April 15–18, 2024
Bari, Italy

Association for
Computing Machinery

Advancing Computing as a Science & Profession

MMVE ’24

Proceedings of the 2024
16th International Workshop on Immersive Mixed and Virtual Environment Systems

Sponsored by:

ACM SIGMM in cooperation with ACM SIGCOMM and ACM SIGMOBILE

Association for
Computing Machinery

Advancing Computing as a Science & Profession

The Association for Computing Machinery
2 Penn Plaza, Suite 701
New York, New York 10121-0701

Notice to Past Authors of ACM-Published Articles
ACM intends to create a complete electronic archive of all articles and/or other material previously published by ACM. If you have written a work that has been previously published by ACM in any journal or conference proceedings prior to 1978, or any SIG Newsletter at any time, and you do NOT want this work to appear in the ACM Digital Library, please inform permissions@acm.org, stating the title of the work, the author(s), and where and when published.

ISBN: 979-8-4007-0618-9

Additional copies may be ordered prepaid from:
ACM Order Department
PO Box 30777
New York, NY 10087-0777, USA

Phone: +1 800 342-6626 (USA and Canada)
+1 212 626-0500 (Global)
Fax: +1 212 944-1318
Email: acmhelp@acm.org
Hours of Operation: 8:30 am–4:30 pm ET

Foreword

We are pleased to present the technical program of the 16th ACM International Workshop on Immersive Mixed and Virtual Environment systems (MMVE) 2024. This workshop has always embraced a multidisciplinary approach, exploring not only the evolution of immersive experiences but also the crossroads where immersive technology intersects with diverse domains. Co-located with ACM Multimedia Systems Conference (MMSys) 2024, MMVE allows the gathering and interaction of researchers in the field of immersive technology, from both academia and industry, with multimedia system researchers.

This year MMVE received an impressive number of 21 high-quality submissions spanning a broad spectrum of multimedia topics, including virtual reality, multisensory experience, point cloud compression, quality of experience, social virtual reality platform, avatar design, and gaming. Thanks to the hard and valuable work of the 24 Technical Program Committee (TPC) members, each submission underwent a rigorous review process and most of them received three high-quality reviews. As a result, 13 full papers and 2 short ones will be presented in the workshop. Keeping the tradition of MMVE as an interactive and discussion-oriented workshop that serves as an inclusive and interdisciplinary forum, the program has been structured to facilitate engagement and collaboration. Two oral sessions will serve as an opportunity to showcase 9 of the accepted papers, while a poster session for the remaining works will provide a further chance to discuss and interact among authors and participants.

We would like to take this opportunity to thank all the people who have contributed to the success of MMVE 2024, including all the authors for submitting their research efforts, the TPC members for their valuable feedback during the review process, essential to create a high quality technical program. We would also like to thank the MMVE Steering Committee and organisers of MMSys 2024 for their support and help in shaping MMVE 2024.

We hope that MMVE 2024 will be an engaging, informative, and enjoyable experience for all participants.

MMVE 2024 Organizing Committee

Silvia Rossi, CWI, The Netherlands – *General Chair*
Débora Christina Muchaluat-Saade, UFF, Brazil – *Technical Program Chair*
Thomas Röggla, CWI, The Netherlands – *Web Chair*

Organization

GENERAL CHAIR:
Silvia Rossi, CWI, The Netherlands

TECHNICAL PROGRAM CHAIR:
Débora Christina Muchaluat-Saade, UFF, Brazil

WEB CHAIR:
Thomas Röggla, CWI, The Netherlands

TECHNICAL PROGRAM COMMITTEE:
Evangelos Alexiou - Xiaomi Technology, The Netherlands
Roberto Azevedo - Disney Research, Switzerland
Federica Battisti - Università degli Studi di Padova, Italy
Jean Botev - University of Luxembourg, Luxembourg
Pablo Cesar - CWI, The Netherlands
Alexandra Covaci - University of Kent, United Kingdom
Joel dos Santos - CEFET, Brazil
Herman Engelbrecht - Stellenbosch University, South Africa
Mylene Farias - UNB, Brazil
George Ghinea - Brunel University, United Kingdom
Alan Guedes - UCL, United Kingdom
Jesus Gutierrez - Universidad Politécnica de Madrid, Spain
Marina Josué - Fluminense Federal University, Brazil
Conor Keighrey - Technical University of the Shannon, Ireland
Tanja Kojic - TU Berlin, Germany
Yao Liu - Rutgers University, United States of America
Niall Murray - Technical University of the Shannon, Ireland
Wei Tsang Ooi - National University of Singapore, Singapore
Marta Orduna - Nokia, Spain
Pablo Perez - Nokia Bell Labs, Spain
Celso Alberto Saibel Santos - UFES, Brazil
Sam van Damme - Ghent University, Belgium
Irene Viola - CWI, The Netherlands
Sara Vlahovic - University of Zagreb, Croatia

