

2.4.4. Data analysis

Outlier detection was performed separately for each test T1 and T2, according to ITU-T Recommendations P.913 [30]. The recommended threshold values $r_1 = 0.75$ and $r_2 = 0.8$ were used. One outlier was found in test T1, and the corresponding scores were discarded. No outliers were found in the scores collected for test T2.

After outlier detection, the Mean Opinion Score (MOS) was computed for each stimulus, independently per viewing condition. The associated 95% Confidence Intervals (CIs) were obtained assuming a Student's t-distribution. Additionally, the Differential MOS (DMOS) was obtained by applying HR removal, following the procedure described in ITU-T Recommendations P.913 [30].

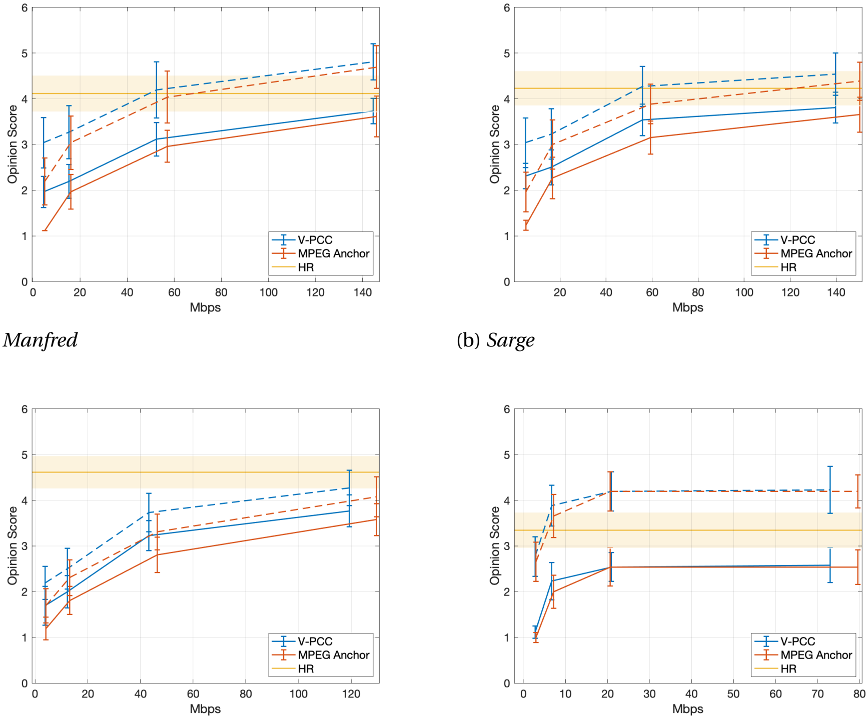
Non-parametric statistical analysis was applied to understand whether statistical differences could be found among variables, using the MATLAB Statistics and Machine Learning Toolbox, along with the ARTool package in R [31].

2.5. Results

2.5.1. Subjective quality assessment

Figures 2.2, 2.3 and 2.4 2.5 shows the results of the subjective quality assessment of the contents comprising test T1 and test T2, respectively, for both 3DoF and 6DoF viewing conditions. In particular, the MOS scores associated with the compressed contents are shown with solid lines, along with relative CIs, whereas the dashed lines represent the respective DMOS scores. The HR scores for each content are represented with a solid line to indicate the mean, and a shaded plot for the corresponding CIs.

To assess whether significant differences could be found between the two visual conditions under test, we ran a Wilcoxon signed-rank test on the scores obtained in the two DoF scenarios. The Wilcoxon test was chosen as the gathered data was not found to be normally distributed, according to the Shapiro-Wilk normality test ($W = 0.90$, $p < .001$ and $W = 0.91$, $p < .001$ for tests T1 and T2, respectively). Results of the Wilcoxon signed-



(a) *Manfred*

(b) *Sarge*

(c) *Despoina*

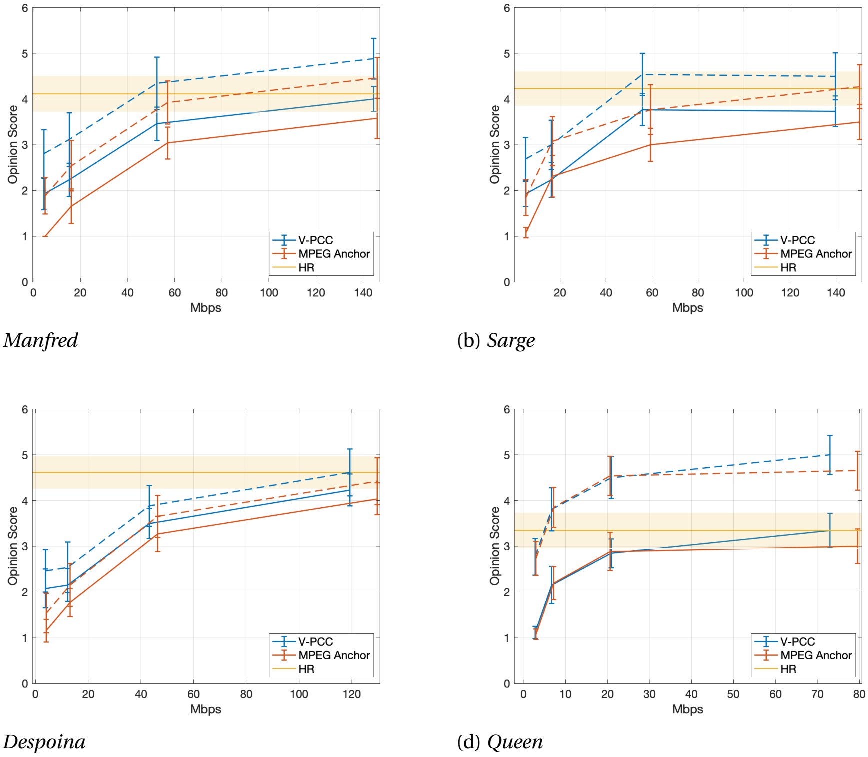
(d) *Queen*

Figure 2.2: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps in the 3DoF viewing condition in test T1. HR scores are shown using a shaded yellow plot.

rank test showed statistical significance for DoF for test T1 ($Z = 2.97$, $p = 0.0029$, $r = 0.07$), whereas for test T2, no significance was found ($Z = -1.96$, $p = 0.0502$, $r = 0.05$). Values seems to indicate an effect of the DoF in test T1; however the small r-value indicates that while the effect apparently exists, it is small.

It can be observed that codec V-PCC has generally a more favorable performance with respect to the MPEG anchor. This is especially evident for the contents acquired through photogrammetry (see Fig. 2.5), for which the gap among the two codecs is more pronounced. Wilcoxon signed-rank test confirmed statistical significance for the two codecs (T1: $Z = 9.87$, $p < .001$, T2: $Z = 20.18$, $p < .001$), albeit with different effect sizes between test T1 and T2 ($r = 0.24$ and $r = 0.50$, respectively).

A Friedman rank test performed on the scores revealed a significant effect of the content on the final scores, for both sets of contents (T1: $\chi^2 = 57.38$, $p < .001$, T2: $\chi^2 = 17.31$, $p < .001$). Table 3.3 shows the results of the post-hoc test conducted using Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/6$). Contents *Manfred*, *Sarge* and *Despoina* all show statistical significance with respect to content *Queen* ($p < .001$, $r > 0.20$ for all pairs). Statistical significance has also been observed between content *Man-*



(a) *Manfred*

(b) *Sarge*

(c) *Despoina*

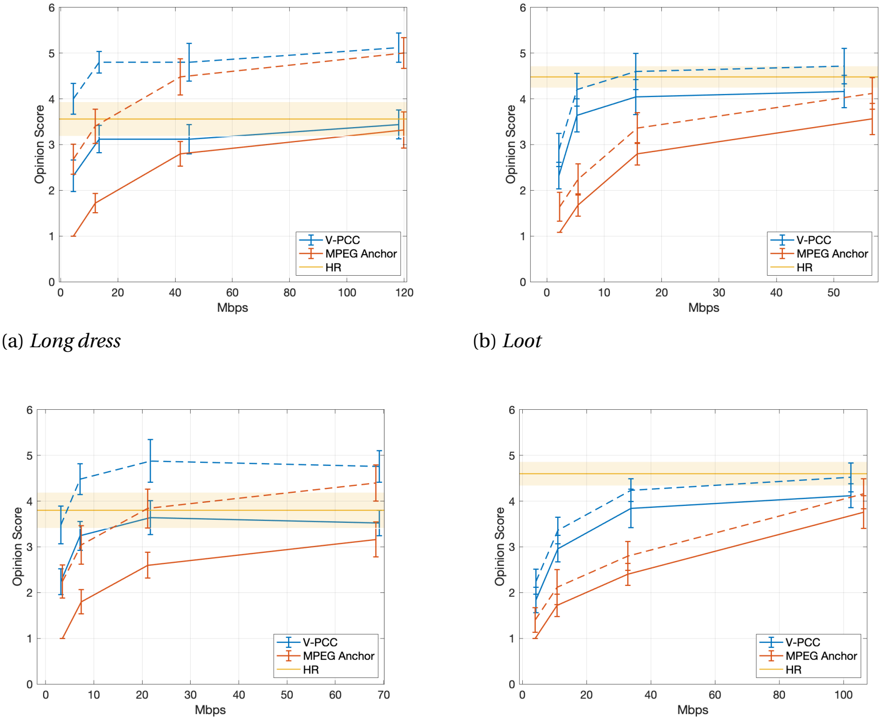
(d) *Queen*

Figure 2.3: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps in the 6DoF viewing condition in test T1. HR scores are shown using a shaded yellow plot.

fred and *Sarge*, albeit with a smaller effect size ($p < .001$, $r = 0.12$). For contents acquired through photogrammetry, statistical significance was found between contents *Long dress* and *Loot*, and *Loot* and *Red and black* ($p < .001$, $r > 0.20$ in both cases), as well as between contents *Long dress* and *Soldier*, *Loot* and *Soldier* ($p < .001$, $r > 0.10$), and *Red and black* and *Soldier* ($p = 0.0019$, $r = 0.10$). Results corroborate our previous statements on how contents *Long dress* and *Red and black* appeared to be given different scores with respect to contents *Loot* and *Soldier*.

We also ran a Friedman rank test on the scores to assess whether the selected bit-rates were showing statistical significance. Results confirmed that the bit-rates have a significant effect for both tests (T1: $\chi^2 = 682.29$, $p < .001$, T2: $\chi^2 = 667.39$, $p < .001$). Post-hoc analysis using Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/6$), shown in Table 2.2 further confirmed that all pairwise comparisons were statistically significant, for both test T1 and T2 ($p < .001$, $r > 0.30$ for all pairs).

In order to further analyze the effect of DoF conditions, contents, codecs and bit-rates, and relative interactions, on the gathered scores, we fitted a full linear mixed-effects model on the data, accounting for randomness introduced by the participants.



(a) *Long dress*

(b) *Loot*

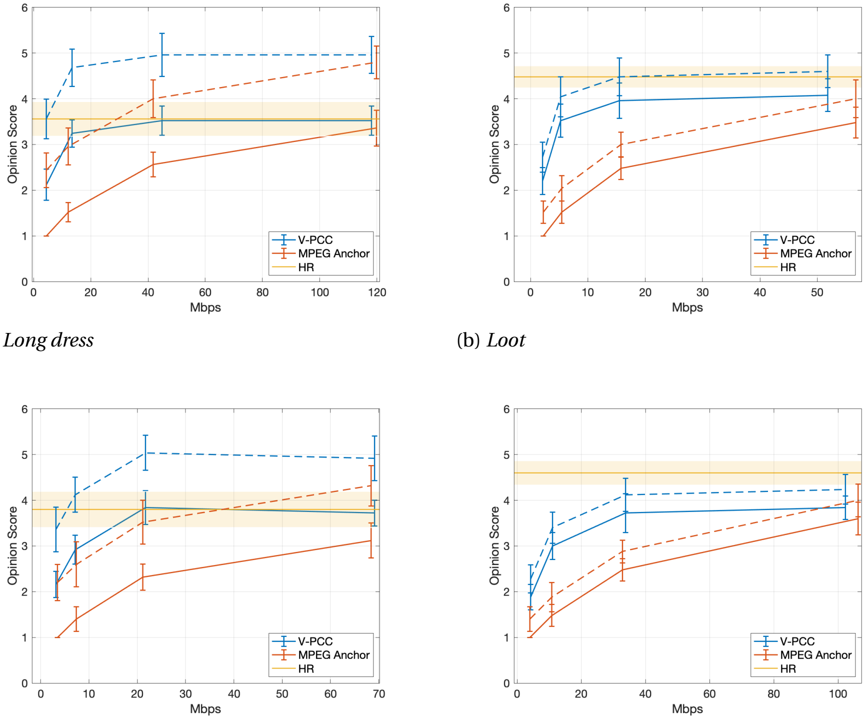
(c) *Red and black*

(d) *Soldier*

Figure 2.4: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps in the 3DoF viewing condition in test T2. HR scores are shown using a shaded yellow plot.

Due to the non-normality of our data, the aligned rank transform was applied prior to the fitting [32]. Since the transform is designed for a fully randomized test, it is not suitable for the scores collected during the test, as the HR addition makes the design matrix rank deficient. However, the transform can be applied to the differential scores used to obtain DMOS, as it follows a fully randomized design. Thus, it was decided to perform the analysis on the differential scores.

For test T1, analysis of deviance on the full mixed-effects model showed significance for main effects Content ($F = 48.14$, $df = 3$, $p < .001$), Codec ($F = 51.01$, $df = 1$, $p < .001$) and bit-rate ($F = 375.35$, $df = 3$, $p < .001$), but not for DoF ($F = 0.0003$, $df = 1$, $p = 0.988$). Moreover, significant interaction effects were found for *DoF - Content* ($F = 4.31$, $df = 3$, $p = 0.005$), *Content - bit-rate* ($F = 5.88$, $df = 9$, $p < .001$) and *Codec - bit-rate* ($F = 4.73$, $df = 3$, $p = 0.003$). Post-hoc interaction analysis with Holm p-value adjustment indicates that the difference between 3DoF and 6DoF has statistical significance at 5% level when comparing contents *Manfred* and *Queen* ($\chi^2 = 10.34$, $p = 0.008$), as well as *Inspector* and *Queen* ($\chi^2 = 8.35$, $p = 0.019$). In other words, the relative difference in scores between contents *Manfred* and *Queen* (and *Inspector* and *Queen*) was



(a) *Long dress*

(b) *Loot*

(c) *Red and black*

(d) *Soldier*

Figure 2.5: MOS (solid line) and DMOS (dashed line) against achieved bit-rate, expressed in Mbps in the 6DoF viewing condition in test T2. HR scores are shown using a shaded yellow plot.

not found to be statistically equivalent in 3DoF with respect to 6DoF. This indicates that the DoF might have an effect on how contents are scored with respect to one another, for example by increasing or reducing their differences. Regarding the interaction effect between contents and bit-rates, post-hoc interaction analysis with Holm p-value correction showed statistical significance in differences between contents *Manfred* and *Queen* at bit-rates R2 and R4 ($\chi^2 = 29.52$, $p < .001$), between contents *Sarge* and *Despoina* at bit-rates R2 and R4 ($\chi^2 = 11.00$, $p = 0.028$), between *Sarge* and *Queen* at bit-rates R2 and R4 ($\chi^2 = 11.56$, $p = 0.022$), and between *Despoina* and *Queen* at bit-rates R1 and R2 ($\chi^2 = 13.75$, $p = 0.007$), R2-R3 ($\chi^2 = 13.59$, $p = 0.007$) and R2-R4 ($\chi^2 = 45.13$, $p < .001$). Results can be explained considering that the low HR scores given to content *Queen* meant a narrower range of ratings. Thus, bit-rate point R2, for example, presents relatively higher differential scores for *Queen* with respect to the rest of the contents, whereas for bit-rate point R4, due to the HR removal, all contents have similar ratings. This is reflected in the statistical analysis conducted on the scores. Finally, post-hoc interaction analysis with Holm p-value adjustment on differences between codecs and bit-rates shows that the difference among codecs is statistically significant at 5% level

Table 2.1: Pairwise post-hoc test on the contents for test T1 and T2, using Wilcoxon signed-rank test with Bonferroni correction.

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

only between R1 and R2 ($\chi^2 = 10.51$, $p = 0.007$), R1 and R3 ($\chi^2 = 7.09$, $p = 0.031$), and R1 and R4 ($\chi^2 = 10.17$, $p = 0.007$). This indicates that the differences between codecs remain constant at all bit-rates, except for R1. This is in line with what observed in Fig. 2.2, which show similar trends for codec V-PCC with respect to the MPEG anchor, except for the lowest bit-rate point, for which V-PCC achieves better performance.

Results of analysis of deviance on the full mixed-effects model for test T2 showed significance for main effects Content ($F = 139.41$, $df = 3$, $p < .001$), Codec ($F = 692.24$, $df = 1$, $p < .001$) and bit-rate ($F = 485.11$, $df = 3$, $p < .001$), but not for DoF ($F = 2.57$, $df = 1$, $p = 0.115$), similarly to what was seen for test T1. Interactions were found significant at 5% level between Content and Codec ($F = 3.81$, $df = 3$, $p = 0.01$), Content and bit-rate ($F = 3.03$, $df = 9$, $p = 0.001$), and Codec and bit-rate ($F = 39.40$, $df = 3$, $p < .001$). The lack of significance in interactions involving DoF is in line with the results of the Wilcoxon signed-rank test, which showed no significance for DoF in test T2 ($Z = -1.96$, $p = 0.0502$, $r = 0.05$). Post-hoc interaction analysis with Holm p-value adjustment shows significance at 5% level for the differences among codecs for content *Long dress* with respect to content *Loot* ($\chi^2 = 10.09$, $p = 0.009$). This confirms what can be seen in Fig. 2.4: the gap among codecs is more prominent for content *Loot* with respect to *Long dress*, probably due to the reduced range associated with a low-rated HR. Post-hoc analysis on the interaction between contents and bit-rates indicates statistical significance at 5% level for differences among contents *Long dress* and *Soldier* when considering differences between bit-rates R1-R4 ($\chi^2 = 17.03$, $p = 0.001$) and R2-R4 ($\chi^2 = 11.81$, $p = 0.021$), and among contents *Red and black* and *Soldier* for differences between R1 and R4 ($\chi^2 = 11.80$, $p = 0.021$). Again, this can be explained considering that both *Long dress* and *Red and black* received remarkably lower scores, which resulted in a narrower rating range. Thus, differences among lowest and highest bit-rates are quite different between those two contents and *Soldier*, which benefited from a larger rating span. Lastly, post-hoc analysis on the interaction between codecs and bit-rates reveals statistical significance at 5% level for all pairwise comparison, except R1-R3 (R1-

Table 2.2: Pairwise post-hoc test on the bitrates for test T1 and T2, using Wilcoxon signed-rank test with Bonferroni correction.

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

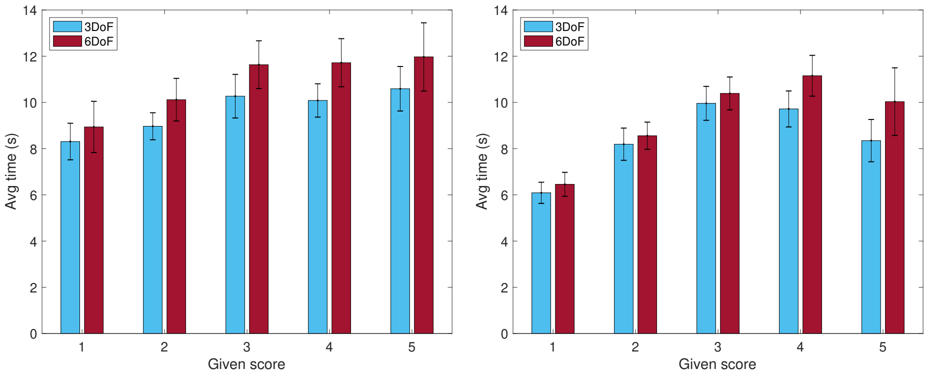
R2: $\chi^2 = 14.60$, $p < .001$, R1-R4: $\chi^2 = 46.58$, $p < .001$, R2-R3: $\chi^2 = 13.81$, $p < .001$, R2-R4: $\chi^2 = 113.34$, $p < .001$, R3-R4: $\chi^2 = 48.02$, $p < .001$). Indeed, in Fig. 2.4 it is quite evident that the curves for the two codecs follow different trends. In particular, codec V-PCC seems to saturate between R2 and R3, whereas a steeper slope is observed for the MPEG anchor.

2.5.2. Additional questionnaires and interaction data

IPQ & SSQ Questionnaires

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

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



(a) T1

(b) T2

Figure 2.6: Average time spent looking at the sequence (in seconds) and relative CIs, against score given to the sequence, for 3DoF (blue) and 6DoF (red), in test T1 (left) and T2 (right).

INTERACTION TIME

Interaction time was found to be strongly correlated with MOS values in a study conducted on light field image quality assessment [33]. In particular, it was found that users tended to spend more time interacting with contents at high quality, whereas for low quality scores, less time was spent looking at the contents. In order to see whether similar trends could be observed in our data, we compared the average time spent watching the sequence in 3DoF and 6DoF, separately for each quality score given by the participants. Results are shown in Fig. 2.6. A positive trend can be observed between the given score and the average time spent looking at the sequence, with the exception of score 5, which for test T2 shows a negative trend with respect to the time. However, it should be considered that on average, a small percentage of scores equal to 5 were given in test T2 (10% of the total scores), thus, variations may be due to the difference in sample size. It is also worth noting that, on average, participants spent more time looking at the sequences in 6DoF, with respect to the 3DoF case. Indeed, several participants pointed out that the lowest scores were the fastest to be given, whereas for higher quality, it was harder to decide on the rating.

INTERVIEWS

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

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

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

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

2.5.3. Analysis of results

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

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

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

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

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

2.6. Discussion

2.6.1. Datasets

Despite the rich literature in point cloud acquisition and compression, few point cloud datasets are publicly available. This is especially true when considering point cloud datasets depicting photo-realistic humans. One of the most popular and widely used full-body dataset, created by 8i Labs [23], consists of only 4 individual contents, whereas the HHI Fraunhofer dataset has 1 individual content [27]. In the context of point cloud compression, such scarcity of available data may lead to compression solutions being designed, optimized and tested while considering a considerably narrow range of input data, thus leading to algorithms that are overfitted to the specifics of the acquisition method used to obtain the contents. The consequences of such a scenario are reflected in our results. Whereas for the contents assessed in test T2 a large difference was observed between codec V-PCC and the MPEG anchor, for the contents in test T1 the gap was markedly lower, and indeed the significance of the effect of the codec selection had a smaller effect size for test T1 with respect to test T2, as seen in section 2.5.1. Test T2 consisted of contents that had been used in multiple quality assessment experiments [16, 17, 34, 35], notably including the performance evaluation of the upcoming MPEG standard [10]. On the other hand, test T1 included contents that have not been used so far in assessment of point cloud compression solutions. The discrepancies in the results of the subjective quality assessment campaign indicate that performance gains may vary considerably when new contents are evaluated. A larger body of contents depicting digital humans, involving several acquisition technologies, is needed in order to properly design, train and evaluate new compression solutions in a robust way.

2.6.2. PERSONAL PREFERENCES AND BIAS

Subjective evaluation experiments are complicated by many aspects of human psychology and viewing conditions, such as participants' vision ability, translation of quality perception into ranking scores, adaptations and personal preferences for contents. Through carefully following the ITU-T Recommendations P.913 [30], we are able to control some of the aspects. For example, eliminate the scores given by the participants with vision problems; train participants to help them understand the quality levels; randomize the stimuli and viewing conditions to minimize the order effects. However, we noticed that personal preferences towards certain contents are difficult to control. Satgunam et al. [36]) found that their participants were divided into two preference groups: prefer sharper content versus smoother content. Similarly, Kortum and Sullivan [37] found that the "desirability" of participants had an impact on video quality responses, with a more desirable video clip being given a higher rating. In our experiments, content *Queen* is generally given lower ratings with respect to the other contents. In the interviews, many participants (27%) expressed dislike towards *Queen*, because of her lifeless look and static gestures; 40% showed their preference towards *Soldier*, due to his high-resolution facial features, unitoned clothes and natural movements. This observation suggests that quality assessment may need to be adjusted based on content and viewer preferences, and offering training with different contents.

2.6.3. TECHNOLOGICAL CONSTRAINTS AND LIMITATIONS

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

One of the main shortcoming of both compression solutions lays in their inability to reach visually-lossless quality, as demonstrated by our results. Achieving a visually pleasant result is of paramount importance for the market adoption of the technology; indeed, poor visual quality might lead consumers to tune off from the experience altogether [38]. Visual perception should be taken into account when designing compression solutions, especially at high bitrates, to ensure that in absence of strict bandwidth constraints, excellent quality can be achieved.

2.6.4. PROTOCOLS FOR SUBJECTIVE ASSESSMENT IN VR

It is critical to select the appropriate subjective evaluation methodology based on the target modality and distortions being evaluated, as this can have a significant effect on the collected quality scores. Single stimulus methodologies, in particular, lead to larger CIs with respect to double stimulus methodologies, and are more subject to be influenced by individual content preference [30]. An early study comparing single and double stimulus methodologies for the evaluation of colorless point cloud contents indicated that the latter was more consistent in recognizing the level of impairment, as relative differences facilitate the rating task [39]. However, the study pointed out that the single stimulus methodology shows more discrimination power for compression-like artifacts, albeit at the cost of wider CIs.

Double stimulus methodologies, while commonly used in video quality assessment and widely adopted in 2D-based quality assessment of point cloud contents [10, 16, 17], are tricky to adopt in VR technology, due to the difficulties in displaying both contents simultaneously in a perceptually satisfying way [40], while ensuring a fair comparison between the contents under evaluation. When dealing with interactive methodologies, in particular, synchronous display of any modification in viewport is usually enforced, to ensure that the two contents are always visible at the same condition [17, 33]. This is clearly challenging to implement in a 6DoF scenario, in which users are free to change their position in the VR space at any given time. Positioning the two contents side by side in the same virtual space would mean that, at any given time, they are seen from two different angles. Temporal sequencing is also unsuitable, as the participant can view the paired sequences from different angles and distances. A toggle-based method like the one proposed in [40] is not applicable to moving sequences, as different frames would be seen between stimuli.

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

2.7. CONCLUSION

In this chapter, we presented two subjective quality evaluation protocols. The first was based on viewing prerecorded videos on a 2D screen and was employed in the MPEG point cloud compression standardization activity. The second was a novel evaluation of dynamic point clouds in an immersive viewing environment. We use the second approach in subsequent chapters as it is more realistic considering our target application of VR remote communication. Using this approach, we compared the performance of the point cloud compression standard V-PCC against an octree-based anchor codec (MPEG anchor). Participants were invited to assess the quality of digital humans represented as dynamic point clouds, in both 3DoF and 6DoF conditions. The results indicate that codec V-PCC performs better than the MPEG anchor, especially at low bit-rates. At the highest bit-rate that we tested, the two codecs are often statistically equivalent. Results indicate that the content under test has a significant influence on how the scores are distributed; thus, new data sets are needed in order to comprehensively evaluate compression distortions. Moreover, current encoding solutions, while efficient at low bitrates, are unable to provide visually lossless results, even when large volumes of data are available, revealing significant shortcomings in point cloud compression.

We provide a protocol to subjectively assess the quality of point cloud human reconstructions in immersive environments to address **R1** *(How can we measure the perceived quality of dynamic point cloud user reconstructions in immersive environments?)*. The results indicate a small effect of the immersive viewing condition on the recorded quality scores to address *R1.1 (What is the impact of immersive viewing environments on subjective quality assessment?)*. In general, both visualization methods led to similar trends in MOS values however, participants indicated that the 6DoF viewing condition felt more realistic and interactive. To address *R1.2 (What factors do users rely on while performing subjective quality assessment of dynamic point clouds?)*, we provide qualitative insights on the factors that participants used to evaluate point cloud quality through the interview conducted at the end of the session. 56% of experiment participants reported that they assigned a quality score based on object outline, patterns of distortion on the body or clothes, the naturalness of movements and visual artefacts such as blockiness and blurred regions.

In the next chapter, we implement a framework to adaptively stream prerecorded dynamic point clouds and play them back in an immersive viewing environment. We evaluate the quality of these streams using objective image distortion metrics as well as subjective quality assessment using the protocol described in this chapter. In all subsequent chapters, due to the performance constraints of real-time systems, we utilize the MPEG anchor codec to adaptively deliver dynamic point cloud reconstructions.

REFERENCES

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

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

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

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

[5] J. Constine, *Facebook animates photo-realistic avatars to mimic VR users' faces,* (2018).

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

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

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

[9] 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**, 828 (2017).

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

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

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

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

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

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

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

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

[18] 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)* (IEEE, 2017) pp. 1–6.

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

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

[21] R. D. Queiroz and P. A. Chou, *Compression of 3d point clouds using a region-adaptive hierarchical transform,* IEEE Transactions on Image Processing 25 (2016).

[22] 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 (2016).

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

[24] V. Baroncini, P. Cesar, E. Siahaan, I. Reimat, and S. Subramanyam, *Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression,* ISO/IEC JTC1/SC29/WG11 M41786 (2017).

[25] ITU-T P.910, *Subjective video quality assessment methods for multimedia applications,* International Telecommunication Union (2008).

[26] E. S. Jang, M. Preda, K. Mammou, A. M. Tourapis, J. Kim, D. B. Graziosi, S. Rhyu, and M. Budagavi, *Video-based point-cloud-compression standard in mpeg: From evidence collection to committee draft [standards in a nutshell],* IEEE Signal Processing Magazine **36**, 118 (2019).

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

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

[29] 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**, 203 (1993).

[30] 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 (2016).

[31] M. Kay and J. Wobbrock, *mjskay/artool: Artool 0.10.6,* (2019).

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

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

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

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

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

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

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

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

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

3

Investigating user centered adaptive streaming of dynamic point clouds with low complexity tiling

In the previous chapter, we presented a protocol to perform subjective point cloud quality evaluation in an immersive viewing environment. This chapter proposes an adaptive streaming approach for individual point cloud human reconstructions with a low complexity approach to segment the point cloud into tiles. We define an auxiliary utility function, and we employ established methods from the literature and newly-proposed schemes for distributing the available bandwidth across the tiles. Both the objective and subjective evaluations confirm that considerable gains can be obtained by employing user-adaptive streaming. The results indicate adaptive streaming can deliver bitrate gains of upto 57% *to deliver comparable objective quality and gains of upto* 65% *to deliver comparable subjective quality with respect to a non-adaptive approach. This work is meant to investigate the potential gains from adaptive streaming before it is applied to stream live captured point cloud user reconstructions in VR remote communication in the next chapter.*

This chapter is based on the following publications:

1. ***Shishir Subramanyam**, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 3669–3677. https://doi.org/10.1145/3394171.3413535*
2. ***S. Subramanyam**, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic and P. Cesar. 2022 Subjective QoE Evaluation of User-Centered Adaptive Streaming of Dynamic Point Clouds. 2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*

3.1. INTRODUCTION

In the last chapter, we presented a protocol and methodology for subjective quality evaluation of dynamic point clouds in immersive VR viewing environments. In this chapter, we present a tiled adaptive streaming approach and evaluate it using this subjective evaluation methodology along with an image distortion-based objective evaluation.

Compared to traditional video, 6DoF social VR applications introduce additional challenges in the pipeline for delivering and representing content in order to offer greater levels of interaction. User navigation can be exploited to optimize the delivery of volumetric data, as only parts of the content are visible at any given time. User-adaptive streaming solutions can be deployed to reduce the bandwidth allocation for parts of the content that are outside of the field of view, ensuring a better Quality of Experience (QoE) for the parts that are visible. In the past, adaptive streaming solutions for point cloud contents have focused on optimizing delivery when multiple point clouds need to be transmitted [1, 2]. User evaluation of adaptive streaming strategies for point cloud contents in VR has largely been absent from the literature, due to the real-time rendering requirements that such an evaluation requires [3]. As such, subjective assessment for user-adaptive streaming has relied on prerecorded videos shown on 2D screens [4, 5]. In this chapter, we employ the subjective evaluation protocol proposed in the previous chapter along with image distortion metrics to subjectively and objectively assess user-adaptive streaming to quantify the rate-distortion gains.

The contributions of this chapter are three-fold:

- We demonstrate the validity of user-centered, real-time adaptive streaming of point clouds based on objective and subjective evaluation
- We propose a utility function to assign bit rate to tiles from a single point cloud object
- We present two new rate allocation strategies for independently decodable point cloud tiles

In this chapter, we demonstrate that user-centered tiled streaming can consistently deliver a higher QoE with respect to non-adaptively encoded content, under the same codec and bit rate. This will help to address the research question *R2.1 (Does user-centered tiled adaptive streaming offer significant gains to reconstruction quality and bandwidth savings?)* and evaluate quality gains through both objective and subjective evaluation. We conduct the quality evaluation experiment with our open source, standards-compliant point cloud processing library. In particular, we implement a live playback environment to render synchronized tiled streams of point clouds, that supports adaptive quality selection based on user movements in 6DoF. We also address *R2.2 (What are the acceptable quality levels amongst adjacent tiles to maximize the quality of the final point cloud reconstruction?)* and determine acceptable quality differences through the various tile rate allocation strategies that are evaluated. Moreover, we present an analysis of user interactions and navigation patterns during the quality assessment task, drawing useful insights

In the remainder of the chapter, we first discuss the related work on point cloud streaming and quality evaluation. We then propose a user-adaptive streaming approach comprising a tile utility function and combining existing tile rate allocation strategies with two novel strategies. We then describe the software platform created to render dynamic point clouds and perform the evaluation. This is followed by a description of the evaluation methodology and experimental setup used. We then present the results of the objective and subjective evaluation along with a statistical analysis of the quality scores. This is followed by a discussion of the results including implications for future research and the conclusions drawn from this work.

.

3.2. Related Work

3.2.1. Point Cloud Compression and Streaming

In recent years, point cloud compression has received significant research attention [6]. The MPEG launched new compression standards [7] for static and dynamic point clouds, namely, G-PCC and V-PCC, respectively. G-PCC relies on octree-decomposition (Octree) for geometry encoding, with a potential triangular surface reconstruction step enabled (TriSoup) [8] after reconstruction, while color encoding is performed using the RAHT [9] or Lifting modules. On the other hand, V-PCC maps the point cloud to a 2D grid and uses legacy video coding tools to compress the geometry and attributes of the point cloud. This codec achieves a high compression ratio, however, it is not suitable for real-time applications due to its high encoding complexity and latency [1]. The original MPEG Anchor codec [10], proposed by Mekuria et al. [11], uses an octree space-partitioning structure to encode point cloud geometry. Point cloud attributes, such as color, are then encoded by mapping them to a 2D grid and applying JPEG compression. This codec design can be employed for both static and dynamic content. Moreover, it achieves low encoding/decoding complexity, which makes it suitable for real-time systems. Hence, in this work, we use the V-PCC codec as a benchmark for rate-distortion performance, and the MPEG Anchor codec to evaluate the gains offered by adaptive streaming of dynamic sequences of encoded point clouds.

Advances in low-latency streaming and volumetric point cloud delivery mechanisms have led to the emergence of novel teleimmersion systems that allow distributed remote users to communicate as themselves in a shared environment with realistic user reconstructions. Microsoft released the RoomAlive Toolkit for creating interactive Augmented Reality (AR) experiences [12, 13]. Mekuria et al. proposed a teleimmersive system that blends avatar representations and photo-realistic reconstructions of users in a shared virtual environment [14]. Cernigliaro et al. proposed a point cloud multi-point control unit for optimizing holo-conferencing systems [2]. Gunkel et al. introduced VR-Comm [15], a web based social VR communication system using photo-realistic user reconstructions that was evaluated using both simulations and subjective studies. Jansen et al. [16] proposed a pipeline for volumetric videoconferencing using low latency DASH with photo-realistic point cloud user reconstructions. In this chapter, we propose a standards-compliant mechanism for adaptive streaming that can be integrated in such teleimmersive systems.

Delivering content in a format that is suitable for real-time interactivity in 6DoF applications, introduces additional constraints in encoding and media representation.

Hooft et al. provides a summary of open challenges in [17]. For scenes with multiple point clouds, Hosseini et al. [18] propose DASH-PC; a dynamic, adaptive, and view-aware point cloud streaming system. They present three algorithms to spatially subsample point cloud assets in a scene, creating multiple representations. The density of each representation is known to the client, which selects a corresponding version, based on human visual acuity. However, they do not account for the orientation of the underlying point cloud surfaces. Park et al. [19] present a streaming framework based on 3D tiles. They define a utility function per tile based on the user's proximity, underlying quality of the tile and the user's display device resolution. To account for real-time interactions, the authors propose a window-based design for the Client Buffer Manager. The representation for each tile is decided using a greedy approach to maximize utility. Hooft et al. [1] propose PCC-DASH, a standards-compliant means for HTTP adaptive streaming of scenes, comprised of multiple point cloud reconstructions. The authors propose rate adaptation heuristics to select representations for each object in the scene, based on viewport position, available bandwidth, and current buffer status. This approach uses a single quality for every object. He at al. [20] propose view-dependent streaming over hybrid networks. Each point cloud frame is projected onto the six faces of a bounding cube, with a color and a depth video created per face. The videos are transmitted using digital broadcasting. The user can request videos that correspond to particular faces of the cube in high quality, reconstructing the point cloud from the downloaded depth and color videos. In this chapter, we perform objective and subjective evaluations in a live playback VR environment, in order to assess the gains in the perceived quality.

3.2.2. Segmentation

Segmentation is the division of the point cloud into clusters of points that are homogeneous with respect to a selected characteristic. In order to adaptively stream a point cloud object based on the orientation of the user's viewport, we need to cluster points based on the orientation of their underlying surface. There are two classes of methods to estimate normals to the surface for each point [21]. The most accurate method is to reconstruct the surface and create a watertight mesh from the point cloud frames. The normals to the surface can then directly be associated with each point. The second approach is based on inferring the underlying surface based on the point local neighbourhood. Nguyen et al. [22] present a taxonomy of five classes of segmentation algorithms: edge-based, region-based, attribute-based, model-based, and graph-based. Attribute-based segmentation methods can be used to tile point clouds, and account for surface orientation. These methods rely on estimating additional attributes for each point, e.g., normals, to obtain the surface orientation before clustering. Thus, these solutions are not suitable for a real-time system, as inferring the surface normal requires repeated eigendecomposition for the local neighbourhood of each point.

In this chapter, we propose a simple low complexity approach to estimate surfaces based on visibility from four virtual cameras. This approach is suitable for real-time applications and allows adaptive streaming based on the orientation of the user's viewport.

3.2.3. Subjective Quality Assessment

The JPEG and MPEG activities for the establishment of standards have raised interest in subjective evaluation of point cloud contents. Initially, most studies were focused on evaluating the quality of static content on 2D displays. Zhang et al. [23] evaluated the impact of degradation such as down-sampling and noise generation. Their results indicate that human visual perception can tolerate colour noise more than geometry or coordinate noise. Alexiou et al. [24] compared the Double Stimulus Impairment Scale and the Absolute Category Rating methodology for quality assessment of geometric degradations. Subjective evaluations using the Screened Poisson surface reconstruction for rendering purposes were conducted in [25], with results indicating different rating behaviours when compared to evaluations using raw point clouds. In [26], a comprehensive study of the rate-distortion performance of the entire set of encoders in the MPEG standard [7] across the rate points used in development [10] is presented, along with an evaluation of objective quality metrics. In [27], a subjective evaluation campaign of a subset of MPEG point cloud codecs was issued, across four independent laboratories with different testing equipment, revealing high correlation. Zerman et al. [28] performed a subjective evaluation of the V-PCC codec and found no correlation between perceived quality and reference point cloud density. The authors also conducted a subjective evaluation of perceived quality between textured meshes and point clouds using state-of-the-art codecs [29]. The results using their rendering solution showed that meshes are able to provide better quality at high bit rates (over 50 Mbit/sec) and point clouds are well suited for applications with limited bandwidth. Javaheri et al. [30] evaluated both subjectively and objectively the quality of point cloud contents under geometry artefacts occurring by different compression approaches. The same authors also provide an evaluation of point cloud rendering techniques and codecs [31] to compare point-based rendering methods (with and without recolouring) and surface reconstruction based mesh rendering against the PCL [21], V-PCC and G-PCC codecs. The results indicate that surface reconstruction can mask some of the artefacts introduced by the codecs. Moreover, Dumic et al. [32] conducted a subjective evaluation on point cloud rendering and display devices, showing that users do not have a preference between 2D or 3D displays while preferring inspection of raw point clouds against meshes created using Poisson surface reconstruction. Hooft et al. [4] compared subjective and objective evaluation results from adaptive streaming of multiple point clouds using algorithmically generated camera paths and the adaptation schemes described in [1], with obtained videos shown to users on a 2D screen. In this work, a point-based rendering approach is employed for the user-adaptive point cloud tiles, which are evaluated in a virtual environment and consumed by means of an HMD.

Quality evaluation in immersive environments has recently gained popularity in the research community. Mekuria et al. [14] evaluated quality based on factors such as immersiveness, togetherness and realism between users in a social VR experience, which are represented using point cloud reconstructions as well as avatars. In [33], subjective quality assessment of point cloud contents was performed in Augmented Reality (AR) using HMDs, while in [34], the authors compare the collected quality scores with subjective ratings from experiments using a 2D monitor. The PointXR toolbox for subjective evaluation of static point clouds in VR was proposed in [35], and was employed to assess the performance of color encoders that have been integrated in G-PCC. Tran et al. [36] found that when evaluating quality in immersive environments, factors like cybersickness and presence should not be overlooked. However, the approach presented in [3, 35] was based on loading a fixed point cloud representation on physical memory, making it unsuitable for evaluating adaptive streaming.

In this chapter, we introduce a software framework for rendering and evaluation of adaptive streams of dynamic point clouds, based on real-time user interactions in 6DoF VR environments.

3.3. OBJECTIVE QUALITY EVALUATION

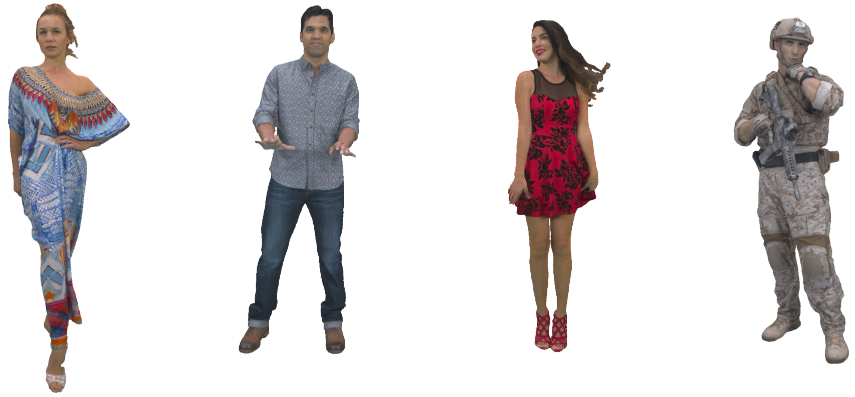
3.3.1. TILING STRATEGY

3D dynamic point clouds are usually captured using a multi-camera setup surrounding the object to be scanned. In order to consistently align each of the 3D views into a complete model, a registration step is required in order to identify a transformation matrix for each of the cameras. In a real-time capture scenario with multiple depth sensors, the camera visibility of each point in the point cloud can be recorded. This can serve as a proxy for the orientation at each point without explicitly performing surface reconstruction. We can assign an orientation vector to each camera location using the transformation matrix for each camera from the registration step. The assumption here is that the camera locations are a proxy for the end user's viewport location. This allows us to assign an orientation to each segment or tile without any additional computational overhead. We then treat each tile as a separate point cloud track. The server creates multiple representations by encoding each tile at multiple qualities. We also include additional metadata to each tile in the adaptation set such as the orientation vector of each tile. The client then uses this information to request each representation at the optimal quality based on the orientation of their viewport. This allows the client to adapt to interactions such as moving the viewport.

In this work, we propose a low-complexity tiling approach based on drawing a vector from the object centroid to the surface and comparing this to four virtual cameras placed around the object in the XZ plane in order to divide the object spatially by assigning points to each of the virtual cameras. We restrict the number of tiles to four for compression efficiency and lower metadata. Here we assume that the point cloud object presents a smooth convex hull. In the Media Presentation Description (MPD) document we include the tile metadata in the adaptation set for each tile. This can be used by the Client Buffer Manager to optimize the representation per tile based on the user's viewport. The tile metadata includes the components of the camera orientation vector and the coordinates of the tile centroid.

3.3.2. TILE RATE ALLOCATION

In order to compare different tile selection algorithms and study the potential impact of tiling on optimizing delivery we assume that the location and orientation of the user's viewport for the next frame is always known. In practice a probability density function can be modelled based on past navigation patterns and the probability of each possible next viewport can be computed.



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

Figure 3.1: Sequences in the 8i Voxelized Full Body Dataset (from left to right: *Longdress*, *Loot*, *Red and Black*, *Soldier*) [37]

The viewport of the user is represented by a location V_l and an orientation V_o. We define the utility U of a tile Ti as $U(T_i) = \vec{T}_i \cdot \vec{V}_o$. The bit rate budget is divided amongst available tiles, based on this utility. A representation for each tile can then be selected, and the final representation vector for a frame is retrieved from the server. In this work, we use three allocation schemes to select representations for each of the tiles, as originally proposed by Hooft et al. [1]. We first sort the tiles based on their utility, defining visible tiles as the ones having a positive $U(T_i)$.

1. **Greedy** bit rate allocation: The highest quality representation is first set for the highest utility tile and then we move on the next highest ranked tile until the bit rate budget is spent.

2. **Uniform** bit rate allocation: The representation of tiles are increased one step at a time starting with the highest utility tile.

3. **Hybrid** bit rate allocation: The representations of visible tiles are first uniformly increased in order of utility. The representations of the remaining tiles are then uniformly increased until the bit rate budget is spent.

3.3.3. Dataset Preparation

In order to perform our evaluation, we use the 8i Voxelised Full Body Dataset shown in figure 3.1, and place four virtual cameras around the object on the XZ (floor) plane (at (1,0,0), (0,0,1), (-1,0,0) and (0,0,-1)). We assume a smooth convex hull with no occlusions, similar to real-time capture with multiple depth sensors. We draw a vector from the centroid of the point cloud to every point on the surface, and we use the vector dot product of these vectors with the 4 virtual cameras to assign a tile number to each point in the cloud.

To evaluate the impact of user adaptation, we select the MPEG anchor codec proposed by Mekuria et al. [11] which uses the popular octree space partitioning structure to encode point cloud geometry. Attributes like colour are then encoded by mapping them to a 2D grid and applying JPEG compression. This codec design allows for low-delay encoding and decoding, making it suitable for real-time applications. Thus, it represents a viable solution for evaluating our tiling strategies. The adaptation set is prepared using the MPEG anchor codec in an all-intra configuration. Each tile is encoded at octree depths from 6 to 11, with the JPEG quantization parameter varying from 55 to 95 in increments of 10. To measure the performance of viewport adaptive streaming we encode the source point clouds with the same codec configuration on the MPEG anchor codec. Additionally, we compare the performance of our tiling approaches against the MPEG V-PCC standard [7]. The codec is based on extending legacy video compression techniques by mapping the point cloud geometry and attributes to a 2D grid, and using video compression to encode both the geometry and the attributes separately. The current implementation of this approach has high encode complexity, making it unsuitable for real-time applications. Moreover, its compression performance is expected to decrease when using tiling approaches, which reduces the amount of data that can be packed in the 2D grid. We use the V-PCC codec to provide a baseline, indicating the state-of-the-art rate-distortion performance. We encode the source point clouds using Release 7.0 of the MPEG V-PCC codec, using the configurations provided in the Common Test Conditions for Category 2 All Intra (C2AI) encoding. We selected the rate points 1, 3 and 5 and extend it to an additional rate point, using a Texture QP of 9, a geometry QP of 12 and an occupancy precision of 2.

We used static bitrates targets, to remove the effect of network adaptation. We define the maximum bit allocation budget based on the encoded bitstream size for the Common Test Condition compression profiles supplied with the MPEG V-PCC codec, for each sequence and rate point.

Finally, the navigation patterns recorded during the experiment described in the last chapter were then used to set the camera position and rotation in Unity. To render the tiles, we first compute the utility of each tile by comparing the recorded user viewport orientation with the four virtual cameras used to create the tiles. We then select a representation for each tile, render the point cloud and record a screenshot from the viewport.

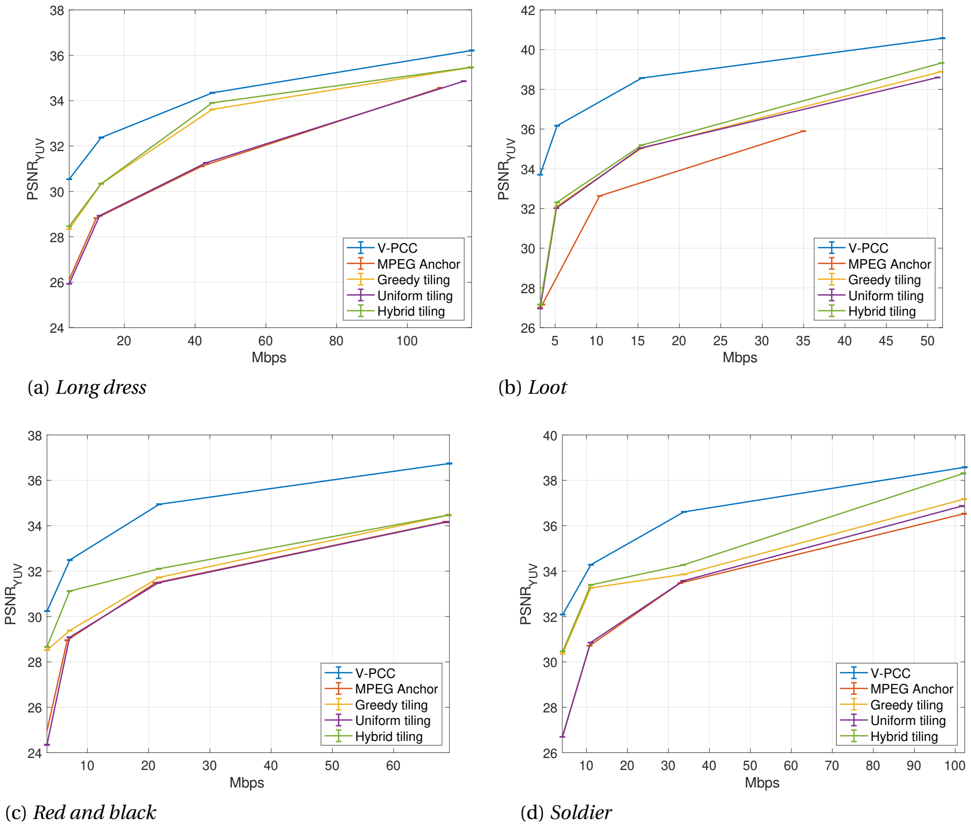
3.3.4. Metrics

In order to obtain an assessment of the visual distortion of a given point cloud content, several point cloud objective quality metrics have been developed, and serve as a benchmark for evaluating different compression solutions [38]. Most of the state-of-the-art objective metrics, however, evaluate the visual quality of the entire point cloud content, thus including parts that would likely not be observed by users. The tiling approach aims at exploiting user attention in order to dedicate a larger portion of the bandwidth to parts of the point cloud that are actively visualized, while assigning less bits to parts that are effectively hidden to the users. Thus, common point clouds metrics, such as point-to-point or point-to-plane approaches, would not be suitable to assess the bit-rate gains brought by user-centered tiling approaches. In this work, we thus decided to evaluate the visual quality of the point cloud contents as they would be seen from the users, using common video metrics to estimate their quality. In particular, we use the camera positions recorded in the experiment detailed in Chapter 2, in order to simulate users visualizing point cloud contents. Each frame of each content is rendered in Unity, and the scene is captured in a resolution of $1920x1080$, to be as close as possible to the resolution of the HMD. The background color of the scene is set to $RGB = [0,177,85]$. The color was selected to provide maximal contrast with respect to the content as this color is not commonly found in human skin tones or clothing. For encoded contents and tiles with an octree depth less or equal to 7, we increase the offset to 0.016 units, to obtain watertight surfaces.