Contents

Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments

Pegah Salehi[*]
SimulaMet, Norway

Syed Zohaib Hassan[†]
SimulaMet, Norway

Gunn Astrid Baugerud
OsloMet, Norway

Martine Powell
Griffith University, Australia

M. Cayetana López Cano
OsloMet, Norway

Miriam S. Johnson
OsloMet, Norway

Ragnhild Klingenberg Røed
OsloMet, Norway

Dag Johansen
UiT The Arctic University of Norway

Saeed Shafiee Sabet
SimulaMet, Norway

Michael A. Riegler[†]
SimulaMet, Norway

Pål Halvorsen[†]
SimulaMet, Norway

ABSTRACT

The current study aims to evaluate and compare the subjective quality of an AI-based training system developed for conducting child interviews, focusing on the distinction between immersive 3D (using virtual reality) and 2D desktop environments. To this end, a structured user study was conducted, involving 36 participants who were exposed to these two distinct environments. The study evaluated various aspects of user experience, namely *presence*, *usability*, *visual fidelity*, *emotion*, *responsiveness*, *appropriateness*, and *training effectiveness*. The findings reveal significant differences in user experience between the 2D and 3D environments. Notably, the 3D environment enhanced *presence*, *visual fidelity*, *training effectiveness*, and *empathy*. In contrast, the 2D environment was favored for *usability*. The study highlights the potential of immersive VR while also pointing out the need to improve the system response and emotional expressiveness of the avatars.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Virtual reality**;
• **Computing methodologies** → *Machine learning*.

KEYWORDS

Virtual Reality (VR), Immersion, Quality of Experience (QoE), Large language model (LLM)

ACM Reference Format:
Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, M. Cayetana López Cano, Miriam S. Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A. Riegler, and Pål Halvorsen. 2024. Immersive Virtual Reality in Child Interview Skills Training: A Comparison of 2D and 3D Environments. In *Proceedings of the international workshop on immersive mixed and virtual environment systems (MMVE '24).* ACM, New York, NY, USA, 7 pages. https://doi.org/3652212.3652219

1 INTRODUCTION

Child abuse is a critical global issue that negatively affects children's development, mental and physical well-being. Meta-analysis studies have estimated that 22.6% of children experience physical abuse and 11.8% experience sexual abuse before reaching adulthood [36]. In particular, less than 15% of child sexual abuse (CSA) cases are corroborated by physical evidence [42], and in 70% of these cases, the child is the sole witness [11]. Properly conducted investigative interviews are crucial for prosecution, as children can be reliable witnesses when interviewed according to best-practice guidelines [4]. Such guidelines encourage communication with the child through open-ended questions to obtain elaborate and relevant evidence details [7]. Unfortunately, these guidelines are not often followed despite investments in training programs [18]; interviewers tend to use too many suggestive and closed questions, and directive questions generate short, rather than elaborate, responses [1, 2, 8]. This is due to training programs that contain inadequate practice opportunities (with feedback) that shape interviewer performance in the use of open questions [44]. Recent studies, including a decade-long Norwegian study, indicate that there is no significant advancement in interview quality despite training innovations. Mock interviews with trained respondents have been shown to be highly effective in shaping interviewer performance [27].

However, face-to-face training with a trained respondent is expensive and cumbersome to set up. It requires the availability of both parties and considerable investment to ensure that the trained respondent responds in a way that effectively shapes the performance of the trainee interviewer. To potentially improve the update of practice opportunities, we introduced an innovative AI-based training platform to conduct investigative interviews [14, 31]. This platform integrates VR technology with an advanced avatar system. The avatar is designed to emulate the responses of children in high-fidelity abuse scenarios, enhancing the realism of the digital training environment. The core of our system lies in the seamless

[*] Also affiliated with UiT The Arctic University of Norway
[†] Also affiliated with OsloMet, Norway

MMVE '24, April 15-18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04…$15.00
https://doi.org/3652212.3652219

integration of state-of-the-art natural language processing (NLP) and vision technologies. NLP empowers the avatar with the ability to understand and express human language, while vision technologies enable them to present visually accurate and responsive representations of human-trained respondents. This dual-technology approach effectively simulates the complexities involved in interviewing child abuse victims, providing a crucial training tool for professionals in this field.