The image metric PSNR is used to provide an estimation of the quality of the distorted point clouds with respect to the uncompressed reference, as rendered in Unity. The computation is performed in the YUV space, as it was proven to be better correlated with human perception, and it is averaged across the channels using the weights proposed in [39]. The weights proposed in [39] are meant to account for the human visual system being more sensitive to changes in luminance. To reduce the impact of the background on the metric computation, we decided to perform the metric computation only on the parts of the acquired image which contain the point cloud content. In particular, we define a Region of Interest (RoI) by excluding points whose *RGB* values are equal to the background. Reference and distorted contents will likely have different number of points and occupancy grids, resulting in potentially mismatching RoI. Thus, we decided to use the intersection of the RoIs to compute our metrics, as suggested in [38]. To avoid biasing the results, we exclude from the computation frames whose RoI covers less than 0.1% of the entire frame.

3.3.5. RESULTS

Figure 3.2 shows the weighted PSNR computed on the YUV channels against the achieved bitrate, averaged across frames and navigation paths, separately for each content. While, as expected, the V-PCC codec achieves the best overall performance, it can be observed that, among the tiling approaches, the hybrid tiling approach yields the best results, closely followed by the greedy approach. Both outperform the MPEG Anchor by a notable margin. Among the approaches, the uniform approach is the one leading to the smallest gains in terms of PSNR. In fact, its performance is comparable with the MPEG anchor, with the notable exception of content *Loot*, for which the obtained results are in line with the other tiling approaches. Due to the restrictions imposed by the maximum bit budget, which was defined based on the V-PCC bitrates, for content *Loot* at the highest bitrate target, it was not possible to select a better representation for the MPEG Anchor, as it would have led to overshooting the bandwidth allocation. This leads to a loss in performance, as a large portion of the bit budget is left unused. The tiling approaches, on the other hand, are able to exploit the available bandwidth in a more efficient way, conducting to a better performance. Table 3.1 reports the Bjontegaard rate savings for each tiling approach, computed with respect to the MPEG Anchor. It can be seen that the largest bitrate saving is obtained with the hybrid approach, closely followed by the



(a) *Long dress*

(b) *Loot*

(c) *Red and black*

(d) *Soldier*

Figure 3.2: PSNR computed on the YUV channels against achieved bit-rate, expressed in Mbps, averaged across frames and navigation paths.

greedy approach, while the uniform approach yields more modest gains.

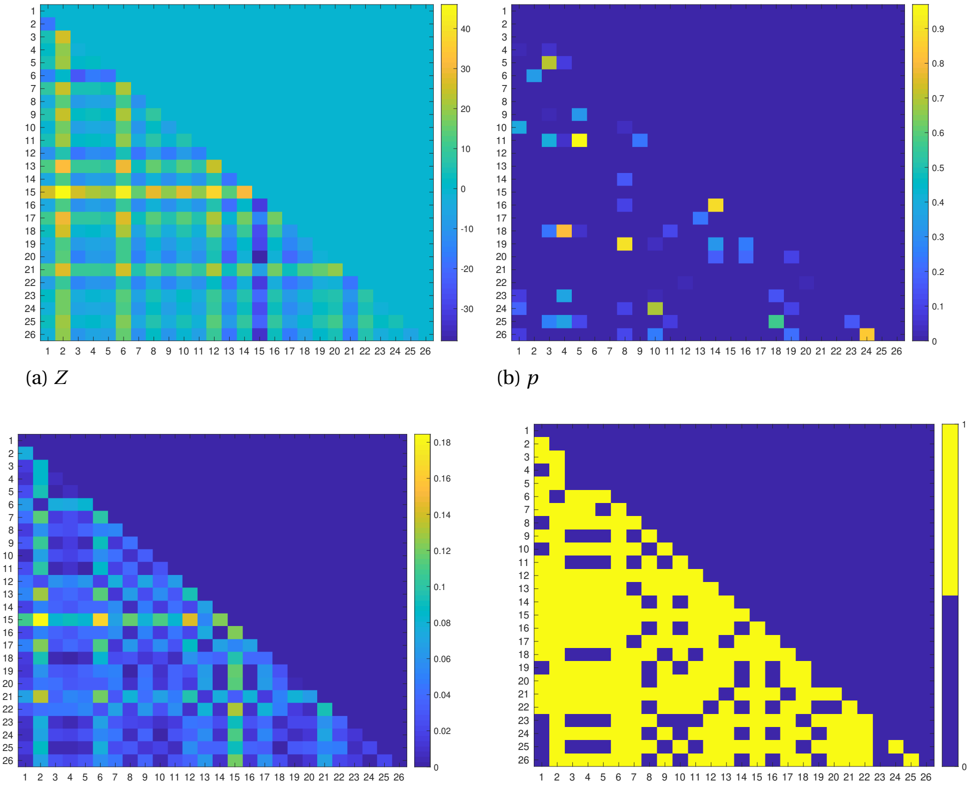
In order to understand whether the collected PSNR scores indicated significant differences among the conditions under test, we performed statistical tests on the data. Before conducting our analysis, we performed a Kolmogorov-Smirnov normality test on the entire set of PSNR scores ($N = 782760$), which rejected the null hypothesis that our data was normally distributed ($p < .001$). Thus, non-parametric tests were selected to analyse our data. Friedman's rank test performed on the scores revealed a significant effect of the codec selection on the final set of scores ($\chi^2 = 570782.56$, $p < .001$). To confirm that our results were not biased by the large number of scores involved, we performed random sampling on the data, selecting $N = 1000$ samples per codec across 1000 sampling runs (seeds) to reduce the dimensionality. We combined probabilities using Fisher's method [40, 41], concluding that the codecs have a significant effect on the scores ($\chi^2 = 944340$, $p < .001$). Results of the post-hoc Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/10$) are reported in Table 3.7. Results confirm that the codecs all show statistical difference with respect to each other, with a sizable effect

Table 3.1: Bjontegaard rate savings for each tiling approach, with respect to the MPEG Anchor.

	W1	**W2**	**W3**
Long dress	-54.39%	1.87%	**-57.15%**
Loot	-47.75%	-46.76%	**-50.43%**
Red and black	-12.84%	3.57%	**-42.85%**
Soldier	-35.15%	-2.77%	**-46.04%**
Average	-37.53%	-11.02%	**-49.12%**

size (r 0.60), indicating that the effect of the codec selection on the obtained scores is considerable.

Similarly, we performed statistical analysis on the scores to understand whether the content selection had a significant effect on the final set of scores. Due to the difference in navigation paths among different contents, we selected the unpaired Kruskal-Wallis



(a) Z

(b) p

(c) r

(d) $p < \alpha$

Figure 3.3: Results of the pairwise post-hoc test on the navigation paths, using Mann-Whitney U test with Bonferroni correction. Results are shown on the lower triangular matrix.

Table 3.2: Pairwise post-hoc test on the codecs under test, using Wilcoxon signed-rank test with Bonferroni correction.

	Z	p	r
V-PCC - MPEG Anchor	342.65	<.001	0.61
V-PCC - W1	342.23	<.001	0.61
V-PCC - W2	342.65	<.001	0.61
V-PCC - W3	338.20	<.001	0.60
MPEG Anchor - W1	-342.65	<.001	0.61
MPEG Anchor - W2	-333.21	<.001	0.60
MPEG Anchor - W3	-342.66	<.001	0.61
W1 - W2	340.25	<.001	0.61
W1 - W3	-328.72	<.001	0.59
W2 - W3	-342.66	<.001	0.61

Table 3.3: Pairwise post-hoc test on the contents, using Mann-Whitney U test with Bonferroni correction.

	Z	p	r
Long dress - Loot	-153.61	<.001	0.25
Long dress - Red and black	28.05	<.001	0.05
Long dress - Soldier	-162.55	<.001	0.26
Loot - Red and black	182.41	<.001	0.29
Loot - Soldier	33.79	<.001	0.05
Red and black - Soldier	-195.63	<.001	0.31

test. Results showed statistical significance of the content selection ($\chi^2 = 61297.24$, $p <$ .001), which was confirmed by random sampling of the data ($N = 1000$, with 1000 seeds) and combining probabilities with Fisher's method ($\chi^2 = 307470$, $p < .001$). Table 3.3 reports the results of post-hoc Mann-Whitney U test with Bonferroni correction ($\alpha = .05/6$). Statistical significance can be found between all the pairs; however, effect sizes for the pairs *Long dress - Red and black* and *Loot - Soldier* ($r = 0.05$ in both cases) indicate that effect, while apparently existing, is small. Similar results were reported in [3] for subjective tests on the same contents, indicating a difference in quality between the two groups of contents.

To assess whether statistical significance could be seen among the selected bit-rates, we ran a Friedman rank test on the scores. As expected, results confirmed that the bit-rates have a significant effect on the PSNR values ($\chi^2 = 587068.8$, $p < .001$), even when random sampling with $N = 1000$ and 1000 seeds was applied on the data (combined probabilities: $\chi^2 = \infty$, $p < .001$)[1] . Post-hoc analysis using Wilcoxon signed-rank test with Bonferroni correction ($\alpha = .05/6$) showed statistical significance, with large effect sizes, for all pairs ($Z = -383.10$, $p < .001$, $r = 0.61$ for all pairs).

Finally, a Kruskal-Wallis test revealed a significant effect of the navigation paths on the PSNR scores ($\chi^2 = 4621.5$, $p < .001$). To account for the difference in sample length, we performed random sampling ($N = 1000$), with 1000 seeds, and we aggregated the

[1] As $p = 0$ for all the sampled pairs, the aggregated χ^2 results equal to infinity.

Table 3.4: Results of ANOVA on the linear model PSNR ~ Path*Content*Codec*Rate.

	df	**MS**	*F*	*p*
Path	25	1973.40	3897.27	<.001
Content	3	302310.70	597034.58	<.001
Codec	4	436070.93	861198.19	<.001
Rate	1	7574782.73	14959468.11	<.001
Path:Content	75	582.54	1150.47	<.001
Path:Codec	100	27.62	54.54	<.001
Content:Codec	12	16083.88	31764.11	<.001
Path:Rate	25	511.29	1009.74	<.001
Content:Rate	3	47865.63	94530.01	<.001
Codec:Rate	4	41393.25	81747.70	<.001
Path:Content:Codec	300	15.66	30.94	<.001
Path:Content:Rate	75	99.50	196.50	<.001
Path:Codec:Rate	100	15.96	31.52	<.001
Content:Codec:Rate	12	9058.22	17889.11	<.001
Path:Content:Codec:Rate	300	6.42	12.69	<.001
Error	781720	0.51	1	0.5

probabilities using Fisher's method. Results confirmed the significant effect of the navigation paths ($\chi^2 = 561800$, $p < .001$). Results of the post-hoc test using Mann-Whitney tests with Bonferroni correction ($\alpha = .05/325$) are reported in Figure 3.3. The pairwise post-hoc test results are shown in the lower triangular matrix. Figure 3.3 (a) and (b) report the test statistic and p-value respectively. Out of 325 possible pairs, 246 of them (75.69%) presented statistical significance after Bonferroni correction as shown in Figure 3.3 (d). Results indicate that the choice of navigation path has a significant impact on the PSNR scores. The effect size (reaching max $r = 0.19$, Figure 3.3 (c)) suggests that the effect is not big, as it is expected that the choice of navigation path will not have same impact on the collected scores as the codec or rate selection. However, the statistical significance of a large combination of the pairs reveal that particular care should be put in differentiating navigation paths when performing the evaluation of compression solutions, as it appears to have an impact on the collected PSNR scores. While the difference in sample length might explain some of the statistical significance, it is noteworthy that the significant effect is maintained when sampling the same number of frames for each navigation path. The varying length is an important feature of the collected navigation paths, as different users had variable experiences in visualizing the contents. Yet, even when reducing the paths to the same length, statistical differences are observed among the paths, indicating that the varying ways of consumption among the users lead to significantly different scores.

A linear regression was applied on the data to understand the ability of navigation paths, contents, codecs and rates to predict the PSNR scores, using a full interaction model. Navigation paths, contents and codecs were treated as categorical variables. The results of the ANOVA conducted on the fitted model are reported in Table 3.4. The adjusted R^2 for the fitted model is 0.965, indicating that our independent variables are able to account for 96.5% of the variance of the model. Moreover, results of the ANOVA confirm that the main effects and the interactions all contribute significantly to the model. Based on the results from this experiment, we modified our approach and then performed a subjective evaluation as described in the next section.

3.4. SUBJECTIVE QUALITY EVALUATION

In this section, we present the subjective evaluation of tiled adaptive point cloud streaming. We create the point cloud tiles using the methodology described in the previous section. Based on the lessons learned from the objective evaluation and pre-trials with colleagues we modified the utility function to better account for distances between the user's viewport and each of the tiles as well as the tile orientation. We also introduce additional tile rate allocation schemes as described in the following subsections.

3.4.1. UTILITY FUNCTION DEFINITION

Let us consider a point cloud P, partitioned into N non-overlapping segments (tiles) T_i, $i = \{1, 2, ..., N\}$. For simplicity, we assume the segmentation to be operated on the XZ floor plane; that is, every segment spans the entire Y axis. An extension to full spatial segmentation is straightforward. We assume N to be even, and that each tile T_i has an associated tile centroid $T_i^{(c)}$ and a tile orientation vector $\vec{T}_i$; that is, $T_i = \left[T_i^{(c)}, \vec{T}_i\right]$. Such a vector could be estimated using the surface normal at each point, or obtained after surface reconstruction. For real-time capturing system, an approximation using the transformation matrix of each camera to construct an orientation vector was proposed and validated in the last section. For our tiling allocation strategies, we assume the orientation vectors to be such that $\forall i, \exists j : \vec{T}_i \cdot \vec{T}_j = -1$, with $i \neq j$; i.e., for each tile, there exists an opposite facing tile.

For any given user visualizing the point cloud from an external advantage point, with associated viewport $V = \left[V^{(c)}, \vec{V}\right]$ in which $V^{(c)}$ defines the center of the viewport, and $\vec{V}$ the orientation, we define the absolute utility $\left|u\left(V, T_i\right)\right| = \left|\vec{V} \cdot \vec{T}_i\right|$. Such a quantity considers the viewing angle, assigning higher utility to tiles that are facing the users, and lowest utility to tiles that are orthogonal with respect to the user's viewing direction. However, it fails to consider the impact of the users' location on the visibility and importance of the tiles. To incorporate the location information in the tiling utility, for each set of tiles with equal absolute utility (i.e., opposite-facing) (T_i, T_j), we compute the Euclidean distances $d(\cdot)$ between the viewport location $V^{(c)}$ and the tile centroids $T_i^{(c)}$ and $T_j^{(c)}$. The utility is computed as follows:

$$u(V, T_i) = \begin{cases} \left|\vec{T}_i \cdot \vec{V}\right|, & \text{if } d(V^{(c)}, T_i^{(c)}) < d(V^{(c)}, T_j^{(c)}) \\ -\left|\vec{T}_i \cdot \vec{V}\right|, & \text{otherwise.} \end{cases} \tag{3.1}$$

The resulting utility can then be used to divide the bit rate budget among the available tiles; a larger utility will correspond to higher visibility, whereas a smaller utility indicates lower visibility. As we compute utility using unit vectors, all values lie between

-1 and 1. A representation for each tile can then be selected, and the final representation vector for a frame is retrieved from the server.

3.4.2. TILE RATE ALLOCATION

In the subjective experiment, we consider five allocation schemes to select representations for each of the tiles. The first three were used in the objective evaluation described in the previous section and the remaining two are novel schemes we created based on the lessons learned from the objective evaluation and pre-trials with colleagues. By introducing additional novel allocation schemes, we aim to increase the variation in quality differences amongst adjacent tiles to identify acceptable limits for perceived quality. We sort the tiles based on their utility, defining visible tiles as the ones having a positive $u(V, T_i)$ or tiles that are closer to the user's viewport. Below, we provide a description of the bit rate allocation strategies that are employed.

1. **Greedy (W1)** [1]: The highest quality representation is first set for the highest utility tile and then we move on the next highest-ranked tile until the bit rate budget is spent.

2. **Uniform (W2)** [1]: The representation of tiles are increased one step at a time, starting with the highest utility tile.

3. **Hybrid (W3)** [1]: The representations of visible tiles are first uniformly increased in order of utility. The representations of the remaining tiles are then uniformly increased until the bit rate budget is spent.

4. **Top $N/2+1$ (W4)**: The representation of the best $N/2+1$ tiles are uniformly increased one step at a time. The remaining bit rate budget is then spent on the lowest utility tiles. The approach is inspired by the fact that, given a polyhedron with central symmetry, at any given viewport location, at most half of the polyhedron's faces are visible; we consider $N/2+1$ to compensate for the irregularity of the actual point cloud surfaces.

5. **Weighted Hybrid (W5)**: The bit rate budget is first split based on the weighted utility of each tile, defined as $\frac{u(V,T_i)+1}{\sum(u(V,T_i)+1)}$. The highest possible representation is then set for each tile based on the budget allocated to each tile. The remaining bit rate budget is then used to uniformly increase the representation of each tile in the order of their utility.

3.4.3. SUBJECTIVE EVALUATION PLATFORM

Real-time user interactions in 6DoF VR applications require low latency systems for selecting and rendering content at the client side. While previous research has focused on objective evaluation [18, 19, 42] or offline loading of point clouds on physical memory for subjective evaluation [3, 35], in this study, we conduct the subjective assessment of tiled streams of point clouds under realistic playback conditions and user interactions in real-time.

We implemented the playback environment using the Unity game engine and develop a serialized binary point cloud reader. This environment allows real-time playback from disk at 30 frames per second (fps), a tile synchronization component and a low latency rendering solution for dynamic point cloud sequences.

Point Cloud Software Library

We have developed an open-source software library to work on point cloud data available here: `https://github.com/cwi-dis/cwipc/`. Our software builds on the popular Point Cloud Library [21], Open3D [43] and other vendor specific libraries. It enables point clouds to remain opaque, similar to libraries that work on images and audio samples, allowing transfer across implementation language boundaries, while minimizing deep copies, making is suitable for real-time applications. The core of the library is written in C++ with most of the utilities written in Python.

Point Cloud Reader

To allow rendering at high frame rates that ensure comfortable viewing, and high responsiveness with user interactions, a fast reader for pre-recorded point cloud data is essential. Thus, we introduce a new feature in our software library, by creating a custom serialized binary point cloud format, that is suitable for reading point cloud frames of over a million points from disk for playback at 30 frames per second. The point coordinates and attributes are serialized and the frame number, timestamp and point size are placed in the header. The format is optimized to a form that is bit compatible with the library's point cloud data structure. This allows us to map the entire frame without an additional parsing step and thus, read point clouds at near disk speeds. In Tables 3.5 and 3.6, the read speeds and the corresponding file sizes for each point cloud sequence in the 8i Voxelized Full Body Dataset [37], are presented. For the largest sequence in terms of point count, *Soldier*, we are able to achieve an average read time of 22.3ms, which denotes a 94.9% and 84.5% reduction in the time needed for the popular Open3D library [43] to read the same sequence in the PLY ASCII and the PLY Binary format, respectively. The corresponding file size is on average 6.7% larger than the Binary PLY format, while remaining 27.8% smaller with respect to the PLY ASCII format.

Table 3.5: Mean ± standard deviation of reading time in milliseconds (ms), for each dynamic point cloud sequence across formats.

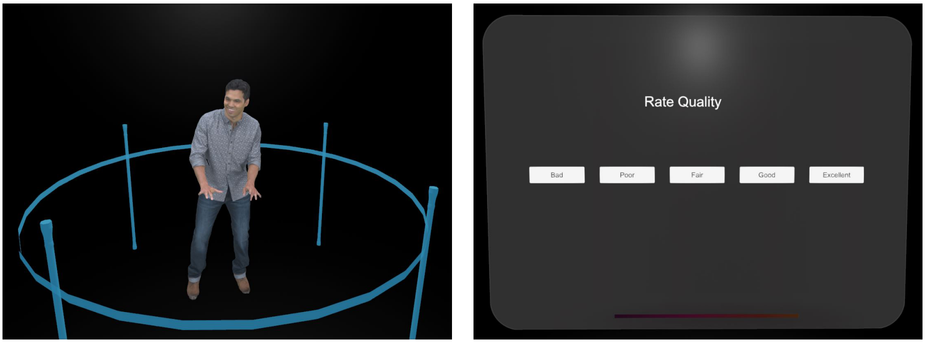
Content	PLY ASCII	PLY Binary	Serialized
Longdress	360.12 ± 15.31	114.27 ± 5.05	17.94 ± 1.46
Loot	337.00 ± 5.82	112.00 ± 7.75	17.33 ± 2.00
Red and Black	310.11 ± 16.78	100.46 ± 5.27	15.42 ± 2.32
Soldier	451.53 ± 9.91	147.76 ± 6.16	22.95 ± 5.24

Point Cloud Rendering

To render each point cloud frame, we store the point coordinates and attributes in a vertex buffer and draw procedural geometry on the GPU. Each point in the frame is rendered as a camera facing quad with the offset determined by the point size. We determine the point size for each point cloud frame offline, based on the average distance of

Table 3.6: Mean ± standard deviation of file size in MBytes, for each dynamic point cloud sequence across formats.

Content	PLY ASCII	PLY Binary	Serialized
Longdress	18.07 ± 0.80	11.94 ± 0.53	12.73 ± 0.57
Loot	16.35 ± 0.41	11.36 ± 0.19	12.11 ± 0.20
Red and Black	14.94 ± 0.82	10.40 ± 0.57	11.09 ± 0.61
Soldier	22.73 ± 0.35	15.38 ± 0.24	16.41 ± 0.25



(a) *Playback scene*

(b) *Rating scene*

Figure 3.4: Scenes from the playback application.

each point to its 5 nearest neighbors, following previous research on the field [26, 30, 31, 35].

Tiles Synchronization

The synchronization component was developed to playback tiled point cloud sequences with tiles of varying sizes and quality in a consistent manner. The unity renderer objects are per-tile, and there is no synchronization between the renderer objects except for the implied Unity rendering loop. Early in the rendering loop cycle, during the *Update()* call, the renderers report the potential frames they could render: the currently displayed frame and the next one if it is available. When all renderers have reported, the synchronizer selects the best frame number, so that all tiles are played back synchronized and at the correct time. This frame number is reported back to all renderers. For each frame rendered in the game engine the *Update()* methods are all called first after their completion the *LateUpdate()* methods are called. During *LateUpdate()*, each renderer then obtains the desired frame and passes it to the GPU. The synchronizer is extensible to other media types such as audio or video, different frame rates per stream, network streaming and jitter buffering.

Virtual Scenes

The evaluation environment, inspired by PointXR [35], consists of two 3D scenes to playback and evaluate tiled point cloud content. The *playback scene* places the point cloud sequence at the centre and spawns the user at a location in front of the point cloud. Tiled streams of point clouds are synchronized before playback. The player position is determined by the HMD trackers and players are free to walk and look around as they inspect the point clouds. To ensure participant safety and consistent boundaries for movement within the scene, we included a blue ring of 2.3 meters diameter to mark the extent of the physical play area as shown in Figure 3.4 (a). Players are advised at the start not to move outside this boundary. The *rating scene* places a fixed canvas on the wall with the rating scale. Participants can register a score by pointing with the HMD motion controller as shown in Figure 3.4 (b).

Data Recording

During playback, all interactions and decisions are logged in a single statistics file written to disk at regular intervals of 10 seconds. The data logged includes the position (expressed in world coordinates) and orientation (expressed as three Euler angles about the principal axes) of the user at each rendered frame. Each component in the rendering pipeline reports performance statistics independently for each tile, including registered quality ratings, adaptation decisions and bandwidth usage. The logs were written out without affecting the system performance, in a format that can be easily parsed into a JSON object.

3.4.4. Subjective Evaluation Methodology

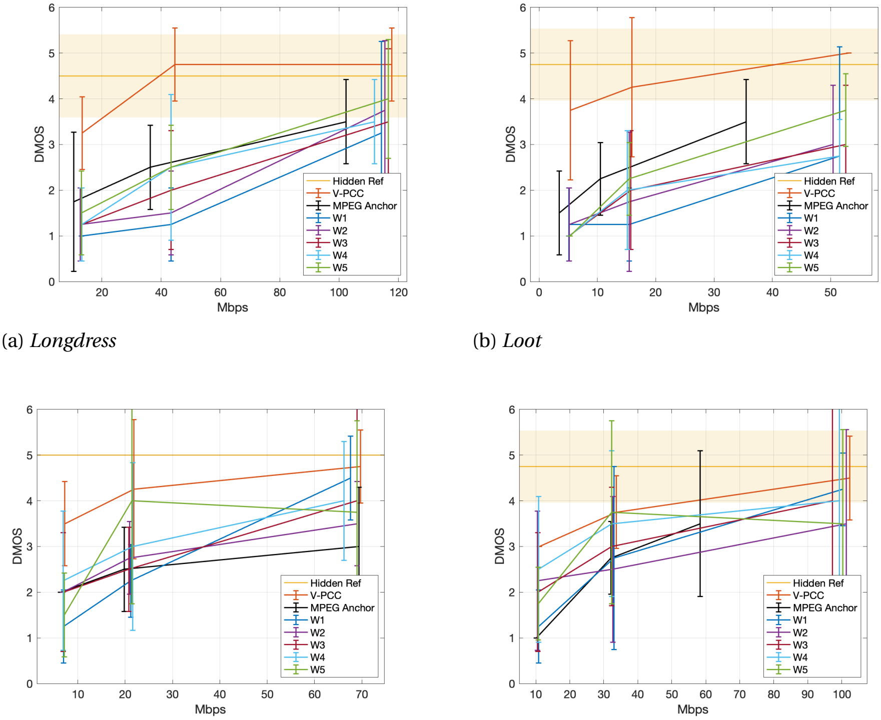
Dataset Preparation

In this work, we use 4 contents from the 8i Voxelized Full Body Dataset [37] shown in figure 3.1, namely *Longdress*, *Loot*, *Red and Black*, and *Soldier*. Similarly to the objective evaluation presented in the last section, we create four non-overlapping tiles by placing virtual cameras around the object on the XZ (floor) plane, at (1,0,0), (0,0,1), (-1,0,0) and (0,0,-1). We draw a vector from the centroid of the point cloud to every point on the surface, and we use the minimum vector dot product of these vectors with the 4 virtual cameras to assign a tile number to each point in the cloud.

The MPEG Anchor codec [11] is used to evaluate the impact of user adaptation. For each adaptation set, the tiles are encoded using the all-intra configuration, with octree depths from 7 to 10, and the JPEG quantization parameter varying from 25 to 95 in increments of 10. To measure the performance of viewport adaptive streaming, the source point clouds are additionally encoded under the same codec configurations. Moreover, we use V-PCC to provide a baseline for our tiling approaches with state-of-the-art rate-distortion performance. In particular, we encode the source point clouds using the Release 9.0 of V-PCC, and the configurations provided in the Common Test Conditions (CTC) for Category 2 All Intra (C2AI) encoding [44]. We select the rate points 2, 3 and 5 and extend it to an additional rate point, using a Texture QP of 9, a geometry QP of 12 and an occupancy precision of 2. The size of the resulting encoded bitstreams are used to set target rate points and bit rate budgets labelled as R1-R4, separately for each of the 4 sequences in the dataset.