In our previous work [14], we conducted a user study that evaluated the efficacy of different interactive platforms, including VR, 2D desktop, audio, and text chat; however, it lacks crucial elucidation aspects of realism, such as *emotion* and *visual fidelity*. Also, the study may not have possessed ideal statistical power to discern certain subtle effects. This study focuses on a detailed assessment of qualitative feedback within both 2D and 3D visual environments, using a comprehensive questionnaire to investigate various evaluation aspects. In this paper, we present the following main contributions:

- Fine-tuning of the GPT-3 [6] using interview data from abused children and seamlessly integrating the dialogue model with a Unity3D framework.
- Demonstration of the potential of virtual reality (VR) for professional training in real-world scenarios.
- A comparison of trainee interviewers' perceptions across different evaluative aspects in 2D and 3D environments, including *presence*, *usability*, *visual fidelity*, *emotion*, *responsiveness*, *appropriateness*, *training effectiveness*, and *empathy*.

2 RELATED WORK

Child avatar training systems aimed at improving interview quality exist, but none are fully automated. One uses prerecorded child responses manually selected by an operator [9], another [26] employs a probabilistic rule-based algorithm for response selection after a person has categorized the question manually, that still requires human intervention. The existing systems do not utilize advanced large language models such as GPT-3 [6]. Instead, they are constrained and limited in generating responses, relying on predetermined sentences from which to select a response. Also, constrained by their dependency on human operators, these systems can have potential issues with flexibility, operational costs, and error rates.

Parallel to these developments, the emergence of VR as a dynamic educational tool is significant [25]. Several studies have shown that it enhances direct learning experiences and helps in memory retention [29, 30]. It increases the engagement and motivation of the learner [24], and improves the understanding of spatial and visual concepts [17]. VR also facilitates better decision-making in simulations [30], although its effectiveness depends on the application of sound pedagogical methods [12].

Furthermore, the adaptability and acceptability of VR, particularly in specialized training scenarios, are noteworthy [13]. Furthermore, the efficacy of VR in job interview training for individuals with serious mental illness underscores its wide applicability in various educational and training contexts [41].

The comparative analysis between 2D and 3D VR environments further emphasizes these benefits. Research indicates that the 3D environment generates stronger emotional reactions than the 2D interfaces when the same content is presented [19, 34]. This increased participation in VR is correlated with its proven efficacy in enhancing educational outcomes, as evidenced in studies focusing on vocabulary acquisition and memory retention [20, 21]. However, Madden et al. [22] did not identify significant differences in learning about moon phases, indicating that the effectiveness of VR may depend on the specific nature of the task.

Current research trends in this domain include the use of electroencephalogram (EEG) measurements to gain more profound insights. These studies reveal a greater sense of presence in 3D environments and a reduced cognitive load in comparison to 2D tasks [35]. Notably, cognitive load appears higher in 2D tasks, suggesting that 3D technologies might provide cognitive benefits in learning contexts due to their reduced cognitive demands [10]. In support of this notion, Tian et al. [38] experienced that VR's stereoscopic vision can lead to increased emotional arousal, further differentiating 3D experiences from 2D ones.

This study aims to obtain feedback on the effectiveness of our tools and to gain a deeper understanding of the experiences and perceptions of the interviewers associated with child abuse.

3 SYSTEM ARCHITECTURE

The architecture of our interactive child avatar system, which supports both 2D and 3D simulation environments, is shown in Figure 1. The system is structured into three main components: i) the language component, which features large language models (LLMs) based on OpenAI's GPT-3 [6]; ii) the speech synthesis component leveraging IBM Watson services for effective speech-to-text (STT) and text-to-speech (TTS) capabilities; and iii) a Unity-based user interface supporting two distinct interactive modalities: a 3D virtual environment using the Oculus Quest2 Head Mounted Display (HMD), and a 2D virtual environment.