Experimental Setup

The experimental setup we used to run the environment was a workstation with a GeForce GTX 2080 Super for the GPU and an Intel Core i9 Skylake-X 2.9GHz CPU. We use Unity



(a) *Longdress*

(b) *Loot*

(c) *Red and Black*

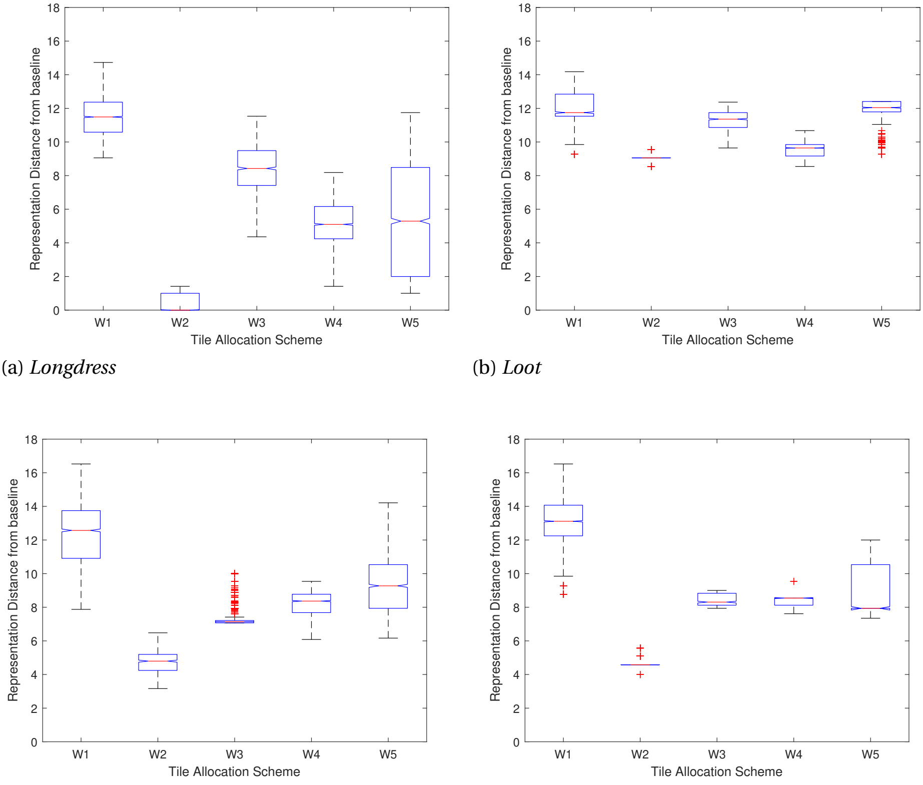
(d) *Soldier*

Figure 3.5: DMOS (solid line) and hidden reference scores (shaded area) against target bit rate from the tile allocation selection study.

2019.4.29f1 (LTS) to support all Open-VR head mounted displays (Oculus Rift or Windows MR) as well as Steam VR enabled HMDs (HTC Vive) with the unity Steam VR SDK. For the experiment we used an HTC Vive Pro eye with an OLED display with 1440 x 1600 pixels per eye and a 110 degree field of view. We chose this HMD as it offers good tracking quality using the Valve base station 2.0 for both the HMD and the motion controllers. The Unity application contained two scenes shown in Figures 3.4 (a), (b). Participants were required to manually trigger scene transitions after every play back loop and register a rating using the HMD motion controllers.

Experimental Design

Prior to the main experiment, we conduct a preliminary user study in order to decide on a subset of bit rate allocation schemes presented in section 3.4.2. This pre-selection is made in order to reduce the total number of stimuli shown to the participants at the main experiment, maintaining a within-subjects design and preventing fatigue. In this case, 116 stimuli were generated, considering all combinations of contents (4), tile allocation strategies plus the baselines (5+2), and the target rate points (4), including hidden



(a) *Longdress*

(b) *Loot*

(c) *Red and Black*

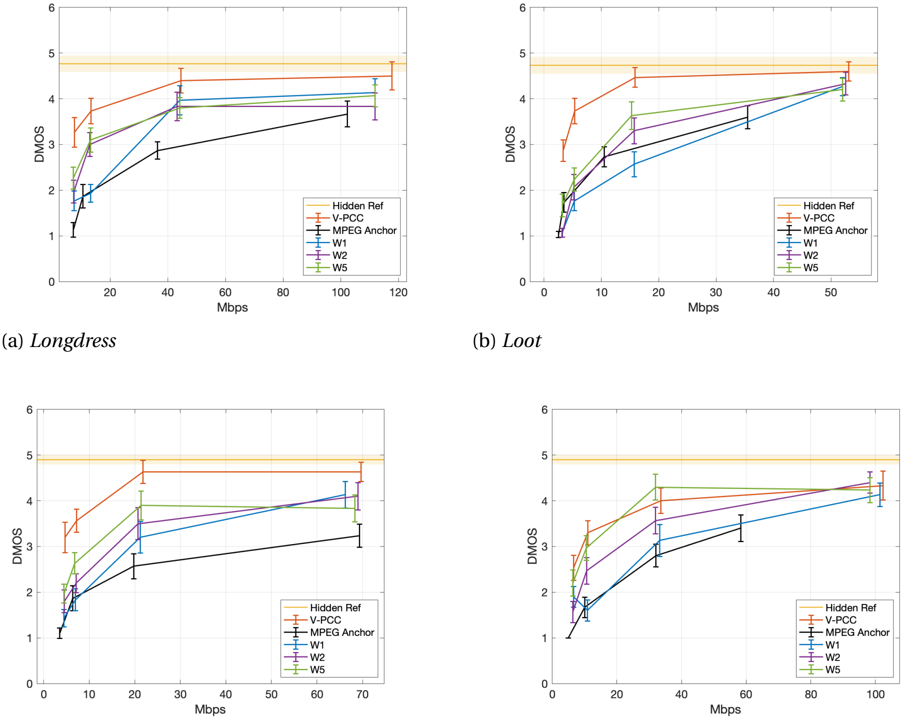
(d) *Soldier*

Figure 3.6: Distance between representation vectors generated by each codec for all users at rate R2 in the tile allocation selection study.

references (4). A total of 5 participants were recruited for the pre-pilot.

After selecting the three best-performing and most diverse approaches, we proceed to the main experiment, with a total of 84 stimuli. The test was divided into two sessions of 42 stimuli, each separated by a 10 minute break to reduce fatigue and motion sickness, based on participant feedback from the pre-pilot study. The participants were requested to fill in the Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [45] before beginning the experiment and after each session. The SSQ was developed to measure cybersickness in computer simulations and was derived from a measure of motion sickness [45]. A total of 30 participants were recruited (16 males, 14 females), with 15 reporting 1-3 prior VR experiences, 6 participants never used a VR HMD before, and 9 participants declaring to be very experienced with VR.

For both subjective experiments, we choose the Absolute Category Rating test method with Hidden References (ACR-HR), following the ITU-T Recommendation P.910 [46]. The point cloud sequences were presented to the participants one at a time and were rated



(a) *Longdress*

(b) *Loot*

(c) *Red and Black*

(d) *Soldier*

Figure 3.7: DMOS (solid line) and Hidden Reference (shaded area) against achieved bit rate, expressed in Mbps.

independently. The participants were asked to observe a loop of 300 frames of each dynamic point cloud sequence, played back at 30 fps for a minimum of 10 seconds, and rate the corresponding visual quality on a scale from 1 to 5 (*1-Bad*, *2-Poor*, *3-Fair*, *4-Good*, and *5-Excellent*). Each sequence was rendered with a randomized initial rotation to prevent bias and encourage user movements.

Every experiment was split in three stages, namely, screening, training and testing. During screening, the color vision of the participants was checked using the Isihara chart, according to the ITU-T Recommendation P.910 [46]. During training, 3 versions of a content not shown at the test were employed, depicting examples of *1-Bad*, *5-Excellent* and *3-Fair*. This stage was essential in order to help the participants to familiarize with the viewing conditions and the test setup, demonstrating also the use of the HMD motion controller to register a rating. During testing, the order of the displayed stimuli was randomized per participant and per session, and the same content was never displayed twice in a row to avoid temporal referencing bias. Three dummy samples were added at the beginning of each testing session in order to ease participants into the task, with the corresponding scores subsequently discarded.

Data Processing

Outlier detection was performed on the individual quality scores collected by the subjects, according to the ITU-T Recommendation P.913 [46], with the recommended threshold values $r_1 = 0.75$ and $r_2 = 0.8$. After outlier detection, a Mean Opinion Score (MOS) was computed for each stimulus, independently per viewing condition. The associated 95% Confidence Intervals (CIs) were computed assuming a Student's t-distribution. Additionally, the Differential MOS (DMOS) was obtained by applying hidden reference removal, following the procedure described in the ITU-T Recommendation P.913 [46].

3.4.5. Results

In this section, we present the results of the quality evaluation experiments. We motivate the selection of stimuli for the experiment. We present an analysis of the objective and subjective scores by codecs, contents and rate points. Finally, we analyze user navigation data and provide the results of simulator sickness and fatigue.

Pre-pilot: Tile Rate Allocation Selection

In Figure 3.5, the DMOS collected during the pre-pilot user study across coding conditions and tile allocation strategies, are depicted. Moreover, in Figure 3.6, we present a boxplot of the Euclidean distance between representation vectors obtained for every bit rate allocation scheme and our non-adaptive baseline. In particular, representation vectors are computed for every point cloud frame, and contain the quality levels that are selected for all corresponding tiles. This information is extracted by the recorded data that are logged at playback time and during subjective evaluation of users. The results refer to all users and all contents encoded at a target rate point R2, resulting in bit rates between 5 MBit/sec to 13 Mbit/s.

Based on Figures 3.5 and 3.6, W5 achieves the highest perceived quality, while it exhibits the largest distribution of representation vector distance. The W1 achieves the lowest perceived quality and the highest absolute representation vector distance, whereas W2 achieves medium quality and has the lowest absolute representation vector distance, indicating that it is the closest to the baseline naive codec, as expected. The performance of W3 and W4 is generally higher than W1, but it is outperformed by both W2 and W5 approaches. Based on the results of this study, we assume that W5 represents the optimal compromise, where tiles of similar utility are afforded similar bit budgets. Meanwhile, W1 and W2 schemes represent the maximum and minimum quality differences between adjacent tiles, respectively. Thus, for the user study, we decided to evaluate W1, W2 and W5 schemes, in addition to the baseline encoded content.

Subjective Scores

Based on the collected scores of this experiment, no outliers were identified. Thus, the entirety of the subjective scores are employed in the subsequent analysis. In Figure 3.7, the results of the subjective quality assessment of the contents, are illustrated. In particular, the DMOS associated with the compressed contents are shown with solid lines, along with the relative CIs. The hidden reference scores for each content are represented with a solid yellow line to indicate the average score, along with a shaded area to represent the corresponding CIs. From the charts in Figure 3.7 we observe that the V-PCC codec generally achieves the best quality. Among the MPEG Anchor-based solutions, W5 usually achieves the best quality, whereas W1 has generally the lowest quality. At R1-R3, W1 results in large quality differences amongst adjacent tiles, often at different octree depths; participants reported that they sometimes found the resulting reconstruction unpleasant. Using a utility-weighted distribution of the bit-budget across tiles, as in W5, appears to result in acceptable quality differences across adjacent tiles, while optimizing the representation of tiles facing the user.

In general, we observe that, at lower bit rates, the MPEG Anchor achieves similar scores to W1, indicating that participants prefer lower-quality, uniform content over large quality differences amongst adjacent tiles. W5 achieves the best quality at rate points R1-R3 for most content under test. For all sequences except *Soldier*, we observe that none of the tile allocation strategies achieve the same quality as V-PCC, independently of the rate point. Moreover, for contents *Red and Black* and *Soldier*, none of the compression solutions used in the study reach transparent quality with respect to the reference content. For contents *Longdress* and *Loot*, V-PCC is able to achieve statistically equivalent quality with respect to the uncompressed reference at the highest bit rate.

A Shapiro-Wilk normality test issued on the entirety of the subjective scores indicates that they don't follow a normal distribution ($W = 0.9093$, $p < 0.001$). Thus, non-parametric statistical tools were applied to perform an exploratory data analysis and understand whether statistical differences could be found amongst the different conditions being evaluated. To compare the different codecs and bit rate allocation strategies being tested, we first conduct a Friedman's test to check if there are any groups with significant differences ($\chi^2 = 888.69$, $p < .001$). We then conduct a pairwise post-hoc analysis with the Wilcoxon signed rank test with Bonferroni correction. The results are shown in Table 3.7.

Table 3.7: Pairwise post-hoc test codecs and bite rate allocation strategies, using Wilcoxon signed-rank test with Bonferroni correction.

Codec	Z	p	r
V-PCC – MPEG Anchor	18.261	<.001	0.589
V-PCC – W1	15.665	<.001	0.506
V-PCC – W2	14.412	<.001	0.465
V-PCC – W5	12.557	<.001	0.405
MPEG Anchor – W1	-9.846	<.001	0.318
MPEG Anchor – W2	-14.000	<.001	0.452
MPEG Anchor – W5	-16.016	<.001	0.517
W1 – W2	-6.347	<.001	0.205
W1 – W5	-10.392	<.001	0.335
W2 – W5	-6.110	<.001	0.197

It can be observed that all pairwise codec comparisons exhibit statistically significant differences with varying effect sizes. V-PCC shows a large effect size with respect to all other codecs ($r = 0.405$ to 0.589), as expected, due to its superior rate distortion performance and high encoder complexity. All three adaptive tile selection strategies show a medium to large effect size with respect to the baseline naive MPEG Anchor codec ($r = 0.318$ to $r = 0.517$). W5 in particular shows the largest effect size and Z value with respect to the naive baseline codec. This, combined with results shown in Figure 3.7, indicates that W5 yields significantly better visual quality as compared to the baseline across all sequences in the dataset.

To compare the four sequences and check if there are any statistically significant differences in the dataset, we first ran a Friedman test. The results confirmed that content has a significant effect on the recorded scores ($\chi^2 = 70.24$, $p < .001$). Post-hoc analysis using the Wilcoxon signed-rank test with Bonferroni correction further confirmed that the *Longdress* sequence had statistically significant differences in scores, as shown in Table 3.8, albeit with small to medium effect sizes ($r = 0.176$ to $r = 0.2$). The remaining three contents do not show statistically significant differences. This can be partially explained by the fact that the V-PCC encoding parameters, which were used to define the target bit rates, lead to larger values in terms of bits per reference points for *Longdress* with respect to the other contents, as shown in Table 3.9.

Table 3.8: Pairwise post-hoc test on contents, using Wilcoxon signed-rank test with Bonferroni correction.

Content	Z	p	r
Longdress – Loot	6.086	<.001	0.176
Longdress – Red and Black	7.060	<.001	0.204
Longdress – Soldier	6.933	<.001	0.200
Loot – Red and Black	0.717	0.473	0.021
Loot – Soldier	0.564	0.573	0.016
Red and Black – Soldier	-0.343	0.732	0.01

Table 3.9: Target bit rates expressed in terms of bits per reference point averaged across all frames in a sequence.

Content	R1	R2	R3	R4
Longdress	0.32	0.56	1.87	4.93
Loot	0.15	0.24	0.70	2.34
Red and Black	0.22	0.35	1.05	3.35
Soldier	0.22	0.36	1.09	3.33

In order to assess if the selected bit rates showed statistically significant differences, we first ran a Friedman test. The results confirmed that there is indeed a significant effect of rate point on scores ($\chi^2 = 1366.21$, $p < .001$). Post-hoc analyses using the Wilcoxon signed-rank test with Bonferroni correction revealed statistically significant differences among all pairwise comparisons, with mostly large effect sizes, as reported in Table 3.10. The comparison between R3 and R4 show the lowest effect size ($r = 0.345$), in line with the results in Figure 3.7. For all sequences in the dataset, we observe similar performance for the V-PCC codec at R3 and R4, with the largest performance gap for the *Soldier* sequence.

In order to further confirm the impact of our tiling adaptive strategies on the scores

Table 3.10: Pairwise post-hoc test on bit rate targets, using Wilcoxon signed-rank test with Bonferroni correction.

Target Bit Rate	Z	p	r
R1 – R2	-14.898	<.001	0.43
R1 – R3	-20.754	<.001	0.599
R1 – R4	-21.034	<.001	0.607
R2 – R3	-18.475	<.001	0.533
R2 – R3	-20.187	<.001	0.583
R3 – R4	-11.941	<.001	0.345

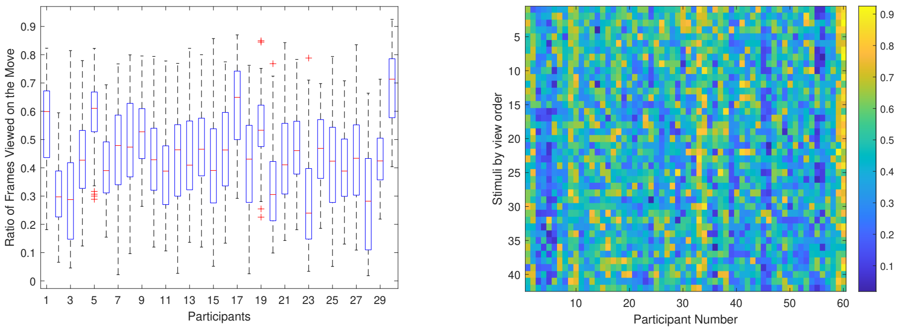
with respect to the baseline MPEG Anchor, we perform a Kruskal-Wallis ANOVA test on the two unmatched groups. A significant effect was found on the scores ($\chi^2 = 93.11$, $p < .001$). This demonstrates the statistically significant overall performance gain across all tiling allocation strategies. In general, W5 yields the best results for content encoded with the baseline MPEG Anchor codec, as shown in Figure 3.7 and Table 3.7.

USER NAVIGATION

We recorded the position and rotation of the user's viewport for each participant and stimuli in the user study. In order to analyze user movement patterns, we defined a motion threshold to quantify the total frames viewed *on the move* for each stimuli. Frames were classified as viewed *on the move* if participants either translated more that 0.05cm from the previous frame or rotated their viewport by more than 0.573 degrees along any axis from the previous frame, based on the findings of Rossi et al. [47].

The overall distribution of the ratio of frames viewed *on the move* by each participant across all 84 stimuli, is shown in Figure 3.8 (a). We observe a lot of variation in the median and the distribution of overall movement, essentially indicating that every participant's motion is unique. In order to see if participants' previous VR experience might have affected their navigation behaviour, we performed a comparison between the reported VR familiarity with respect to the average ratio of frames *on the move*; however, no significant correlation at 5% significance level was found (Spearman's $\rho = 0.264$, $p = 0.160$). In order to check if fatigue had an impact on participant movement over time, we compare the motion ratio for each combination of participant and session number in chronological order, as shown in Figure 3.8 (b). Here we treat each session as separate resulting in a total of 60 participant sessions. The results show that, no particular patterns of reduction in motion ratio can be observed over time, due to participant fatigue.

In order to check if adaptive playback using any of the three bit rate allocation strategies has a statistically significant affect on motion ratios, we compare it against naive playback; this includes the MPEG Anchor codec, V-PCC, and the hidden references. We performed a Kruskal-Wallis ANOVA test and observed a significant affect on motion ratios ($\chi^2 = 22.54$, $p < .001$). We observed that participants exhibit higher motion ratios while viewing adaptive content (median: 0.472, min: 0.018, max: 0.925) as compared to non-adaptive content (median: 0.407, min: 0.022, max: 0.921). When we compare the median motion ratios based on the scores assigned by participants, we observe that all scores have similar median motion ratios of 0.44. However, for stimuli rated as *fair* and



(a) Box plot

(b) Viewing order

Figure 3.8: Ratio of frames viewed *on the move*, across all participants.

good, we observe a larger range of motion ratios (median: 0.442, min: 0.025 and max: 0.925) as compared to all other scores (median: 0.439, min: 0.066 and max: 0.857), indicating that users might have interacted more to distinguish between these two quality ratings.

Simulator Sickness and Fatigue

Based on the SSQ filled by every subject, we observe low post-exposure total severity scores for cybersickness, as defined in [45, 48], after each session of the experiment based on the recommendations in [48, 49]; the [mean, median, std] of baseline: [5.24, 0, 10.29], after session 1: [16.83, 13.09, 16.89], and after session 2: [22.32, 18.70, 19.20].

Looking at some of the individual symptoms reported by participants in the course of the experiment, we found no statistically significant differences in fatigue ($\chi^2 = 4.31$, $p = 0.116$). We do observe statistically significant differences in eyestrain ($\chi^2 = 19$, $p < .001$) and general discomfort ($\chi^2 = 19.18$, $p < .001$). However, no participants reported severe symptoms on any of the SSQ questions after completing the experiment. No participants dropped out of the experiment on account of cybersickness.

In order to check if fatigue influenced the navigation behaviour of the users, we compare the average ratio of frames *on the move* with the total severity as reported in the SSQ, using Pearson's correlation coefficient. We find no significant effect at 5% significance level (Baseline: $\rho = -0.302, p = 0.105$; After Session 1: $\rho = -0.291, p = 0.119$; After Session 2: $\rho = -0.337, p = 0.067$).

3.5. Discussion

Results of our subjective evaluation showed clear benefits in using adaptive streaming strategies when compared with the non-adaptive baseline. Our results seem to indicate that smoother quality transitions between adjacent tiles are to be preferred with respect to greedy approaches, which lead to obvious discontinuities in the appearance of the content. As our evaluation focuses on point cloud contents depicting humans, boundary artifacts might be more annoying if such boundary lies on regions of interest, such as faces, or influences the structural integrity of the reconstruction (e.g., disappearing fingers). We observe the best quality gains using W5, which at high bit rates approaches the quality offered by the state of the art offline codecs such as V-PCC. Such strategy ensures that transitions among adjacent tiles are not too drastic, while maximizing the quality of high utility tiles.

The performance of the adaptive streaming approach is bounded by the codec that is selected to encode each tile. Previous studies have shown that the MPEG Anchor leads to significantly lower visual quality with respect to V-PCC [3]; thus, it is not surprising that the adaptive strategies adopted in this study do not achieve the same visual quality as V-PCC. Moreover, we can observe that our adaptive strategy does not reach transparent quality, i.e., achieving statistically equivalent results with respect to the hidden reference. As mentioned in previous sections, V-PCC is not suitable for real-time encoding [1, 42]; new low-complexity solutions should be devised to further increase the performance gains of adaptive streaming.

In general, segmenting the point cloud spatially leads to larger encoded payload sizes due to a loss of entropy and lower compression efficiency. While this trade-off appears to yield significant improvements in perceived quality using the MPEG Anchor codec, further assessment is required once more low-latency point cloud codecs become available, to ensure performance gains are maintained.

3.5.1. Limitations

In this chapter, we proposed a low-complexity algorithm in order to partition the content in multiple tiles, which can then be encoded at different qualities. Specifically, the selected spatial segmentation is designed for live captured point clouds; as such, it exploits information from acquisition sensor placement to infer surface orientation of spatial segments of a given point cloud content. In this work, due to the lack of point cloud datasets with labelled acquisition information, and to favor reproducibility, we emulated our approach on a popular point cloud dataset [37], which was employed in MPEG standardization activities [7], and has been successfully used in the past to test adaptive streaming for point cloud contents, both objectively [1] and subjectively [4]. We operated under the assumption that the content could be separated into an even number of non-overlapping, opposite-facing tiles, in order to apply our utility function. While this assumption appears to be valid for the prerecorded point clouds used in this study, the performance of our proposed system with real-time point clouds may vary, depending on the accuracy and noise level of the acquisition sensors. Alternative spatial segmentation schemes, such as body part segmentation based on pose estimation, can also lead to further gains in perceived quality, as participants reported preferring content where the reconstruction's face has a high-quality representation.

In the objective evaluation, we assumed that the client has omniscient knowledge of user behaviour, and is able to a) request and b) receive the tiles with the highest utility at any time instance. The screenshots were recorded after all adaptation decisions were taken at the given viewport. Thus, our objective evaluation represents an upper bound of the performance of the tiling approaches.

In the subjective evaluation, the adaptation was performed based on the last known viewport location and orientation. This approach was acceptable as participants were able to navigate the scene using physical movements only. In VR applications where participants can move faster than physical movements through the use of a controller or teleporting to a different location in the scene, our approach will have to be adapted to account for the uncertainty in the user's future viewport location and orientation when the point cloud frame is rendered.

To evaluate the proposed adaptive streaming strategies, the network conditions and available bandwidth were set based on the CTC defined by the MPEG standardization activity [44]. While these cover a wide range of bit rates (3 - 117 Mbit/sec), the bit rate budget was constant for the duration of the playback sequence. The constant bit rate budget was selected to avoid introducing biasing factors in the subjective evaluation, as a variable bit rate with adaptive tiling might have been a confounding factor for both DMOS and SSQ. In order to adequately assess the performance of the proposed adaptive streaming approach, further analyses in adverse varying network conditions are required to ensure the performance gains are maintained.

3.5.2. Display Device Considerations

While developing the frame work and testing playback quality, we found that current Timewarp modules such as the Oculus Asynchronous Spacewarp and SteamVR Motion Reprojection that are used to boost the perceived framerate exacerbate participant's cybersickness while viewing dynamic point cloud content. This is primarily caused by a difference in framerate between the point cloud sequence (30fps) and the Unity game engine in VR mode (90fps). This, combined with the relatively lower tracking accuracy offered by the Oculus Rift DK1 Constellation tracking system, rendered the system unusable on account of cybersickness while using the Oculus Rift DK1 HMD and playing back point clouds larger than a million points per frame. We were able to achieve optimal performance using the HTC Vive pro with the lighthouse tracking system and a target framerate of 90fps with SteamVR VSync turned on. While using a regular monitor, we observe smooth performance with the application running at 30fps and a significant reduction in GPU load.

3.5.3. Towards Deployment in a Social VR Application

The framework we propose is designed to improve real-time transmission of dynamic point cloud contents for teleimmersive applications, such as social VR. In our evaluation approach, we consider an omniscient server: the client requests tiles for the upcoming frame based on the last known viewport location. However, in order to accommodate larger segment sizes and additional transmission latency from the server, viewport prediction is required. Inaccuracy in such a component can further degrade the perceived quality of tiled adaptive streams of point clouds. As such, this work establishes an upper bound on potential performance gains from tiled adaptive point cloud streaming. We envision our approach to be suitable for systems using chunked HTTP transfers and fragmented ISOBMFF/MP4 as a transport stack for live streaming, such as [16]. Codecs that offer progressive decoding, such as [11], can achieve further gains at the server side by creating multiple representations in a single encode cycle. In the next chapter, we apply our framework to a live two-user VR remote communication application.

3.6. Conclusion

In this chapter, we performed an objective and subjective quality evaluation of user-adaptive dynamic streams of point clouds. We also presented a software framework to playback tiled point cloud streams in 6DoF VR and adapt to user interactions in real-time.

To address research question **R2** *(How can we optimally allocate the available bandwidth across independently decodable spatial segments?)*, we proposed novel tile allocation strategies that consider the user's position with respect to the point cloud content to assign different quality levels to independently encoded tiles. To address *R2.1 (Does user-centred tiled adaptive streaming offer significant gains to reconstruction quality and bandwidth savings?)*, we performed a user study comparing tiled adaptive streaming of point clouds to non-adaptive approaches. At the lowest bit rates (3 to 8 Mbit/sec), we observe that a greedy approach to bit rate allocation per tile produces drastic differences in quality in adjacent tiles, resulting in a low perceived quality comparable to the non-adaptive approach. We also demonstrate significant performance gains in objective quality and bitrate savings compared to non-adaptive streaming. The results of the objective evaluation also demonstrate that there is a significant influence of navigation paths on objective quality, indicating a need for evaluating future adaptive streaming solutions using navigation data rather than pre-defined fixed paths. In addition, we released a dataset of user navigation data that is publicly available here: https://github.com/cwi-dis/6DoF-HMD-UserNavigationData for the community to use in future user-centred research. To address research question *R2.2 (What are the acceptable quality levels amongst adjacent tiles to maximize the quality of the final point cloud reconstruction?)* we evaluated several tile rate allocation strategies. A utility-weighted bit rate allocation was observed to provide acceptable quality differences amongst adjacent tiles at the target bit rates used for evaluation and yielded the highest gains over baseline non-adaptive playback across bit rates and contents in the dataset under test.

Our approach is codec-agnostic and standards-compliant, allowing interoperability with existing 6DoF remote communication pipelines. In the next chapter, we will integrate this approach with a point cloud based remote communication system, with real-time reconstructions of remote users, to evaluate tiled user-adaptive streaming in a live multiparty communication scenario, with realistic network conditions and server-side constraints of commodity hardware.

REFERENCES

[1] J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, and H. Hellwagner, *Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression,* in *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19 (Association for Computing Machinery, New York, NY, USA, 2019) pp. 2405–2413.

[2] G. Cernigliaro, M. Martos, M. Montagud, A. Ansari, and S. Fernandez, *PC-MCU: Point Cloud Multipoint Control Unit for Multi-User Holoconferencing Systems,* in *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* (Association for Computing Machinery, New York, NY, USA, 2020) p. 47–53.

[3] 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)* (IEEE, 2020) pp. 127–136.

[4] 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)* (2020) pp. 1–6.

[5] V. Baroncini, P. Cesar, E. Siahaan, I. Reimat, and S. Subramanyam, *Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression,* ISO/IEC JTC1/SC29/WG11 M41786 (2017).

[6] C. Cao, M. Preda, and T. Zaharia, *3d point cloud compression: A survey,* in *The 24th International Conference on 3D Web Technology*, Web3D '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 1–9.

[7] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, *et al.*, *Emerging MPEG Standards for Point Cloud Compression,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 133 (2019).

[8] E. Pavez, P. A. Chou, R. L. de Queiroz, and A. Ortega, *Dynamic polygon clouds: representation and compression for vr/ar,* APSIPA Transactions on Signal and Information Processing **7**, e15 (2018).

[9] R. L. de Queiroz and P. A. Chou, *Compression of 3d point clouds using a region-adaptive hierarchical transform,* IEEE Transactions on Image Processing **25**, 3947 (2016).

[10] MPEG3DG and Requirements, *Call for proposals for point cloud compression,* ISO/IEC JTC1/SC29 WG11 N16732, Geneva, CH (2017).

[11] 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 (2016).

[12] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, *Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units,* in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14 (Association for Computing Machinery, New York, NY, USA, 2014) p. 637–644.

[13] A. D. Wilson and H. Benko, *Projected augmented reality with the roomalive toolkit,* in *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, ISS '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 517–520.

[14] R. Mekuria, P. Cesar, I. Doumanis, and A. Frisiello, *Objective and subjective quality assessment of geometry compression of reconstructed 3D humans in a 3D virtual room,* Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, 95991M (2015).

[15] S. N. B. Gunkel, R. Hindriks, K. M. E. Assal, H. M. Stokking, S. Dijkstra-Soudarissanane, F. t. Haar, and O. Niamut, *Vrcomm: An end-to-end web system for real-time photorealistic social vr communication,* in *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 65–79.

[16] J. Jansen, S. Subramanyam, R. Bouqueau, G. Cernigliaro, M. Martos Cabré, F. Pérez, and P. Cesar, *A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH,* in *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20 (ACM, New York, NY, USA, 2020).

[17] J. v. d. Hooft, M. T. Vega, T. Wauters, C. Timmerer, A. C. Begen, F. D. Turck, and R. Schatz, *From capturing to rendering: Volumetric media delivery with six degrees of freedom,* IEEE Communications Magazine **58**, 49 (2020).

[18] M. Hosseini and C. Timmerer, *Dynamic Adaptive Point Cloud Streaming,* in *Proceedings of the 23rd Packet Video Workshop*, PV '18 (ACM, New York, NY, USA, 2018) pp. 25–30.

[19] J. Park, P. A. Chou, and J. Hwang, *Rate-Utility Optimized Streaming of Volumetric Media for Augmented Reality,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 149 (2019).

[20] L. He, W. Zhu, K. Zhang, and Y. Xu, *View-dependent streaming of dynamic point cloud over hybrid networks,* in *Advances in Multimedia Information Processing – PCM 2018* (Springer International Publishing, Cham, 2018) pp. 50–58.

[21] R. B. Rusu, *3d is here: Point cloud library,* Robotics and Automation (ICRA), 2011 IEEE International Conference (2011).

[22] A. Nguyen and B. Le, *3d point cloud segmentation: A survey,* in *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (2013) pp. 225–230.

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

[24] E. Alexiou and T. Ebrahimi, *On the performance of metrics to predict quality in point cloud representations,* in *Applications of Digital Image Processing XL*, Vol. 10396, edited by A. G. Tescher, International Society for Optics and Photonics (SPIE, 2017) pp. 282 – 297.

[25] 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)* (2018) pp. 1–6.

[26] 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).

[27] S. Perry, H. P. Cong, L. A. da Silva Cruz, J. Prazeres, M. Pereira, A. Pinheiro, E. Dumic, E. Alexiou, and T. Ebrahimi, *Quality evaluation of static point clouds encoded using mpeg codecs,* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020) pp. 3428–3432.

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

[29] 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)* (2020) pp. 1–6.

[30] 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)* (2017) pp. 1–6.

[31] A. Javaheri, C. Brites, F. M. B. Pereira, and J. M. Ascenso, *Point Cloud Rendering after Coding: Impacts on Subjective and Objective Quality,* IEEE Transactions on Multimedia , 1 (2020).

[32] E. Dumic, F. Battisti, M. Carli, and L. A. da Silva Cruz, *Point cloud visualization methods: a study on subjective preferences,* in *2020 28th European Signal Processing Conference (EUSIPCO)* (2021) pp. 595–599.

[33] 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)* (2017) pp. 1–6.

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

[35] E. Alexiou, N. Yang, and T. Ebrahimi, *PointXR: A Toolbox for Visualization and Subjective Evaluation of Point Clouds in Virtual Reality,* in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)* (2020) pp. 1–6.

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

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

[38] E. Alexiou, I. Viola, T. Borges, T. Fonseca, R. 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** (2019), 10.1017/ATSIP.2019.20.

[39] 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 Transactions on circuits and systems for video technology **22**, 1669 (2012).

[40] R. A. Fisher *et al.*, *Statistical methods for research workers.* Statistical methods for research workers. (1934).

[41] J. T. Kost and M. P. McDermott, *Combining dependent p-values,* Statistics & Probability Letters **60**, 183 (2002).

[42] S. Subramanyam, I. Viola, A. Hanjalic, and P. Cesar, *User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling,* in *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 3669–3677.

[43] Q.-Y. Zhou, J. Park, and V. Koltun, *Open3D: A modern library for 3D data processing,* arXiv:1801.09847 (2018).

[44] MPEG 3DG and Requirements, *Complementary PCC Test Material,* ISO/IEC JTC1/SC29 WG11 Doc. N16716, Geneva, CH (2017).

[45] 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**, 203 (1993).

[46] ITU-T P.910, *Subjective video quality assessment methods for multimedia applications,* International Telecommunication Union (2008).

[47] S. Rossi, I. Viola, J. Jansen, S. Subramanyam, L. Toni, and P. Cesar, *Influence of Narrative Elements on User Behaviour in Photorealistic Social VR,* in *Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '21)*, MMVE '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 1–7.

[48] P. Bimberg, T. Weissker, and A. Kulik, *On the Usage of the Simulator Sickness Questionnaire for Virtual Reality Research,* in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020) pp. 464–467.

[49] K. M. Stanney, R. S. Kennedy, and J. M. Drexler, *Cybersickness is not simulator sickness,* Proceedings of the Human Factors and Ergonomics Society Annual Meeting **41**, 1138 (1997), https://doi.org/10.1177/107118139704100292 .

4

Optimizing the delivery of tiled streams of dynamic point clouds in VR remote communication

In the previous chapter, we proposed a low-complexity tiling and adaptive point cloud streaming approach. This chapter deploys this in a fully functional VR remote communication pipeline. The approach is compared against traditional network adaptive streaming and an uncompressed baseline. We propose a novel evaluation methodology along with a training task as part of the protocol. We then evaluate the quality of communication, visual quality, system performance and task completion. We performed a subjective evaluation with 33 users to compare the different streaming conditions with a neck exercise training task. As a baseline, we use uncompressed streaming requiring approximately 300 Mbps and our solution achieves similar visual quality with tiled adaptive streaming at 14 Mbps. We also demonstrate statistically significant gains in the quality of interaction and improvements to system performance and CPU consumption with tiled adaptive streaming. This work is meant to address the core hypothesis of this thesis and determine the quality impact of adaptive delivery optimizations.

This chapter is based on:

1. ***S. Subramanyam**, I. Viola, J. Jansen, E. Alexiou, A. Hanjalic and P. Cesar. Evaluating the Impact of Tiled User-Adaptive Real-Time Point Cloud Streaming on VR Remote Communication. Proceedings of the 30th ACM International Conference on Multimedia, October 10-14, 2022, Lisboa, Portugal*
2. *J. Jansen, **S. Subramanyam**, R. Bouqueau, G. Cernigliaro, M. Cabre, F. Perez, and P. Cesar. 2020. A pipeline for multiparty volumetric video conferencing: transmission of point clouds over low latency DASH. In Proceedings of the 11th ACM Multimedia Systems Conference (MMSys '20). Association for Computing Machinery, New York, NY, USA, 341–344. https://doi.org/10.1145/3339825.3393578*

4.1. Introduction

Immersive Virtual Reality (VR) applications offer an increased sense of presence and immersion. These applications have emerged as a promising alternative for remote communication and telepresence [1–8]. They allow users to employ both verbal and non-verbal communication in a shared virtual space. Previous work in the field has demonstrated that realistic user reconstructions improve immersion and communication [9, 10] as compared to avatars.

In this thesis, we focus on the point cloud format to represent realistic user reconstructions. However, these reconstructions are too large to transmit uncompressed over bandwidth-limited networks at framerates suitable for VR remote communication [11]. From our experiments, we found that we would require ca. 300 Mbps to transmit these reconstructions uncompressed at 15 frames per second (fps). In order to alleviate this requirement, the previous chapter looked into adapting the point cloud stream to the user's viewport location and orientation in order to optimize how the available bandwidth is spent. This is done by prioritizing objects or surfaces facing the viewer and reducing the wastage of bandwidth on surfaces or objects that are occluded or outside the viewport [12–18]. In this chapter, we present a VR remote communication pipeline that has been optimized to deliver point cloud user reconstructions adaptively. We evaluate adaptive streaming for live communication with real-time point cloud capture.

The contributions of this chapter are threefold:

1. We present the first implementation of a two-user social VR system with real-time adaptive streaming of user reconstructions shown in Figure 4.1.

2. We propose and employ a novel evaluation methodology using a training task to assess delivery optimizations.

3. We present an evaluation of the impact of tiled adaptive streaming in terms of quality of communication, visual quality and computational resource consumption

To address research question *R3.1 (How does tiled user-adaptive point cloud streaming impact the perceived quality of remote user reconstruction?)* we set out to assess the impact of tiled adaptive streaming on the quality of communication, visual quality and subjective task related performance. We constructed a social VR pipeline [4, 19] with additional modules for network adaptive and tiled adaptive streaming. Two confederate users were recruited and trained to play the role of a trainer in every experiment session while 33 users (16 females, 17 males) were recruited to play the role of trainee. The participants were asked to learn and perform three neck exercises during the session. We evaluate and compare tiled adaptive streaming (*TA*) with traditional network adaptive streaming (*NA*) and baseline uncompressed streaming in a functional live VR remote communication system. To address research question *R3.2 (What is the computational overhead of using tiled adaptive point cloud streaming?)* we profile system performance and resource consumption during the course of the experiment across the three streaming conditions.

From our results, we observe statistically significant improvements to Quality of Interaction/ Quality of Communication (QoI). Our approach results in significant improvements to visual quality where at 14Mbps we get similar visual quality as compared to uncompressed point clouds streamed at ca. 300Mbps. We also see a reduction in CPU utilization and an improvement to playback performance. On the other hand, we do not observe any statistically significant change to task-related experience.

In the remainder of the chapter, we first discuss the related work VR remote communication systems, point cloud delivery and communication evaluation. This is followed by a description of the study design and the implementation of the social VR pipeline used in the experiment. We then present the results of the user study. This is followed by a discussion of the results and how they can be applied to future VR remote communication systems along with the conclusions drawn from this work.

4.2. Related Work

4.2.1. VR Remote Communication using Point Clouds

Advances in low-latency streaming and volumetric point cloud delivery mechanisms have led to the emergence of novel teleimmersion systems that allow distributed remote users to communicate as themselves in a shared environment with realistic user reconstructions. Microsoft released the RoomAlive Toolkit for creating interactive Augmented Reality (AR) experiences [6, 20]. Mekuria *et al.* proposed a teleimmersive system that blends avatar representations and photo-realistic reconstructions of users in a shared virtual environment [9]. Gunkel *et al.* introduced VRComm [5], a web-based social VR communication system using photo-realistic user reconstructions that was evaluated using both simulations and subjective studies. They present a video-based transmission approach as an initial step toward delivering volumetric content. Cernigliaro *et al.* propose a point cloud multi-point control unit for optimizing holo-conferencing systems [21]. None of these systems implements real-time tiled adaptive streaming of live-captured point clouds. In this work, we present a system design with network adaptive *(NA)* and tiled user-adaptive *(TA)* streaming. We transmit tiled point cloud user reconstructions at fixed target bitrates to assess the experience without the influence of a volatile network.

4.2.2. Point Cloud Delivery

Compression Standards

Point cloud compression has received significant research attention in recent years with the launch of two new MPEG compression standards [22]. The V-PCC standard codec for dynamic point clouds that projects point clouds geometry and attributes onto separate 2D patches that are them packed into video tracks along with the occupancy. These video tracks are then encoded using legacy video codecs making this approach suitable for relatively dense and uniform distribution of points. The G-PCC standard codec uses an octree space partitioning structure to code geometry and can be optionally combined with an additional surface reconstruction step using the TriSoup approach [23]. G-PCC also includes several modules for attribute coding, the lowest complexity coder uses the Region Adaptive Hierarchical Transform (RAHT) [24]. This codec is targeted at irregular sparse distributions of points making it suitable for live captured point clouds. However, both codecs introduce high complexity encoding making them unsuitable for real-time communication. At the start of the MPEG standardization activity an anchor codec proposed by Mekuria *et al.* [3] was introduced. This codec utilizes octree occupancy to code geometry and scans attributes to map them to a 2D grid and encodes them with legacy JPEG image compression. This approach offers low encode and low decode complexity making it suitable for real-time framerate-sensitive applications such as VR remote communication. In this chapter, we use the anchor codec to encode point cloud tiles at multiple quality levels in real time before streaming.

Adaptive Streaming

Initial works on adaptive streaming of point clouds utilized entire point cloud objects as the basic unit of bandwidth allocation in scenes containing multiple point cloud objects. Hosseini *et al.* [12, 25] propose DASH-PC for dynamic adaptive view aware point cloud streaming. They propose three spatial subsampling techniques to create multiple representations of point cloud objects in a scene. The density of each object representation is used by the client for bitrate allocation based on human visual acuity. Hooft *et al.* [14] propose PCC-DASH, a standards-compliant means for HTTP adaptive streaming. They present three heuristics based on the users viewport and distance to the object to allocate bitrate to different objects in the scene. Different ranking metrics and bitrate allocation heuristics had to be selected for different scenes and user navigation paths.

Another approach used in previous work is to split each point cloud object into tiles that are then used as the unit of bandwidth allocation. Park *et al.* [13] define a utility per tile based on the user's proximity, point cloud surface quality and display device resolution. To account for interactions, they propose a window-based design for the Client Buffer Manager with greedy utility maximization. This type of rasterization and pixel occupancy based approach is not suitable as computing this at every frame is computationally expensive. He *et al.* [26] propose view-dependent streaming over hybrid networks. Each point cloud frame is projected onto the six faces of a bounding cube, with a color and a depth video created per face. The user can request videos that correspond to particular faces of the cube in high quality from the edge node of a bidirectional broadband network, reconstructing the point cloud from the downloaded depth and color videos at the receiver end. This approach requires a redundant extra reconstruction step at each receiver. We instead perform the reconstruction on the sender side in order to generate a self view to embody the user and transmit the reconstruction to all receivers. Li *et al.* [27] propose a joint communication and computational resource allocation framework to stream and decode pre-recorded point clouds. They also propose a QoE metric to guide tile selection based on the users viewport, distance to tile and available quality levels. Lee *et al.* [28] propose GROOT a real-time streaming system to reduce decoding overhead by dividing the point cloud into cells defined by the leaf nodes of an octree represented in a parallel decodable tree. Han *et al.* [29] propose ViVo using a similar approach and employ machine learning models to predict viewport movement. Liu *et al.* [30] follow a similar approach, they include an uncompressed base layer and use fuzzy logic based quality selection. This type of approach using the leaf nodes of the octree as an enhancement layer is currently not suitable for real-time systems as it adds an extra surface orientation estimation step that introduces additional delays in the pipeline. In the previous chapter, we build on the ideas presented in PCC-DASH [14] to tile point cloud content using low complexity surface estimation suitable for frame rate sensitive real-time applications. In this chapter, we build on this approach and create tiles based on surfaces visible to multiple depth sensors and estimate their orientation using the transformation matrix associated with each sensor.

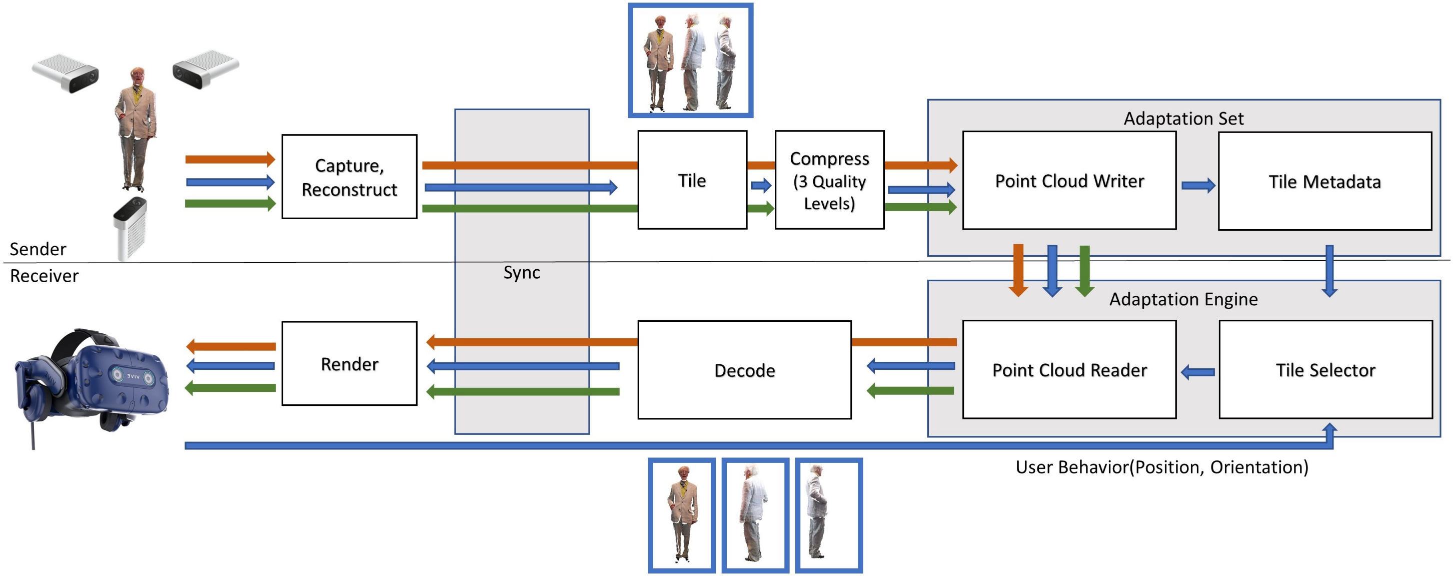


Figure 4.1: Architecture and Data Flow for Baseline (Orange), NA (Green) and TA (Blue) Streaming

4.2.3. VR COMMUNICATION EVALUATION

Quality assessment for remote communication is usually conducted using subjective user studies that are either passive or active. Passive tests involve asking users to rate pre-recorded clips of content. This approach to evaluation is more suited for standardized testing of codecs with offline content and has limited ecological validity in remote communication [31, 32]. Active tests involve multiple remote participants being placed in an interactive live communication system. The International Telecommunication Union published recommendations to define evaluation methods for quantifying the impact of terminal and communication link performance on point-to-point audiovisual communication [33]. The recommendation contains sample tasks such as name guessing, story comparison, picture comparison, object description and building blocks. Schmitt *et al.* [32] utilize the building blocks task to develop and evaluate personalized quality of experiment metrics for multiparty video conferencing at varying bitrates. Smith *et al.* [34] compare face-to-face communication with embodied and unembodied remote VR communication. They propose a task involving negotiating apartment layout and furniture placement based on blueprints. Li *et al.* [35] compare face-to-face, videoconferencing and Facebook spaces VR communication in the context of a photosharing task. They found that Facebook spaces is able to closely approximate face-to-face photosharing. In general, these methods have been used to compare VR remote communication with other technologies and with face to face communication. The tasks proposed either focus on the audio quality or rely on external objects for the evaluation. In this work, we focus on evaluating adaptive streaming within VR remote communication. We define a new visually focused training task where participants are taught a neck exercise from a trainer and are asked to perform the exercise in order to complete the task.

In order to evaluate communication, several questionnaires have been proposed in the literature. Toet *et al.* [36] propose the holistic mediated social communication (H-MSC) framework and associated questionnaire to evaluate the experience of spatial presence as well as social presence. The framework is general enough to be used for any mediated social communication system. Slater *et al.* [37] and Witmer *et al.* [38] have proposed two popular questionnaires aimed at measuring presence in virtual environments. Kangas *et al.* [39] present a pragmatic task related questionnaire that they use to evaluate VR interaction techniques in a rigid object manipulation task. Li *et al.* [35] propose a social VR questionnaire that evaluates Quality of Interaction/Communication *(QoI)*, Presence/Immersion and Social Meaning. In this chapter, we combine the QoI part of this questionnaire along with ITU visual quality questions [33], and task related questions from the pragmatic task related questionnaire [39].

4.3. STUDY DESIGN

4.3.1. EXPERIMENT TASK

The goal was to define a task that could be used to evaluate QoI, visual quality and task experience in VR remote communication. We needed to facilitate a conversation with a fixed general outline and with a focus on the visual aspect of communication. We considered the tasks presented in ITU-T P.920 [33], the building blocks task presented in [32] and the photosharing task presented in [35]. We found that these tasks are either

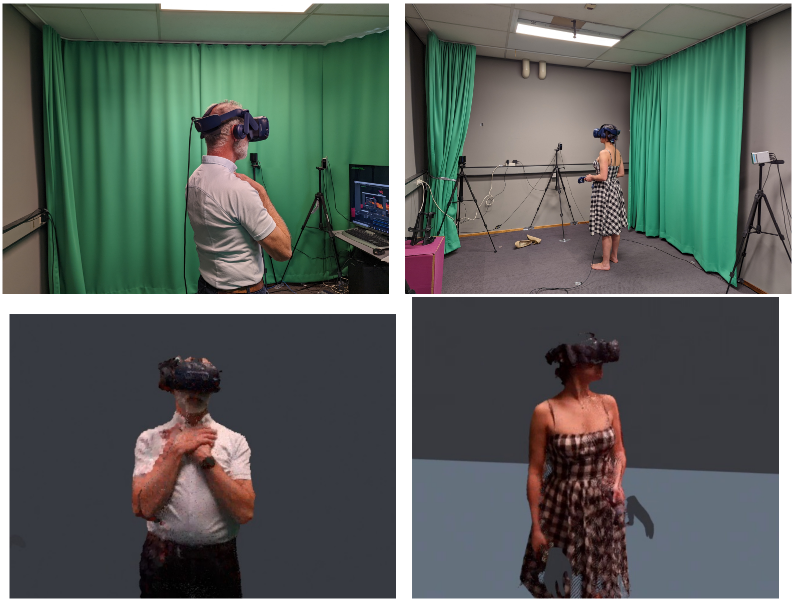


Figure 4.2: Capture Nodes and Real-Time Point Cloud Reconstructions

heavily dependent on the audio quality (story comparison, name-guessing and object description task) or required external objects that had to either be live-captured along with the user or digitally represented in the virtual world with appropriate interaction tools (picture comparison, building blocks, photo-sharing). These objects could occlude and distract from the visual quality of the user reconstruction. In order to assess the impact of tiled adaptive point cloud streaming, we chose to keep the audio quality consistent across sessions and manipulated the quality of the point cloud representation of the participants. Based on pre-trials with seven colleagues, we selected a training task where participants were asked to first learn and then perform a neck exercise. This allowed for a fixed general outline of the conversation and focused on the visual representation of the remote participant. The neck exercises were found not to induce severe motion sickness as validated in section 4.4 and provided coherent results. In our experiment design, we use a confederate user as the trainer. In this approach, all participants (trainees) in the same condition are always paired with the same trainer. This allows us to focus on the individual as the unit of study and we mitigate social context as a confounding variable. In addition, we isolate the basic elements of communication by attempting to hold constant the behavior of one conversation partner. In order to adhere to the recommendations on using confederate users [40], we used an asymmetric communication channel. The trainers are always shown the same quality point clouds of the trainee (octree depth 9) in order to ensure that the confederate users are as naive as possible to the current experimental condition. In addition, the trainers were not briefed on the hypothesis associated with each experimental condition. Confederate behavior could be scripted as these users were always the initiators and addressers as they provided instructions on how to perform the neck exercise. We then evaluate the quality ratings only from the point of view of the trainees in line with the recommendations in ITU-T P.1301 [31].

4.3.2. Adaptive VR Remote Communication System

System

The overall architecture and dataflow of the real-time VR remote communication system are shown in Figure 4.1 with baseline uncompressed, tiled user-adaptive (TA) and traditional network-adaptive (NA) streaming. Apart from the point cloud delivery pipeline described here the actual implementation also contains an audio delivery pipeline and a module for session management. The point cloud capture, and codec modules are implemented in C++, the other modules in C# with overall control in the Unity game engine.

The capture module interfaces with three Azure Kinect Depth sensors as shown in Figure 4.1. The sensors are calibrated in advance to generate the transformation matrix with intrinsics to combine color and depth as well as extrinsics to bring all sensors to a common coordinate system. The color and depth images from the sensors are then transformed and fused in order to reconstruct the point cloud. We set a target capture framerate of 15 fps based on pre-tests with colleagues as this was the maximum achievable framerate at an acceptable baseline quality. The point cloud generated is first sent directly to the renderer in order to generate a self view to embody the participant. The layout of the capture node is shown in Figure 4.2.

In baseline uncompressed, the point cloud is serialized and sent directly to the writer to forward to the receiver. For NA, the point cloud is sent to the encoder where it is compressed to three quality levels and sent to the writer to prepare the adaptation set with the associated encoded size.

For TA, the point cloud is split into tiles based on the contributing sensor. Each tile contains an orientation vector that is derived from the transformation matrix of the sensor and the centroid of it's bounding box. The tiles are then fed to the compression module that launches encoders in parallel for each tile and quality level. In this way we create an adaptation set with multiple representations for each tile. In addition, we prepare a tile meta data structure that contains information on the number of tiles, meta data for each tile, available quality levels and the associated encoded size.

At the receiver, for baseline uncompressed, the point cloud is sent directly to the renderer. For NA, the adaptation engine selects the highest possible quality within the available bandwidth budget. This is then decoded and sent to the renderer. We apply the bitrate budget per frame based on the target capture framerate of 15 frames per second (fps).

For TA, the adaptation engine utilizes both the tile metadata from the sender and the receiver's interactions in the system in terms of viewport position and orientation to select an appropriate representation for each tile within the available bandwidth budget. The tiles are then decoded in parallel and sent to the synchronizer. The synchronizer module was developed to playback tiled point cloud sequences with tiles of varying sizes and quality in a consistent manner. The primary goal of the synchronizer is to playback tiles of the same frame together with a secondary goal of playing back frames at the right time to match the received frame rate. The point clouds are then sent to the renderer.

Finally, the renderer stores the point locations and colors on a vertex buffer and

Table 4.1: System Setup

Hardware	HMD	HTC Vive Pro Eye		
	CPU	Intel(R) Core(TM) i7-7700K @ 4.20GHz (8 CPUs)		
	GPU	NVIDIA GeForce GTX 1080 Ti		
	Memory	32768MB RAM		
	Depth Sensors	3 x Azure Kinect DK		
Display Parameters	Resolution	1440 x 1600 pixels per eye		
	Application Target Framerate	90 Hz		
	Point Cloud Target Framerate	15 Hz		
Fixed Parameters	Audio Codec	Ogg Speex 48 KHz		
	Point Cloud Codec Configs	Octree Depth 6	Octree Depth 7	Octree Depth 9
	Kinect Depth Config	NFOV unbinned 640x576		
	Kinect Color Config	1280x720		
	Point Cloud Capture	ca. 130k points per frame at 15 fps		
Conditions	Target Bitrates	7 Mbps	14 Mbps	
	Streaming Conditions	Network Adaptive	Tiled Adaptive	Uncompressed

draws procedural geometry on the GPU. Points are rendered as camera-facing quads with a fixed offset based on the selected quality level.

Tiling and Tile Selection

We use a modified version of tiling approach presented in Chapter 3. We adapted the approach to three tiles instead of four due to the real-time framerate constraints of our capturing system. We were able to achieve a stable capture framerate of 15 fps with three sensors. We use the forward vector of the contributing sensor to estimate the orientation of the tile surface. We also compute the centroid of the bounding box of the tile rather than the centroid of points in a tile as described in Chapter 3. Each tile T_i has an orientation $\vec{T}_i$ and a bounding box centroid $T_i^{(bc)}$. The current viewport V of the user is defined by a position $V^{(pos)}$ and an orientation $\vec{V}$. The utility of each tile is calculated based on the following formula:

$$u(V,T_i) = \begin{cases} \left|\vec{T}_i \cdot \vec{V}\right|, & \text{if } d(V^{(pos)}, T_i^{(bc)}) < d_{max}(V^{(pos)}, T_j^{(bc)}) \\ -\left|\vec{T}_i \cdot \vec{V}\right|, & \text{otherwise.} \end{cases} \tag{4.1}$$

We use the absolute value of the dot product to identify surfaces directly facing the user. In addition, to account for the position of the user we ensure that the two tiles closest to the user always have a positive utility.

In the next step, the calculated utility is used to allocate the bandwidth budget to each tile. In this work, we re-use the allocation strategies presented in previous research [14, 15]. Based on pre-tests with colleagues, we found that uniform bit rate allocation achieved a higher median score and a lower spread of scores. We use the utility to rank the available tiles. The quality of each tile is then increased one step at a time in order of this ranking.

4.3.3. Experiment Conditions

The primary components determining the point cloud quality are the capture sensors, the codec configurations or adaptation set, target bitrate, streaming condition and the display device used. In order to evaluate adaptive streaming, we vary the streaming condition and target bitrate. The other factors do not change dynamically over a session and are uninteresting for streaming optimization. They are held constant across all participants. In addition, the audio quality is also maintained at the same level across all experimental sessions using the Ogg container format and the Speex codec with an ultra wide band sampling rate. In order to conduct a pre-test and set the experimental conditions, we used the dataset published by Reimat *et al.* [41] sub-sampled to three cameras (1,5,6) as this most closely resembles our capture setup. Based on pre-tests on encode time and captured point count we use three codec configurations on the MPEG anchor codec shown in Table 4.1. All codec configurations were encoded at JPEG QP 75. We also set two target bitrates exclusively for remote user point cloud reconstructions at 7Mbps and 14Mbps based on the approximate bandwidth requirement for full point clouds using the two highest quality levels from the pre-tests. We selected three neck exercises that were taught to all participants. In the experiment, we evaluate three streaming conditions: baseline uncompressed, NA and TA.

4.3.4. Experiment Design and Protocol

With three neck exercises and three streaming conditions, we used a Greco-Latin square design to randomize and counter balance the different levels of each variable. In order to avoid fatigue, we separated the target bitrates so each participant only took part in three sessions at a fixed target bitrate. We recruited two confederate users to play the role of trainer for each target bitrate. We recruited 16 and 17 participants for the two target bitrates of 7Mbps and 14Mbps.

Upon arrival, participants were led to the experiment room and were briefed about the purpose of the study, after which they were asked for written consent for data gathering. Participants were asked to provide some background information on themselves and to take a Ishihara test [42] for color perception. Participants were then asked to fill the simulator sickness questionnaire before starting and again after each experiment session. We then conducted a brief training session where participants were shown the highest and lowest available quality of the remote user point cloud to serve as an anchor. Participants were then taught how to use the HMD controllers to teleport and navigate the virtual space. Participants were informed that during the experiment they are free to move about the virtual space. Participants then entered the first session, in each session there was a brief introduction by the trainer, a training stage where the trainer demonstrated the exercise technique and finally a performance where participants were asked to repeat the exercise three times in order to complete the session. Each session took 2 to 3 minutes to complete. Participants were asked to fill in a questionnaire to report their experience after each session. Participants completed the experiment in ca. 30 minutes.

4.3.5. Experiment Setup

Participants took part in the study in a separate room from the trainer. Each room had a workstation, an HTC Vive Pro Eye HMD (with controllers and base stations) and three azure kinect depth sensors. The configuration of the setup used by participants is shown in Table 4.1. Figure 4.2 shows the setup used to capture each user and the resulting point cloud. The two labs were connected using a dedicated gigabit ethernet connection in order to control the connection quality for the duration of the study. Each of the azure kinect cameras was set to use a depth mode of NFOV unbinned with a resolution of 640x576 and the lowest supported color resolution of 1280x720. This was done as the color image is later mapped to the depth image in order to reconstruct the point cloud similar to the method proposed by Reimat *et al.* [41].

During each experiment session, the system resource consumption was logged at the trainee node using the Resources Consumption Metrics (RCM) measurement tool [43]. This tool is a windows application that allows for the capture of CPU, GPU and memory usage for a given process in a 1-second interval. The conversation audio was recorded with audio-only capture using the Open Broadcaster Software tool. The VR remote communication application recorded log files that contained performance information on each component including latency and framerate.

4.3.6. Data Collection

After each condition participants filled in a questionnaire about the experience they just had. The first six questions were related to QoI [35] with a 5-point Likert scale to address. The next four questions were about the visual quality of the point cloud representation [33] with a 5-point ACR scale to address. The last five questions were about task related experience [39] with a 7-point Likert scale to address. These questionnaires together will address research question **3.1**. In addition, on the trainee node we record system resource consumption and log playback performance to address **3.2**.

4.4. Results

4.4.1. Performance Results

We evaluate system performance based on resource consumption, framerate and latency. The machine used for the evaluation is described in Table 4.1 and the results are shown in Table 4.2. As expected, we see higher memory consumption for large uncompressed point clouds. At 7Mbps we observe ca. 10% reduction in CPU usage on account of additional parallelization in encoding and decoding in TA as compared to NA with similar GPU and memory utilization. At 14Mbps we observe similar results with a 9% reduction in CPU usage.

In addition, our VR application logs the capture to render latency and framerate while accounting for clock sync between the trainee and trainer nodes. The latency results are shown in Figure 4.3. The selfview latency describes the time needed for reconstructing and rendering only. The selfview is rendered with a median latency of 75ms across all experiment sessions. We observe the largest range of latency for baseline un-

Table 4.2: System Resource Consumption (1000+ samples each, mean values)

	Streaming Condition	Target Bitrate	CPU (%)	GPU(%)	Memory (MB)
Uncompressed	Baseline	-	31.1% (SD 7.8)	46.2% (SD 6.8)	3146 (SD 1423)
Compressed	Network Adaptive	7Mbps	49.7% (SD 5.9)	45.2% (SD 6.9)	1025 (SD 112)
	Tiled Adaptive	7Mbps	39.2% (SD 10.7)	46.9% (SD 5.4)	1088 (SD 94)
	Network Adaptive	14Mbps	57.4% (SD 4.6)	45.7% (SD 4.8)	1071 (SD 68)
	Tiled Adaptive	14Mbps	48.4% (SD 6.4)	46.5% (SD 4.7)	1096 (SD 89)

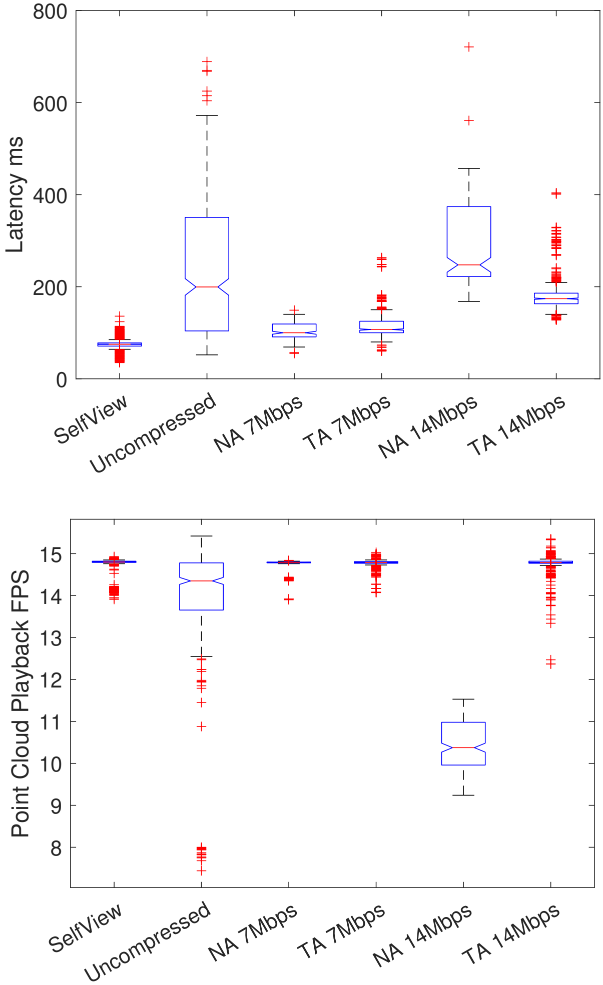


Figure 4.3: System Performance in terms of latency (ms) and framerate (fps)

compressed streaming with the largest point clouds requiring ca. 300Mbps to transmit and render. At 7Mbps we observe similar latency across the two streaming conditions. However, at 14Mbps we observe a 74ms increase in median latency for NA. This is caused by larger encode and decode times required for the highest quality point clouds in our adaptation set. In case of TA we have some performance gains due to parallel encoding and decoding of tiles and generally smaller point clouds decoded at the receiver.

The application runs at a near consistent 90 fps with motion reprojection. The point cloud target capture framerate is set to 15 fps based on the capability of the system. On the receiver end, we observe a drop in median framerate to 10.4 fps for NA at 14Mbps caused by the encode and decode times required for the highest quality full point clouds in our adaptation set. For the remaining streaming conditions we generally observe similar performance of ca. 15 fps point cloud playback with a larger range for uncompressed point clouds as shown in Figure 4.3. To summarize, we observe significant gains in playback performance (framerate and latency) by employing TA with lower CPU usage as compared to NA.

4.4.2. Subjective Results

Quality of Communication

In this section of the questionnaire we included the QoI questions from [35]. These questions are meant to assess four types of experience: (1) feeling understood (2) engaging in conversations (3) sensing other's emotion and (4) feeling comfortable in the environment. The overall QoI scores are obtained by adding up the scores for each of the six questions. We split the analysis for each target bitrate of 7Mbps and 14Mbps.

Table 4.3: Pairwise post-hoc test QoI and streaming condition at 7Mbps

Streaming Condition	Z	p	r
NA – TA	-2.2340	0.0025	0.3950
NA – Baseline	-2.6410	0.0083	0.4670
TA – Baseline	-1.7230	0.0849	0.3050

For the 7Mbps case a Shapiro-Wilk normality test issued on the entirety of the scores indicates that they do not follow a normal distribution ($W = 0.9163$, $p = 0.003$). We use non-parametric statistical tools to perform an exploratory data analysis and check if statistical differences could be found amongst the different streaming conditions. To compare the QoI across the different streaming conditions, we first conduct a Friedman's test to check if any groups exist with statistically significant differences ($\chi^2 = 12.04$, $p = 0.0024$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 4.3

We observe statistically significant differences in two of the comparisons. TA outperforms NA with a medium effect size ($r = 0.39505$). As expected, baseline uncompressed streaming outperforms NA with a large effect size ($r = 0.467$). On the other hand, we do not observe statistically significant differences between TA and baseline uncompressed indicating similar performance in terms of QoI.

For the 14Mbps case, a Shapiro-Wilk normality test indicates that the scores are not

Table 4.4: Pairwise post-hoc test QoI and streaming condition at 14Mbps

Streaming Condition	Z	p	r
NA – TA	-3.3720	<.001	0.3780
NA – Baseline	-3.3310	<.001	0.5710
TA – Baseline	2.1650	0.0304	0.3710

normally distributed ($W = 0.9484$, $p = 0.002$). In order to check if any of the groups exhibit statistically significant differences we run Friedman's test ($\chi^2 = 24.96$, $p < 0.001$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 4.4.

This time we observe statistically significant differences in all comparisons. TA outperforms NA with a medium effect size ($r = 0.3780$). Baseline uncompressed streaming outperforms NA with a large effect size ($r = 0.5710$) and outperforms TA with a medium effect size ($r = 0.3710$). In general, we observe that TA leads to statistically significant gains in terms of QoI with respect to NA across both bitrates.

In order to validate the three exercises we used, we checked if they led to different QoI scores and we found no statistically significant differences using the Friedman test ($\chi^2 = 3.16$, $p = 0.206$) at 7Mbps and ($\chi^2 = 2.28$, $p = 0.3198$) at 14Mbps.

VISUAL QUALITY

In order to assess the visual quality we include a question about the visual quality of the trainer's point cloud representation. Participants were asked to indicate the quality on a scale from 1 to 5 (*1-Bad*, *2-Poor*, *3-Fair*, *4-Good*, and *5-Excellent*). We analyzed these scores separately for each target bitrate.

Table 4.5: Pairwise post-hoc test visual quality and streaming condition at 7Mbps

Streaming Condition	Z	p	r
NA – TA	-2.5840	0.0098	0.4570
NA – Baseline	-2.5550	0.0106	0.4520
TA – Baseline	-1.2650	0.2059	0.2240

For the 7Mbps case, a Shapiro-Wilk normality test issued on the scores indicates that they do not follow a normal distribution ($W = 0.8106$, $p < 0.001$). To compare the remote user visual quality across the different streaming conditions, we first conduct a Friedman's test to check if any groups exist with statistically significant differences ($\chi^2 = 9.8$, $p = 0.0074$). We then conduct a Wilcoxon signed-rank test with Bonferroni correction. The results are shown in table 4.5. We observe statistically significant differences in two pairwise comparisons. TA outperforms NA with a medium effect size ($r = 0.4570$). Baseline uncompressed streaming outperforms NA with a medium effect size ($r = 0.4520$). We do not observe statistically significant differences between TA and baseline indicating similar performance in terms of visual quality.

For the 14Mbps case, a Shapiro-Wilk normality test indicates the scores do not follow a normal distribution ($W = 0.8622$, $p < 0.001$). Friedman's test reveals that there are

Table 4.6: Pairwise post-hoc test visual quality and streaming condition at 14Mbps

Streaming Condition	Z	p	r
NA – TA	-2.8140	0.0049	0.4830
NA – Baseline	-3.1400	0.0017	0.5390
TA – Baseline	-2.1210	0.0339	0.3640

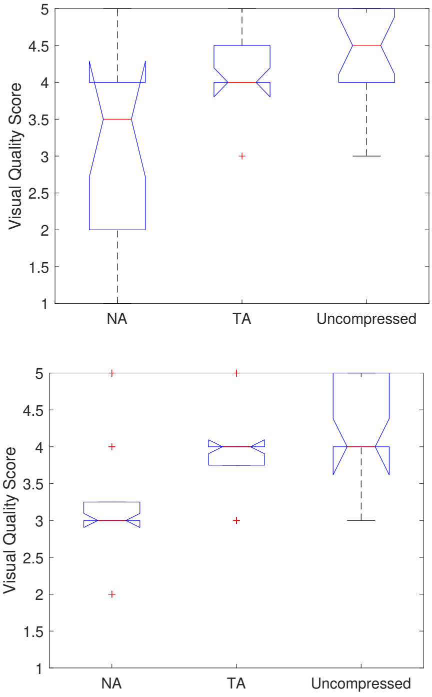


Figure 4.4: Point cloud quality at 7Mbps and 14Mbps

groups with statistically significant differences ($\chi^2 = 19$, $p < 0.001$). The results of the Wilcoxon signed-rank test with bonferroni corrections are shown in Table 4.6. This time we observe statistically significant differences in two of the pairwise comparisons. TA outperforms NA with a medium effect size ($r = 0.4830$). Baseline uncompressed outperforms NA with a large effect size ($r = 0.5390$). We do not observe statistically significant differences between TA and baseline uncompressed streaming indicating similar performance in terms of visual quality at 14Mbps. The distribution of the scores is shown in Figure 4.4. We observe significant gains in perceived visual quality by employing TA over NA. At 14Mbps we even observe the same median score as baseline uncompressed streaming that required ca. 300Mbps.

Task Related Experience

For the task-related experience questions, we used the questionnaire presented in [39]. The questions are meant to assess the participant's confidence in using the system and if the system was natural and easy to use while performing the task. We compute an overall score by adding up the five questions from this section. We split the analyses by target bit rate. A Shapiro Wilk normality test issued on the scores revealed that they are not normally distributed; ($W = 0.9019$, $p = 0.0013$) for 7Mbps and ($W = 0.9653$, $p = 0.014$) for 14Mbps. We then checked if there are any groups with statistically significant differences using Friedman's test ($\chi^2 = 2.1$, $p = 0.3504$) at 7Mbps and ($\chi^2 = 2.56$, $p = 0.278$) at 14Mbps. We found no statistically significant differences amongst the different streaming conditions at both target bitrates. This indicates participants were able to adapt their behavior to compensate for changes in the point cloud quality and were able to complete the task within the same time regardless. We observe no gains in task experience by employing TA. This can be explained by the relative simplicity of completing our training task as compared to tasks used in other works [32, 33, 35].

SIMULATOR SICKNESS QUESTIONNAIRE

We calculate the total severity of cybersickness based on the SSQ questionnaire for all 33 participants across the three experiment sessions. We observe low post-exposure total severity scores for cybersickness, as defined in [44, 45] ([mean, median, std] baseline: [10.54, 3.74, 18.22], after session 1: [9.63, 3.74, 14.88], after session 2: [9.41, 3.74, 16.14], after session 3:[11.22, 3.74, 16.95]). In general, no users reported severe symptoms after participating in any of the experiment sessions.

4.5. DISCUSSION

In general, tiled adaptive streaming techniques have received significant research attention for omnidirectional videos [46–48] and point clouds [11, 18, 49–51]. However, further studies are required for live real-time human point cloud reconstructions. Although the tiles are independently decodable, their quality cannot be optimized in isolation based on available bandwidth and viewport. Some participants reported that seeing artifacts in body extremities and in the face of the reconstruction was unpleasant. Further studies about body part segmentation and quality perception are required to optimize tiling and tile selection strategies for humans engaging in conversation as compared to prerecorded content, objects and scenes.

In our pre-trials, we observed a higher median score and a lower spread of scores with uniform tile allocation. This is different from the results with a prerecorded dataset reported in the previous chapter, where hybrid tile allocation was shown to yield higher perceived quality. The point clouds used in that study were captured offline. They were dense and voxelized with ca. 1 million points per frame and the adaptation set comprised of 30 quality levels. In this work, we used real-time live captured point clouds with ca. 130K points with an adaptation set of 3 quality levels defined by octree depths.

In this work, participants were trained on how to navigate the virtual space with a controller-based teleport. During the session, we did not force participants to move in order to keep the interaction more natural. Future studies on VR remote communication should account for this trade-off. Movement within the scene is important to evaluate the visual quality of adaptation from different view angles but scripting or forcing these movements tends to break the flow of the interaction making it difficult to evaluate the quality of communication.

4.6. Conclusion

In this chapter, we presented a VR remote communication system with tiled adaptive real-time point cloud streaming using commodity hardware. We present an evaluation framework and a training task to evaluate the impact of adaptive streaming on QoI, visual quality and task-related experience to address research question **R3** *(How can we optimize the delivery of streams of dynamic point clouds in VR remote communication?)*. To address research question *R3.1 (How does tiled user-adaptive point cloud streaming impact the perceived quality of remote user reconstruction?)*, we demonstrate that our system at 14Mbps was able to achieve similar visual quality as compared to uncompressed streaming at ca. 300Mbps. We also demonstrate statistically significant improvements to QoI as compared to traditional network adaptive streaming. To address research question *R3.2 (What is the computational overhead of using tiled adaptive point cloud streaming?)*, we demonstrate improvements to playback performance with up to 10% reduction in CPU usage. We were able to validate the adaptive streaming approach proposed in the previous chapter in a live communication scenario with a novel evaluation methodology.

There is a need for new standardized tasks that can be used to evaluate VR remote communication. The existing ITU recommendations are insufficient to handle novel interaction techniques and immersive content inherent to VR communication. In this work, we utilized a neck exercise training task as it was more visually focused and the interaction was repeatable with a confederate trainer. Further study into other use cases and scenarios are required to evaluate emerging VR communication systems. To this end, ITU-T has recently launched a new activity [52] that cites our work as one potential approach to develop assessment methods for extended reality meetings. In the next chapter, we conclude the thesis by revisiting the research questions and summarizing the various discussion items. In addition, we also present some applications created using the delivery pipeline presented in this chapter along with other publicly available resources created during the course of this thesis.

REFERENCES

[1] Z. Yang, W. Wu, K. Nahrstedt, G. Kurillo, and R. Bajcsy, *Enabling multi-party 3d tele-immersive environments with <i>viewcast</i>,* ACM Trans. Multimedia Comput. Commun. Appl. **6** (2010), 10.1145/1671962.1671963.

[2] H. Fuchs, A. State, and J. Bazin, *Immersive 3d telepresence,* Computer **47**, 46 (2014).

[3] 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 (2016).

[4] J. Jansen, S. Subramanyam, R. Bouqueau, G. Cernigliaro, M. Martos Cabré, F. Pérez, and P. Cesar, *A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH,* in *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20 (ACM, New York, NY, USA, 2020).

[5] S. N. B. Gunkel, R. Hindriks, K. M. E. Assal, H. M. Stokking, S. Dijkstra-Soudarissanane, F. t. Haar, and O. Niamut, *Vrcomm: An end-to-end web system for real-time photorealistic social vr communication,* in *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 65–79.

[6] A. D. Wilson and H. Benko, *Projected augmented reality with the roomalive toolkit,* in *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, ISS '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 517–520.

[7] J. Lawrence, D. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong, *Project starline: A high-fidelity telepresence system,* ACM Trans. Graph. **40** (2021), 10.1145/3478513.3480490.

[8] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi, *Holoportation: Virtual 3d teleportation in real-time,* in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 741–754.

[9] R. Mekuria, P. Cesar, I. Doumanis, and A. Frisiello, *Objective and subjective quality assessment of geometry compression of reconstructed 3D humans in a 3D virtual room,* Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, 95991M (2015).

[10] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch, *The effect of avatar realism in immersive social virtual realities,* in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17 (Association for Computing Machinery, New York, NY, USA, 2017).

[11] J. v. d. Hooft, M. T. Vega, T. Wauters, C. Timmerer, A. C. Begen, F. D. Turck, and R. Schatz, *From capturing to rendering: Volumetric media delivery with six degrees of freedom,* IEEE Communications Magazine **58**, 49 (2020).

[12] M. Hosseini and C. Timmerer, *Dynamic Adaptive Point Cloud Streaming,* in *Proceedings of the 23rd Packet Video Workshop*, PV '18 (ACM, New York, NY, USA, 2018) pp. 25–30.

[13] J. Park, P. A. Chou, and J. Hwang, *Rate-Utility Optimized Streaming of Volumetric Media for Augmented Reality,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 149 (2019).

[14] J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, and H. Hellwagner, *Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression,* in *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19 (Association for Computing Machinery, New York, NY, USA, 2019) pp. 2405–2413.

[15] S. Subramanyam, I. Viola, A. Hanjalic, and P. Cesar, *User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling,* in *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 3669–3677.

[16] S. Gül, D. Podborski, A. Hilsmann, W. Morgenstern, P. Eisert, O. Schreer, T. Buchholz, T. Schierl, and C. Hellge, *Interactive volumetric video from the cloud,* (2020).

[17] J. Son, S. Gül, G. S. Bhullar, G. Hege, W. Morgenstern, A. Hilsmann, T. Ebner, S. Bliedung, P. Eisert, T. Schierl, T. Buchholz, and C. Hellge, *Split rendering for mixed reality: Interactive volumetric video in action,* in *SIGGRAPH Asia 2020 XR*, SA '20 (Association for Computing Machinery, New York, NY, USA, 2020).

[18] L. Wang, C. Li, W. Dai, S. Li, J. Zou, and H. Xiong, *Qoe-driven adaptive streaming for point clouds,* IEEE Transactions on Multimedia , 1 (2022).

[19] I. Reimat, Y. Mei, E. Alexiou, J. Jansen, J. Li, S. Subramanyam, I. Viola, J. Oomen, and P. Cesar, *Mediascape xr: A cultural heritage experience in social vr,* in *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22 (Association for Computing Machinery, New York, NY, USA, 2022) p. 6955–6957.

[20] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, *Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units,* in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14 (Association for Computing Machinery, New York, NY, USA, 2014) p. 637–644.

[21] G. Cernigliaro, M. Martos, M. Montagud, A. Ansari, and S. Fernandez, *PC-MCU: Point Cloud Multipoint Control Unit for Multi-User Holoconferencing Systems,* in *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* (Association for Computing Machinery, New York, NY, USA, 2020) p. 47–53.

[22] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, *et al.*, *Emerging MPEG Standards for Point Cloud Compression,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 133 (2019).

[23] E. Pavez, P. A. Chou, R. L. de Queiroz, and A. Ortega, *Dynamic polygon cloud compression,* Microsoft Research Technical Report (2016).

[24] R. L. de Queiroz and P. A. Chou, *Compression of 3d point clouds using a region-adaptive hierarchical transform,* IEEE Transactions on Image Processing **25**, 3947 (2016).

[25] M. Hosseini, *Adaptive rate allocation for view-aware point-cloud streaming,* (2017), 10.13140/RG.2.2.23436.26244.

[26] L. He, W. Zhu, K. Zhang, and Y. Xu, *View-dependent streaming of dynamic point cloud over hybrid networks,* in *Advances in Multimedia Information Processing – PCM 2018* (Springer International Publishing, Cham, 2018) pp. 50–58.

[27] J. Li, C. Zhang, Z. Liu, W. Sun, and Q. Li, *Joint communication and computational resource allocation for qoe-driven point cloud video streaming,* (2020).

[28] K. Lee, J. Yi, Y. Lee, S. Choi, and Y. M. Kim, *Groot: A real-time streaming system of high-fidelity volumetric videos,* in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (Association for Computing Machinery, New York, NY, USA, 2020).

[29] B. Han, Y. Liu, and F. Qian, *Vivo: Visibility-aware mobile volumetric video streaming,* in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (Association for Computing Machinery, New York, NY, USA, 2020).

[30] Z. Liu, J. Li, X. Chen, C. Wu, S. Ishihara, Y. Ji, and J. Li, *Fuzzy logic-based adaptive point cloud video streaming,* IEEE Open Journal of the Computer Society **1**, 121 (2020).

[31] ITU-T P.1301, *P.1301 : Subjective quality evaluation of audio and audiovisual multiparty telemeetings,* International Telecommunication Union (2017).

[32] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar, *Towards individual qoe for multiparty videoconferencing,* IEEE Transactions on Multimedia **20**, 1781 (2018).

[33] ITU-T P.920, *P.920 : Interactive test methods for audiovisual communications,* International Telecommunication Union (2000).

[34] H. J. Smith and M. Neff, *Communication behavior in embodied virtual reality,* in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18 (Association for Computing Machinery, New York, NY, USA, 2018) p. 1–12.

[35] J. Li, Y. Kong, T. Röggla, F. De Simone, S. Ananthanarayan, H. de Ridder, A. El Ali, and P. Cesar, *Measuring and understanding photo sharing experiences in social virtual reality,* in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 1–14.

[36] A. Toet, T. Mioch, S. N. Gunkel, O. Niamut, and J. B. van Erp, *Holistic framework for quality assessment of mediated social communication,* (2021).

[37] M. Slater, M. Usoh, and A. Steed, *Depth of presence in virtual environments,* Presence: Teleoper. Virtual Environ. **3**, 130–144 (1994).

[38] B. G. Witmer and M. J. Singer, *Measuring presence in virtual environments: A presence questionnaire,* Presence: Teleoper. Virtual Environ. **7**, 225–240 (1998).

[39] J. Kangas, S. K. Kumar, H. Mehtonen, J. Järnstedt, and R. Raisamo, *Trade-off between task accuracy, task completion time and naturalness for direct object manipulation in virtual reality,* Multimodal Technologies and Interaction **6** (2022), 10.3390/mti6010006.

[40] A. Kuhlen and B. SE, *Language in dialogue: when confederates might be hazardous to your data,* Psychon Bull Rev. **20**, 54 (2013).

[41] 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*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 300–306.

[42] J. H. Clark, *The ishihara test for color blindness.* in *American Journal of Physiological Optics* (1924) p. 269–276.

[43] M. Montagud, J. A. De Rus, R. Fayos-Jordan, M. Garcia-Pineda, and J. Segura-Garcia, *Open-source software tools for measuring resources consumption and dash metrics,* in *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 261–266.

[44] 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**, 203 (1993).

[45] P. Bimberg, T. Weissker, and A. Kulik, *On the Usage of the Simulator Sickness Questionnaire for Virtual Reality Research,* in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020) pp. 464–467.

[46] C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, *A survey on 360° video streaming: Acquisition, transmission, and display,* ACM Comput. Surv. **52** (2019), 10.1145/3329119.

[47] M. Zink, R. Sitaraman, and K. Nahrstedt, *Scalable 360° video stream delivery: Challenges, solutions, and opportunities,* Proceedings of the IEEE **107**, 639 (2019).

[48] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, *An http/2-based adaptive streaming framework for 360° virtual reality videos,* in *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17 (Association for Computing Machinery, New York, NY, USA, 2017) p. 306–314.

[49] S. Petrangeli, G. Simon, H. Wang, and V. Swaminathan, *Dynamic adaptive streaming for augmented reality applications,* in *2019 IEEE International Symposium on Multimedia (ISM)* (2019) pp. 56–567.

[50] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li, and Y. Ji, *Point cloud video streaming: Challenges and solutions,* IEEE Network **35**, 202 (2021).

[51] F. Qian, B. Han, J. Pair, and V. Gopalakrishnan, *Toward practical volumetric video streaming on commodity smartphones,* in *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, HotMobile '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 135–140.

[52] ITU-T P.QXM, *QoE assessment of eXtended Reality (XR) meetings,* International Telecommunication Union (2022).

5

Conclusion

Throughout the main body of the thesis, we conducted a series of studies exploring user adaptive streaming of dynamic point clouds in VR remote communication. We started by defining and employing a subjective quality evaluation protocol for dynamic point clouds viewed in immersive environments in Chapter 2. We then proposed a real-time tiling and adaptive streaming approach for prerecorded point clouds in Chapter 3 and evaluated the quality gains offered by this approach. We then constructed a live VR remote communication application with live captured point cloud user reconstructions and tested our adaptive streaming approach in Chapter 4. In this final chapter, we first revisit the research questions formulated in the introductory chapter. After that, we discuss the lessons learned in the course of this thesis and present the resources created that can benefit future work in the field. At last, we reflect on the implications of the research presented in this thesis and highlight the limitations and areas for future work.

This chapter is based on:

1. *I. Reimat, E. Alexiou, J. Jansen, I. Viola, **S. Subramanyam**, and P. Cesar. 2021. CWIPC-SXR: Point Cloud dynamic human dataset for Social XR. In Proceedings of the 12th ACM Multimedia Systems Conference (MMSys '21). Association for Computing Machinery, New York, NY, USA, 300–306. https://doi.org/10.1145/3458305.3478452*
2. *A. Chatzitofis, L. Saroglou, P. Boutis, P. Drakoulis, N. Zioulis, **S. Subramanyam**, B. Kevelham, C. Charbonnier, P. Cesar, D. Zarpalas, S. Kollias, P. Daras, "HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media," in IEEE Access, vol. 8, pp. 176241-176262, 2020, doi: 10.1109/ACCESS.2020.3026276.*
3. *I. Reimat, Y. Mei, E. Alexiou, J. Jansen, J. Li, **S. Subramanyam**, I. Viola, J. Oomen, P. Cesar, Mediascape XR: A Cultural Heritage Experience in Social VR, Proceedings of the 30th ACM International Conference on Multimedia, October 10-14, 2022, Lisboa, Portugal*
4. *A. Revilla, S. Zamarvide, I. Lacosta, F. Perez, J. Lajara, B. Kevelham, V. Juillard, B. Rochat, M. Drocco, N. Devaud, O. Barbeau, C. Charbonnier, P. de Lange, J. Li, Y. Mei, K. Ławicka, J. Jansen, N. Reimat, **S. Subramanyam**, P. Cesar, "A Collaborative VR Murder Mystery using Photorealistic User Representations," 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2021, pp. 766-766, doi: 10.1109/VRW52623.2021.0266.*

5.1. Revisiting the research questions

The main challenge of this thesis is to explore if user-centered adaptive streaming of point cloud user reconstructions can improve the quality of interaction in VR remote communication. To answer this question, we first need to define a protocol to evaluate the perceived quality of dynamic point cloud reconstructions viewed in immersive environments.

In recent years, the standardization body MPEG launched an activity to create new compression standards for point clouds [1]. This included the definition of common test conditions (CTC) for upcoming point cloud technologies [2]. In chapter 2, we first describe the subjective evaluation methodology that we used to evaluate codecs as part of this activity. This approach was based on recording and evaluating 2D videos of decoded dynamic point cloud sequences. We then propose and employ a subjective quality evaluation protocol suitable for dynamic point cloud sequences viewed in realistic immersive environments. We use this protocol to evaluate the quality of decoded dynamic point clouds. We subjectively evaluated the state-of-the-art V-PCC standard codec [1] and the real-time MPEG anchor codec [3] using the CTC to answer research question 1:

R1: How can we measure the perceived quality of dynamic point cloud user reconstructions in immersive environments?

In chapter 2, we review the subjective evaluation methodology employed during the MPEG point cloud compression standardization activity [2]. The approach used is based on recording videos using fixed navigation paths around the point cloud object with a wide range of viewing angles and distances and then subjectively assessing the resulting uncompressed videos on a 2D screen. From our participation in this activity, we learned that the choice of navigation paths can have a drastic impact on the assessment and assessing point cloud content on a 2D screen is not realistic for target applications in VR. In addition, the commonly used double stimulus assessment method is challenging to implement with 3D objects while ensuring consistent viewing in terms of proximity to both stimuli, in immersive viewing environments this can also lead to cybersickness. In general, the double stimulus method can lead to the quality evaluation task getting reduced to difference recognition while insights on factors contributing to perceived quality can be missed.

We then set out to perform an evaluation in an immersive viewing environment with unrestricted user movements. Based on the lessons learned, we employed the absolute category rating with hidden references evaluation methodology based on ITU recommendations [4]. We implemented an immersive VR playback environment for dynamic point clouds and defined an evaluation protocol for this novel environment. We validated this approach by evaluating two-point cloud codecs under the MPEG common test conditions [2] with 52 participants and compared the resulting quality across a range of target bitrates using two datasets. The first was a popular dataset used by the various standards bodies active in point cloud compression [5] and the second is a custom dataset of dynamic point clouds sampled from the surface of a dynamic motion-captured animated mesh sequence. The two codecs under test were the state-of-the-art MPEG V-PCC codec [1] and the low latency MPEG anchor codec [3]. In general, we found that the V-PCC codec achieves better rate-distortion performance as compared to the MPEG anchor.

This research question was further divided into:

R1.1: What is the impact of immersive viewing environments on subjective quality assessment?

We investigated this research question by conducting the subjective evaluation in two immersive viewing conditions based on freedom of movement; 3DoF and 6DoF. In general, we found that the viewing condition had a small effect on the quality scores. Both conditions led to similar trends in terms of relative performance between the two codecs being tested. Experiment participants reported that the 3DoF condition offered a more stable assessment as the point of view and proximity was enforced for all stimuli. They also reported that the 6DoF condition felt more realistic while freely navigating the virtual space. Overall 52% of participants preferred the 6DoF viewing condition as it allowed them to move closer and inspect details in each stimulus. Only 21% of participants preferred the passive and relaxed viewing offered by 3DoF as compared to 6DoF. We concluded that a decision on what viewing condition to use should be made considering the trade-off between immersive, personalized experience, and fairness of comparison between solutions. In all subsequent quality evaluations presented in this thesis, we use a 6DoF viewing condition as it is more realistic for VR remote communication applications.

R1.2: What factors do users rely on while performing subjective quality assessment of dynamic point clouds?

To address this research question, we conducted a semi-structured interview with all experiment participants. We found that 56% of participants used the same three criteria to determine dynamic point cloud quality. The first criterion was based on the overall outline of the point cloud and patterns of distortions on the body and clothes of the reconstruction. The second was based on the perceived naturalness of movements and gestures in each dynamic sequence. The final criterion was visual artefacts such as blocks, blurred textures and extraneous floating points around the reconstruction. 46% of participants reported that accurate facial reconstructions and facial expressions were the most important factor for quality assessment, they felt this was an important cue for social connectedness.

This led us to conclude that there is an opportunity to optimize the delivery of these point cloud reconstructions by segmenting surfaces and prioritizing amongst them based on visibility from the user's viewport at any given instant. This leads us to research question 2:

R2: How can we optimally allocate the available bandwidth across independently decodable spatial segments?

In chapter 3, we proposed a low-complexity adaptive streaming approach based on tiling point cloud human reconstructions. We defined an auxiliary utility function to quantify the utility of each point cloud tile based on the location and orientation of the tile as well as the user's viewport. In order to allocate the available bandwidth across each tile based on utility we employed three bit-rate allocation approaches from the literature [6] along with two novel approaches. We then evaluated the quality of the resulting point cloud reconstruction with both objective and subjective quality evaluation. The objective evaluation was performed by recording screenshots and using image distortion metrics. The subjective evaluation was performed based on the protocol described in Chapter 2 with a user study involving 30 participants. From the results, we found that a utility-weighted approach to bit rate allocation was able to achieve the best rate-distortion performance across all sequences in the dataset.

This research question was further divided into:

R2.1: Does user-centered tiled adaptive streaming offer significant gains to reconstruction quality and bandwidth savings?

We investigated this research question by first evaluating the objective image quality of screenshots of tiled adaptive point cloud frames. We played back the dynamic point cloud sequences using the navigation patterns recorded during the user study presented in chapter 2. Screenshots were recorded for both non-adaptive and adaptive methods using bandwidth allocation streams from existing research [6]. The results from the objective evaluation indicate that adaptive streaming with the best-performing bandwidth allocation approach yields bitrate savings of upto 57% while delivering the same quality as a non-adaptive approach using the same codec. Based on observations from this experiment we then improved the auxiliary utility function to better account for user viewport orientation and added two new bandwidth allocation algorithms. We then conducted a subjective evaluation, the results from this experiment indicate that we are able to achieve bitrate savings of upto 65% with the best-performing bandwidth allocation approach while delivering the same subjective quality as compared to a non-adaptive streaming with the same codec.

R2.2: What are the acceptable quality levels amongst adjacent tiles to maximize the quality of the final point cloud reconstruction?

To address this research question, we compared several tile bandwidth allocation approaches using both objective and subjective evaluation. In general, we found that using a greedy approach to bandwidth allocation results in point cloud representations with drastic quality differences amongst adjacent tiles resulting in lower quality of the final reconstruction with obvious discontinuities at the point cloud surface. As our study focussed on human reconstructions these boundary artefacts were especially annoying if they lay on regions of interest such as the face of the reconstruction. This was especially evident at low bitrates under 10 Mbps where better quality was achieved using a non-adaptive approach. A uniform approach to bandwidth allocation resulted in small improvements to quality over the non-adaptive approach as a larger amount of the available bandwidth was spent on tiles that were not visible to the user. The utility-weighted hybrid bandwidth allocation approach was able to achieve the highest gains in quality as it ensured that quality transitions amongst adjacent tiles were not too drastic while maximizing the quality of high utility tiles. This was especially evident at high bitrates over 60 Mbps where this approach was able to approach the quality offered by offline state-of-the-art codecs like V-PCC.

Based on these results we then adapted this approach to live-captured pointclouds in a VR remote communication application. This leads us to research question 3:

R3: How can we optimize the delivery of streams of dynamic point clouds in VR remote communication?

In chapter 4, we construct a fully functional VR remote communication application with live captured point cloud user reconstructions. In the delivery pipeline, we apply the tiled adaptive streaming approach that we propose in chapter 3. We reduced the number of tiles, the quality levels per tile and adjusted the codec configurations used in order to achieve real-time streaming at framerates suitable for live communication. In addition, we implement modules for uncompressed streaming and traditional network adaptive streaming of point clouds. We then conducted a user study with 33 users and compared the three streaming conditions to assess the feasibility and impact of tiled user-centered delivery optimizations.

This research question was further divided into:

R3.1: How does tiled user-adaptive point cloud streaming impact the perceived quality of a remote user reconstruction?

In chapter 4, we evaluated the quality of tiled live-captured point clouds in real-time communication as opposed to the offline evaluation performed in chapter 3. We investigated this research question during the user study by having participants perform a training task three times in a VR remote communication session. Participants interacted with a live point cloud reconstruction of a trainer. They experienced remote communication under tiled adaptive streaming, network adaptive streaming and uncompressed streaming of the trainer's point cloud. From the data gathered, we demonstrate statistically significant gains to perceived visual quality and quality of interaction by using tiled adaptive streaming as compared to network adaptive streaming at 7 Mbps and at 14 Mbps. With tiled adaptive streaming at a target bitrate of 14 Mbps, we were able to achieve similar visual quality as compared to uncompressed streaming that required ca. 300 Mbps.

R3.2: What is the computational overhead of using tiled adaptive point cloud streaming?

To address this research question, we recorded playback statistics and system resource consumption for each participant during the user study. From the data gathered we observed up to a 10% reduction in CPU consumption with tiled adaptive streaming as compared to network adaptive streaming with similar levels of GPU and memory consumption. This resulted in significant improvements to end to end latency and point cloud playback framerate. At a target bitrate of 14 Mbps, we demonstrated a reduction in latency of ca. 75ms with an average increase in point cloud playback framerate of 4.6 frames per second by employing tiled adaptive streaming.

5.2. Discussion

This thesis focuses on investigating the suitability and the impact of interacting with tiled user-centered adaptive streams of dynamic point clouds. We embed our research in the context of VR remote communication with real-time user reconstructions. To conclude the last chapter of this thesis, we summarize the discussion items highlighted throughout the thesis and present some resources created in the course of this thesis that can prove useful for future research in the field.

5.2.1. Protocols for subjective quality assessment in VR

To evaluate the impact of lossy delivery optimizations on the perceived quality of point cloud content, it is crucial to select an appropriate subjective evaluation protocol. The selected protocol can affect the collected quality scores, the number of stimuli that can be tested and the sample size required.

The double stimulus method is based on simultaneously presenting two stimuli to participants; a reference and a degraded stimulus. This approach is popular in video quality assessment [7] and has also been adopted in point cloud quality assessment on 2D screens using uncompressed prerecorded video by standardization activities [8]. In interactive immersive viewing environments, this approach is challenging to implement as both stimuli need to be simultaneously rendered in a perceptually satisfying manner [9] while ensuring a fair comparison between stimuli in terms of viewing angle and proximity to the user's viewport. Previous research has relied on synchronous viewport movements for both stimuli [10, 11]. This approach is not suitable in 6DoF VR viewing environments where users are free to navigate the VR space without restrictions. Presenting the stimuli in this manner would result in large experiment session times and increases the risk of participants experiencing cybersickness.

Single stimulus methods, on the other hand, present stimuli one at a time to be rated independently. This approach can produce almost double the number of ratings within the same time frame compared to double stimulus methods. However, this approach can confound the impact of visual quality degradation with the participant's preferences in point cloud content [7] and can result in larger confidence intervals implying the need for more experiment participants. To mitigate the impact of content preference, a hidden reference approach can be employed where reference content is shown as a free-standing stimulus within the experiment and the scores are adjusted based on how the reference content is scored.

Previous research comparing the two methods on static point cloud reconstructions confirmed that both are statistically equivalent [12]. The double stimulus method was found to be more consistent in identifying the level of impairment. The single stimulus method was found to better classify geometry degradations such as octree pruning which more closely resemble compression artefacts at the cost of having larger confidence intervals.

In chapter 2, we found the choice of navigation paths had a significant impact on the evaluated quality and we decided to use a realistic immersive VR viewing environment where participants were able to navigate freely without restrictions. We employ the single stimulus method with hidden references. We observed a significant impact of content preference on the collected scores with some content being consistently rated lower including the reference stimulus. Results of the interviews also pointed out that the naturalness of gestures was an important criterion in assessing visual quality. Such components would not be normally evaluated in a double stimulus scenario; however, they are important in understanding how human perception reacts to digital humans. From our experiments and pre-trials, we learned that it is essential to train participants with example stimuli before performing the evaluation. We used content that is not included in the test in order to avoid bias. In addition, we also added dummy stimuli at the start of the test in order to ease participants into the task and the scores for these stimuli were subsequently discarded. To ensure the scores are not influenced by cybersickness we asked participants to fill out the simulator sickness questionnaire [13] before and after each experiment session. In addition, we collected playback statistics including the rendered framerate to ensure consistent viewing experiences across participants and stimuli.

5.2.2. Datasets

To develop and evaluate the performance of all the components in a point cloud delivery pipeline, there is a need for a large diverse body of publicly available point cloud datasets. Specifically, for VR remote communication there is a need for dynamic point cloud full-body human reconstructions captured with a wide range of sensors and capture setups. In the work presented in chapters 2 and 3, we use the most popular and widely used 8i voxelized full bodies dataset [5] with a total of four sequences. Other available datasets include the HHI Fraunhofer dataset [14] with two sequences and the Owlii dynamic human mesh sequence dataset [15] with four mesh sequences and associated sampled point clouds. The scarcity of publicly available datasets can lead to delivery components being designed and optimized to overfit a narrow range of available data and the characteristics of the associated acquisition method. In datasets acquired through photogrammetry we observe dense voxlelized photorealistic reconstructions with some extraneous noise over the dynamic sequence. In datasets acquired through sampling points from meshes, we observe some surfaces with a flat appearance and some of the motion capture resulted in movements that participants identified as unnatural. Point clouds live-captured for real-time applications with consumer depth sensors exhibit more noise, are sparse and generally not voxelized. In our work, we found that live-captured point clouds are more sensitive to depth noise and the lighting used in the capture space can significantly impact the fidelity of the reconstruction.

From the user study presented in chapter 2, we observe significant differences in the perceived quality delivered by the codecs under study in both the datasets used. We see significantly larger differences in the quality delivered by the two codecs when using point clouds captured using photogrammetry as compared to point clouds sampled from the surface of an animated mesh recorded using motion capture. In this case, the

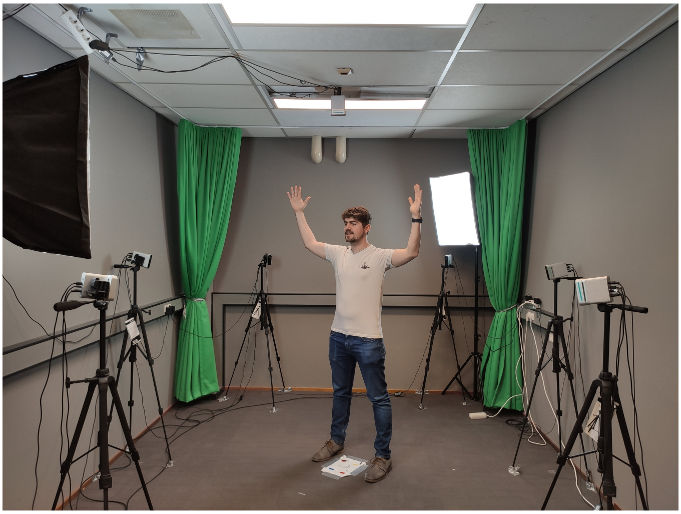


Figure 5.1: Physical camera arrangement from the CWIPC-SXR dataset

nature of the capture method played a role in perceived quality. While applying the user-adaptive streaming approach to the live-capture scenario in chapter 4 we found that a uniform allocation of bandwidth across tiles led to the most consistent gains in visual quality as compared to the best-performing utility-weighted allocation used in chapter 3.

A larger body of dynamic point cloud human reconstructions is needed to ensure consistent and robust optimizations to the various pipeline components. To this end, we contributed two publicly available datasets captured using different capture sensors and sensor fusion methods suitable for evaluating real-time delivery systems. Both datasets were captured using commodity hardware without pre-processing to preserve the noise and stitching artefacts inherent to current consumer depth sensors making them more realistic for developing and testing components used for VR remote communication.

CWIPC-SXR Dataset

The CWIPC-SXR dataset was captured using 7 calibrated and synchronized Azure Kinect DK sensors distributed around the subject [16] as shown in figure 5.1. The dataset presented dynamic point cloud reconstructions with audio targeting four key use cases for social XR namely; “Education and Training”, “Healthcare”, “Communication and Social interactions”, and “Performance and Sports”. The dataset features 45 dynamic sequences recorded using 23 different actors. Each actor was recorded individually along with the associated props including audio from a lavalier mic. All sequences were recorded at 30fps and are presented along with the raw color and depth videos that can be used to test new reconstruction and calibration methods. A sample of the recorded sequences is shown in Figure 5.2. The entire dataset is available at `https://www.dis.cwi.nl/cwipc-sxr-dataset/`.

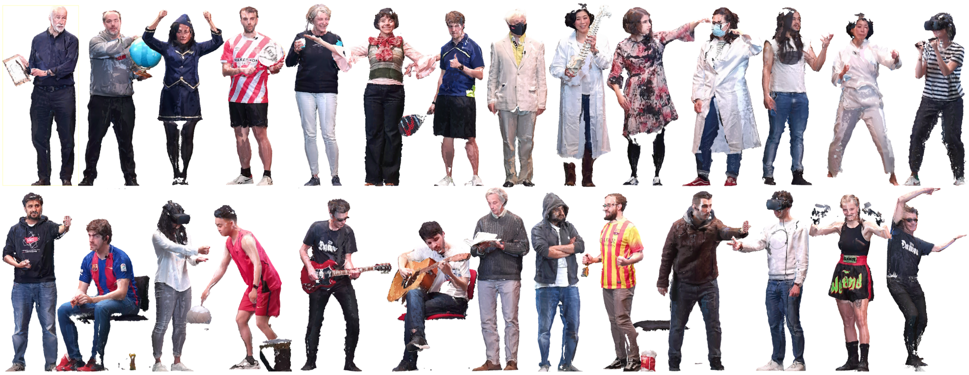


Figure 5.2: Point cloud sequences from the CWIPC-SXR dataset

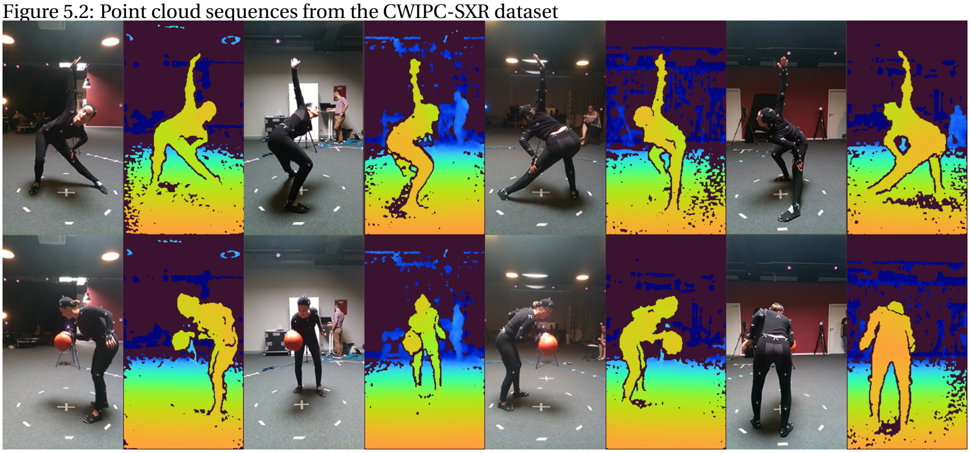


Figure 5.3: Samples of multiview RGBD recorded in the Human4D dataset

HUMAN4D DATASET

The Human4D dataset [17] was captured in a professional motion capture studio with 24 motion capture sensors and 4 Intel Real Sense D415 stereo-based consumer depth sensors. The dataset comprises a total of 66 sequences recorded using 4 professional actors performing a range of physical, social and daily activities. All sequences contain motion-captured dynamic meshes, time-varying meshes, point clouds and raw colour and depth videos recorded at 30fps. Actors were recorded together along with audio. In addition, the dataset includes benchmarking using state-of-the-art pose estimation and 3D compression algorithms. A sample of the actors being recorded and the associated depth maps are shown in Figure 5.3. The entire dataset is available at https://tofis.github.io/human4d_dataset/.

USER NAVIGATION DATA

In addition to point cloud content, there is also a need for large user navigation datasets in 6DoF VR. In other immersive modalities such as omnidirectional video, user interaction patterns have been the driving force to develop adaptive streaming approaches [18]. Previous work on evaluating point cloud compression [8] and adaptive streaming [6] has relied on predefined manually created navigation paths around the object or scene. This approach does not account for how users actually consume this content. While some datasets exist for 6DoF VR navigation in open spaces [19] and AR navigation while viewing volumetric video [20], there are no datasets of user navigation in 6DoF VR while viewing dynamic point cloud content.

In the objective evaluation presented in chapter 3, we observe significant differences in user navigation patterns in terms of viewport position and orientation over time. This results in navigation paths having a statistically significant impact on the recorded objective quality of compression and tiling. We recorded a dataset of user navigation patterns during the user study presented in chapter 2 that was then used in the objective evaluation presented in chapter 3. This dataset includes the position and orientation of the viewport of 26 users viewing 4 dynamic point cloud sequences logged at 30Hz. In chapter 3, we saw that navigation data can have a significant impact on objective visual quality. In some sequences, we observed upto 95% of the frames viewed contained the two frontal tiles of the point cloud object presenting an opportunity to optimize delivery. We also observe that participants had a greater spread of viewport positions around content that they preferred due to clear facial features and slow movements as compared to content that they disliked.

This dataset can help the research community tailor their efforts towards a true user-centered approach, fostering new research in the field. The dataset is made publicly available at the following link: `https://www.dis.cwi.nl/6dof-nav-dataset/`.

5.2.3. VR Remote Communication Application Scenarios

In chapter 4, we constructed a two-user remote communication pipeline in an empty grey room in order to avoid distractions during the evaluation. Users were able to navigate the space freely with physical movements as well as teleporting to cover larger distances. In VR remote communications applications of the future, we envision additional navigation methods and interaction techniques as well as using a blend of user representations based on the context. We believe that point cloud delivery optimizations in the future need to be application dependent. Different activities in VR remote communication can place different demands on how the auxiliary utility function for tiles is defined, for instance, the medical examination application presented in [21] could require the tiles corresponding to the site of the patient's injury to always retain high utility and avoid temporal quality fluctuations while the physician makes the diagnosis. The playback environment is another important factor to consider, virtual objects in the scene can occlude the reconstruction presented to the user presenting an opportunity to further optimize the delivery. Interactive objects in the scene and narrative elements such as virtual characters can also be used to infer a user's visual attention and have been shown to influence user navigation patterns in 6DoF VR [22]. In addition, interaction techniques and methods of locomotion such as teleporting, controller-based movements and physical movements can influence user navigation patterns and affect the uncertainty in predicting viewport locations necessitating more conservative quality adaptations. Delivery optimizations in future will need to evolve to meet the demands of novel social VR application scenarios.

Using some of the infrastructure that we developed for the user study presented in

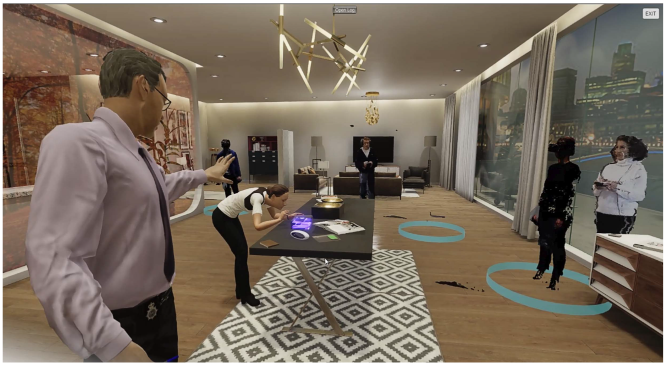


Figure 5.4: Collaborative Murder Mystery Application

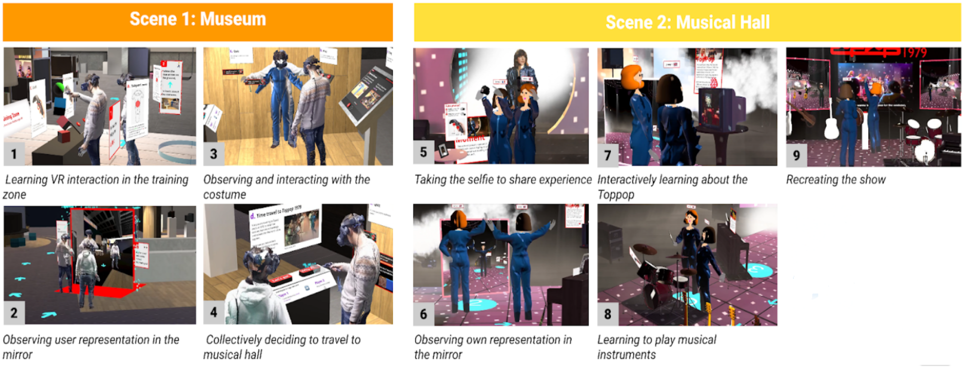
chapter 4, we created demonstrations of two different application scenarios to further showcase the potential of VR remote communication. Both these systems can benefit from the adaptive streaming approach presented in this thesis with modifications to account for user representation formats, display devices and interaction techniques.

COLLABORATIVE MURDER MYSTERY APPLICATION

Four remote users are brought together at a crime scene in an apartment to collaboratively solve a murder mystery. The application supported users in 6DoF VR and users viewing the experience on a 2D screen. All users were reconstructed as point clouds in real-time and could interact in the virtual space as themselves. An example of users in the experience is shown in Figure 5.4. VR users were able to navigate the scene using physical movements and four fixed teleportation targets. 2D screen users were able to freely navigate the scene using a controller. In addition, the participants interacted with three virtual characters that were rendered using synchronized prerecorded animated meshes. In this application, we showcase different viewing devices, navigation and interaction techniques in the same experience. Such applications demonstrate the need for further delivery optimizations that perform user-specific adaptation beyond viewport position and orientation.

CULTURAL HERITAGE APPLICATION

Two remote users are brought together to a virtual museum. Both users were initially represented using live captured point cloud reconstructions in the first scene and later as mesh-based avatars in the second scene. An overview of the different stages of the experience is shown in Figure 5.5. In this application, we showcase a blend of user reconstruction types while allowing users to interact with artefacts in the virtual museum. Users are allowed unrestricted teleportation and need to travel much larger distances to navigate within the museum. In addition, the application features a mirror where users can view their own reconstruction along with remote users. This presents a unique challenge for adaptive streaming of volumetric content where we need to prioritize tiles based on the angle of incidence rather than direct surface orientation. This type of application scenario also presents other challenges such as extended periods where users inspect and interact with museum exhibits rather than remote users.



Learning VR interaction in the training zone

Observing and interacting with the costume

Taking the selfie to share experience

Interactively learning about the Toppop

Recreating the show

Observing user representation in the mirror

Collectively deciding to travel to musical hall

Observing own representation in the mirror

Learning to play musical instruments

Figure 5.5: Cultural Heritage Application

5.2.4. Point Cloud Compression and Streaming

In the work presented in this thesis, we use two point cloud codecs; the MPEG Anchor codec [3] and the state-of-the-art V-PCC standard codec [1]. The MPEG anchor codec is currently the only point cloud codec with low delay encoding and decoding making it suitable for real-time applications. We use this codec to implement our adaptive streaming approach. This codec uses the octree data structure to encode point cloud geometry and the number of points in the decoded point cloud varies exponentially with the tree depth. At low bitrates, this can result in sparse representations requiring large point sizes to appear watertight resulting in a blocky appearance. As the compression is performed in a 3D space this codec design allows for future optimizations based on human perception of 3D objects. The V-PCC standard codec on the other hand projects the point cloud geometry and colors to a 2D plane and leverages existing video codecs. This approach maintains the number of points across target bitrates but is optimized based on human perception of 2D videos. This approach achieves significantly higher rate-distortion performance but suffers from high encode complexity making it suitable for offline applications. In chapters 2 and 3, we use this codec to serve as a benchmark. The adaptive streaming approach presented in chapters 3 and 4 is based on segmenting the point cloud reconstruction into independently decodable tiles. This spatial segmentation leads to a larger encoded payload as compared to a non-adaptive approach due to a loss of entropy lowering the compression efficiency. This trade-off yields significant improvements to visual quality using the MPEG Anchor codec. This trade-off needs to be tested with any future low-delay point cloud codecs that are proposed to ensure quality gains are maintained. In addition, future low-delay codecs with progressive decoding can achieve further gains with a larger adaptation set by encoding multiple representations or quality levels in a single encode cycle in real-time.

We presented a low-complexity tiling approach intended for real-time systems, due to the relative simplicity with which the tiles are created and their utility is computed. In particular, our system aims at exploiting information related to the acquisition system, such as the positions of the cameras used to capture the point cloud contents, to optimize delivery with respect to the user's point of view. This approach is based on the assumption that the point cloud object can be approximated by a convex hull, while this appears to be working in our case, this might not generally be true, which would lead to different quality gains. Nguyen et al. [23] present a taxonomy of five classes of segmentation algorithms: edge based, region based, attribute based, model based, and graph based. Attribute-based segmentation methods can be used to tile point clouds, and account for surface orientation. These methods rely on estimating additional attributes for each point, e.g., normals, to obtain the surface orientation before clustering. Thus, these solutions are unsuitable for real-time applications, as inferring the surface normal requires repeated eigendecomposition for the local neighbourhood of each point. In the user study presented in Chapter 2, some participants reported that seeing artefacts in body extremities and in the face of the reconstruction was unpleasant. Further study into body part segmentation and quality perception is required to optimize tiling and tile selection strategies for humans engaging in conversation as compared to prerecorded content, objects and scenes.

The evaluation of our proposed tiling approach for point cloud streaming was conducted under controlled conditions. In particular, the available bandwidth for each time instance was modelled after the MPEG Test Conditions [2], which are defined to be representative of different degradation levels for the content being studied. Thus, in our evaluation scenario, the bit budget was a) constant for the duration of the dynamic sequence, and b) known in advance. Real network conditions, however, seldom follow such an ideal scenario. Further analysis in more adverse network conditions is needed in order to evaluate the gains a tiling approach can bring when the rate allocation can change unpredictably over time.

5.3. Conclusion and Future Work

Volumetric point cloud content viewed in immersive VR environments presents novel challenges in optimizing the delivery pipeline. In this final section we conclude the thesis and highlight some limitations and avenues for future work in this field. In general, tiled adaptive streaming techniques have received significant research attention for omnidirectional videos [24–26] and point clouds [27–31]. However, further work is required to better understand the case of live real-time human point cloud reconstructions. Although the tiles in the human reconstruction are independently decodable, their quality cannot be optimized in isolation based on available bandwidth and viewport. In future, we will explore body part segmentation and tracking to define better tiles and tile utilities based on visual saliency to better approximate how content is actually consumed in remote communication applications.

In this thesis, we embed our research in the application scenario of VR remote communication with real-time point cloud user reconstructions. In order to optimize content delivery we segment the point cloud into independently decodable tiles allowing us to optimize the quality of the final reconstruction by allocating more bandwidth to tiles visible to the user from a given viewport. Our work on establishing a framework to playback and evaluate user-centered tiled streams of dynamic point clouds uses a low-complexity algorithm in order to partition the content in multiple tiles, which can then be encoded at different qualities. Specifically, the selected spatial segmentation is designed for live captured point clouds; as such, it exploits information from acquisition sensor placement to infer surface orientation of spatial segments of a given point cloud content. In chapters 2 and 3, due to the lack of publically available point cloud datasets with labelled acquisition information we emulated our approach on a popular point cloud dataset[5] that has been successfully used in the past to test adaptive streaming for point cloud contents, both objectively [6, 32] and subjectively [33]. We operated under the assumption that the content could be separated into an even number of non-overlapping, opposite-facing tiles, in order to apply our utility function. While this assumption appears to be valid for the prerecorded point clouds used in the user study, the performance of our proposed system with real-time point clouds may vary, depending on the accuracy and noise level of the acquisition sensors. In chapter 4, we reduced the number of tiles to 3 and reduced the adaptation set to have two quality levels per tile. Future codec optimizations with progressive decoding can lead to an expanded adaptation set with further gains to final visual quality and playback performance. Alternative spatial segmentation schemes, such as body part segmentation based on pose estimation, can also lead to further gains in perceived quality, as participants reported preferring content where the reconstruction's face has a high-quality representation.

To evaluate the proposed adaptive streaming strategies in chapter 3, the network conditions and available bandwidth were set based on the CTC defined by the MPEG standardization activity [34]. While these cover a wide range of bit rates (3 - 117 Mbit/sec), similarly to [32], the bit rate budget was constant for the duration of the playback sequence. In chapter 4, we use a fixed target bitrate of 7 Mbps and 14 Mbps to perform the evaluation. The constant bit rate budget was selected to avoid introducing biasing factors in the subjective evaluation, as a variable bit rate with adaptive tiling might have been a confounding factor for both visual quality and cybersickness. In order to adequately assess the performance of the system, in future we will analyze the performance of out approach under adverse and temporally varying network conditions to ensure the performance gains are maintained.

In chapter 4, we used a direct TCP connection with fixed audio quality between the two remote participants in order to control the network conditions for the experiment. However, we implemented and ensured our approach can scale up to service a larger number of users using chunked HTTP transfers and fragmented ISOBMFF/MP4 as a transport stack for live streaming. This introduces additional latency in the pipeline and in turn latencies in quality transitions over time. Future work in the field should evaluate delivery optimizations by accounting for this uncertainty in the user's viewport location and available bandwidth and adapt the delivery of both point cloud reconstructions and audio.

There is a need for new standardized tasks that can be used to evaluate VR remote communication. The existing ITU recommendations are insufficient to handle novel interaction techniques and immersive content inherent to VR communication. In chapter 4, we utilized a neck exercise training task as it was more visually focused and the interaction was repeatable with a confederate trainer. Further study into other use cases and scenarios are required to evaluate emerging VR remote communication systems.

In this thesis, we first looked at point cloud quality assessment and presented the first approach to subjective evaluation used during the MPEG point cloud compression standardization activity. We also presented the first evaluation of dynamic point clouds in an immersive VR viewing environment. We then focused on user-centered adaptive streaming of pre-recorded content at the client-side. We proposed a low-complexity tiling approach along with a novel tile selection strategy and demonstrate the validity of our approach through both objective and subjective evaluation. Finally, we deployed our approach to a real-time two-user VR remote communication application with live-captured point cloud user reconstructions. We evaluated the impact of user-centered adaptive streaming in the context of this application using a novel evaluation methodology for remote communication.

REFERENCES

[1] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, *et al.*, *Emerging MPEG Standards for Point Cloud Compression,* IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**, 133 (2019).

[2] MPEG3DG and Requirements, *Call for proposals for point cloud compression,* ISO/IEC JTC1/SC29 WG11 N16732, Geneva, CH (2017).

[3] 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 (2016).

[4] ITU-T P.910, *Subjective video quality assessment methods for multimedia applications,* International Telecommunication Union (2008).

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

[6] J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, and H. Hellwagner, *Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression,* in *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19 (Association for Computing Machinery, New York, NY, USA, 2019) pp. 2405–2413.

[7] 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 (2016).

[8] V. Baroncini, P. Cesar, E. Siahaan, I. Reimat, and S. Subramanyam, *Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression,* ISO/IEC JTC1/SC29/WG11 M41786 (2017).

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

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

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

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

[13] 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**, 203 (1993).

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

[15] Y. L. Yi Xu and Z. Wen, *Owlii dynamic human mesh sequence dataset,* ISO/IEC JTC1/SC29/WG11 m41658 120th MPEG Meeting, Macau (2017).

[16] 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*, MMSys '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 300–306.

[17] A. Chatzitofis, L. Saroglou, P. Boutis, P. Drakoulis, N. Zioulis, S. Subramanyam, B. Kevelham, C. Charbonnier, P. Cesar, D. Zarpalas, S. Kollias, and P. Daras, *Human4d: A human-centric multimodal dataset for motions and immersive media,* IEEE Access **8**, 176241 (2020).

[18] X. Corbillon, F. De Simone, G. Simon, and P. Frossard, *Dynamic adaptive streaming for multi-viewpoint omnidirectional videos,* in *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18 (ACM, New York, NY, USA, 2018) pp. 237–249.

[19] J. Chakareski and M. Khan, *Wifi-vlc dual connectivity streaming system for 6dof multi-user virtual reality,* in *Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, NOSSDAV '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 106–113.

[20] E. Zerman, R. Kulkarni, and A. Smolic, *User behaviour analysis of volumetric video in augmented reality,* in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)* (2021) pp. 129–132.

[21] *World's first volumetric video conference (point clouds) over a public 5g network: A medical emergency example,* `https://vrtogether.eu/2020/11/09/worlds-first-volumetric-video-conference-point-clouds-over-a-public-5g-network/`.

[22] S. Rossi, I. Viola, J. Jansen, S. Subramanyam, L. Toni, and P. Cesar, *Influence of narrative elements on user behaviour in photorealistic social vr,* in *Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '21)*, MMVE '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 1–7.

[23] A. Nguyen and B. Le, *3d point cloud segmentation: A survey,* in *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (2013) pp. 225–230.

[24] C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, *A survey on 360° video streaming: Acquisition, transmission, and display,* ACM Comput. Surv. **52** (2019), 10.1145/3329119.

[25] M. Zink, R. Sitaraman, and K. Nahrstedt, *Scalable 360° video stream delivery: Challenges, solutions, and opportunities,* Proceedings of the IEEE **107**, 639 (2019).

[26] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, *An http/2-based adaptive streaming framework for 360° virtual reality videos,* in *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17 (Association for Computing Machinery, New York, NY, USA, 2017) p. 306–314.

[27] S. Petrangeli, G. Simon, H. Wang, and V. Swaminathan, *Dynamic adaptive streaming for augmented reality applications,* in *2019 IEEE International Symposium on Multimedia (ISM)* (2019) pp. 56–567.

[28] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li, and Y. Ji, *Point cloud video streaming: Challenges and solutions,* IEEE Network **35**, 202 (2021).

[29] J. v. d. Hooft, M. T. Vega, T. Wauters, C. Timmerer, A. C. Begen, F. D. Turck, and R. Schatz, *From capturing to rendering: Volumetric media delivery with six degrees of freedom,* IEEE Communications Magazine **58**, 49 (2020).

[30] F. Qian, B. Han, J. Pair, and V. Gopalakrishnan, *Toward practical volumetric video streaming on commodity smartphones,* in *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, HotMobile '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 135–140.

[31] L. Wang, C. Li, W. Dai, S. Li, J. Zou, and H. Xiong, *Qoe-driven adaptive streaming for point clouds,* IEEE Transactions on Multimedia , 1 (2022).

[32] S. Subramanyam, I. Viola, A. Hanjalic, and P. Cesar, *User Centered Adaptive Streaming of Dynamic Point Clouds with Low Complexity Tiling,* in *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 3669–3677.

[33] 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)* (2020) pp. 1–6.

[34] MPEG 3DG and Requirements, *Complementary PCC Test Material,* ISO/IEC JTC1/SC29 WG11 Doc. N16716, Geneva, CH (2017).

Acknowledgements

I would like to express my deepest gratitude to my promoters, Pablo Cesar and Alan Hanjalic, for their invaluable guidance, support, and encouragement throughout my PhD journey. Their expertise and insights have been instrumental in shaping my research and bringing this thesis to fruition. I would like to thank Pablo for giving me the opportunity to work on this exciting and timely research topic and for introducing me to MPEG standardization. When I started this journey, Pablo taught me about the importance of subjective testing in multimedia computing and the difference between fundamental science and engineering. I would not have been able to complete this work without his guidance and support.

A special thanks goes out to my daily supervisor and copromotor Irene Viola, whose constant support, constructive feedback, and unwavering patience have been a source of inspiration and motivation. I especially thank her for the long work days and dedication on so many deadline days. She taught me a lot about subjective quality assessment and statistical techniques for HCI research. I also want to thank Francesca for mentoring me during the first year of my PhD.

I am also grateful to the European Commission for funding the VRTogether project that made this work possible. The project provided me with resources and opportunities to pursue this research and introduced me to excellent researchers who were part of the consortium.

I am indebted to Pablo, Alan, Omar and Simon for pushing me to complete this thesis. I am very grateful for their patience and support in the final period of this PhD.

I would also like to extend my heartfelt thanks to my colleagues, Abdo and Tianyi, for their camaraderie and unwavering support during our days in the trenches. Their friendship and encouragement have made this journey more enjoyable and have provided me with the motivation to persevere through the challenges. I also want to thank Abdo, Tom, Artem and Ke for all the memorable Friday evenings after long work weeks.

I want to thank my colleague Jack Jansen for putting up with my poor version control discipline and for the excellent hacker-jack solutions to the endless sea of bugs and issues we faced during the course of our research. I will always look back fondly on the time that you, Nacho, and I spent tinkering in the lab to try to get the point cloud communication system to behave.

I want to express my gratitude to all my coauthors and collaborators during my PhD years. Irene and Jie made invaluable contributions during our initial work on point cloud quality assessment. Irene and Evangelos mentored me and made invaluable contributions to our work on assessing tiling for point cloud streaming. Evangelos and Fons spent long hours when they volunteered to help me run the final experiment evaluating end-to-end communication.

I would like to thank my fellow PhDs Tianyi, Xuemei and Carlos for the camaraderie and the coffee break talks on politics and the English language. I want to extend my

gratitude to all my excellent colleagues at the DIS group Alina, Kees, Fons, Susie, Nacho, Yanni and Silvia for their support, for participating in my experiments and for agreeing to be guinea pigs back when the point cloud rendering system sometimes doubled as a motion sickness simulator. I also want to thank the old guard of PhD students Simon, Marwin, Jan Willem and Rufael for welcoming me to the group during the early days of my PhD.

Lastly, I want to express my deepest appreciation to my parents and my brother MP, for their unwavering support and encouragement. Their belief in me has been a constant source of strength and motivation and has carried me through the toughest times during this journey.