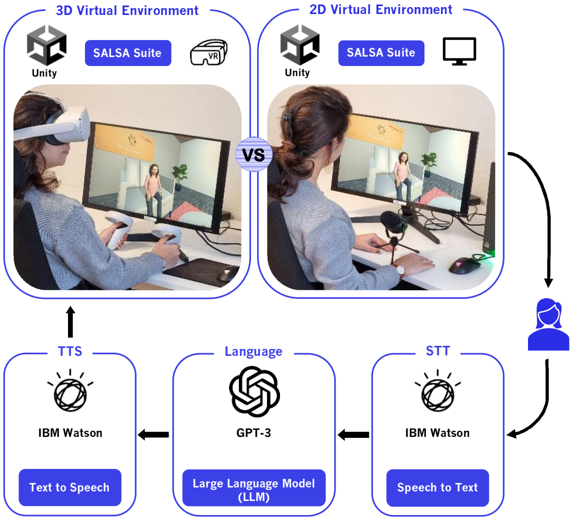


Figure 1: Architecture of the child avatar system for the comparative study.

3.1 Language

The model's language component is designed to process and respond to interviewers' inquiries, emulating a child's conversational patterns. This is based on a dataset of interview transcripts from the Centre for Investigative Interviewing at Griffith University, Australia [28]. The data set includes mock interviews, conducted by trained professionals, that simulate interactions between actors representing children and interviewers from Child Protection Services or law enforcement. In our earlier work [14], our dialogue model was developed using the RASA framework. Now we have fine-tuned the GPT-3 [6] Davinci model for two specific situations: sexual abuse and physical abuse. This was done using 10 simulated forensic interviews with children aged 6 to 8 years who were potential victims. The aim was to produce dynamic and appropriate responses to the questions asked by the interviewers in these contexts. The dialogue model was integrated with a Unity-based user interface using OpenAI API calls.

3.2 Speech Synthesis

IBM Watson's STT and TTS services are used to link the dialogue model with the interactive user interface. The IBM TTS API primarily offers adult voices; however, for a more realistic interaction, we altered the pitch and speed of a female voice to simulate a childlike sound. This modification was based on feedback from a pilot study that indicated that the voices of adults for the children were not well received by the participants [15, 32].

3.3 Visual Environment

This system allows users to interact with a virtual child avatar in two distinct environments. Despite the difference in environments, the system maintains a consistent dialogue model in its back-end, ensuring uniform responses from the avatar in both settings. The system's front-end, developed using the Unity game engine, offered two environments: a 3D environment accessible via an Oculus Quest 2 HMD and a 2D environment displayed on a 24-inch desktop monitor. Both 2D and 3D environments feature the same avatar, created using the Unity Multipurpose Avatar (UMA) open source project [39], which allows the customization of character meshes and textures. For realistic avatar movements, the Salsa Suit asset [40] is employed to synchronize lip, eye, and head movements with a generated voice. To enhance naturalism, prerecorded animations are used to animate the avatar's hands and neck.

4 EXPERIMENT DESIGN

This study involved two groups of participants: experienced investigative interviewers and individuals with backgrounds in psychology, criminology, or related fields but who lacked professional interviewing experience. Recruitment of participants was facilitated through the support of managerial personnel at their respective workplaces, guiding the researchers towards professionals best suited for participation. Eligibility for selection required experience in the criminal justice sector, with a preference for those who had previously worked as investigative interviewers. Out of those approached, all but a few individuals agreed to participate.

The methodology involved simulations in both 2D and 3D settings in random order. Each participant engaged in these environments for a duration ranging from five to ten minutes. Following the simulations, participants were asked to complete a questionnaire. This questionnaire focused on their experiences and observations while interacting with the child avatars in both the 2D and 3D environments. The study was conducted over six separate days without any specific arrangement to distinguish between participants with and without experience in interviewing children.

Our research has received ethical approval from the Norwegian Agency for Shared Services in Education and Research (SIKT) (project number #614272), titled "Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars." Additionally, this study was approved by Griffith University, Australia (project number #2023/501), titled "Evaluation of Child Avatar Interview Simulation Learning Activity."

4.1 Participant Demographics

Our study initially included 39 individuals, but with the exclusion of three participants who only interacted in the 2D environment, the number of participants considered for the final analysis was reduced to 36. Within this group, 24 identified as female, 11 as male, and one participant preferred not to disclose gender. The age distribution revealed that most of the participants, 26 participants, fell within the 30-49 age range, 9 were over 50 years old, and one was under 29 years old. Regarding their professional background in child interviewing, 14 participants had no prior experience, 11 had more than ten years of experience, and the remaining 11 had less than ten years of experience. Regarding their exposure to VR, 16 participants had previous VR experience, while 20 had not used VR technology before.

4.2 Questionnaires

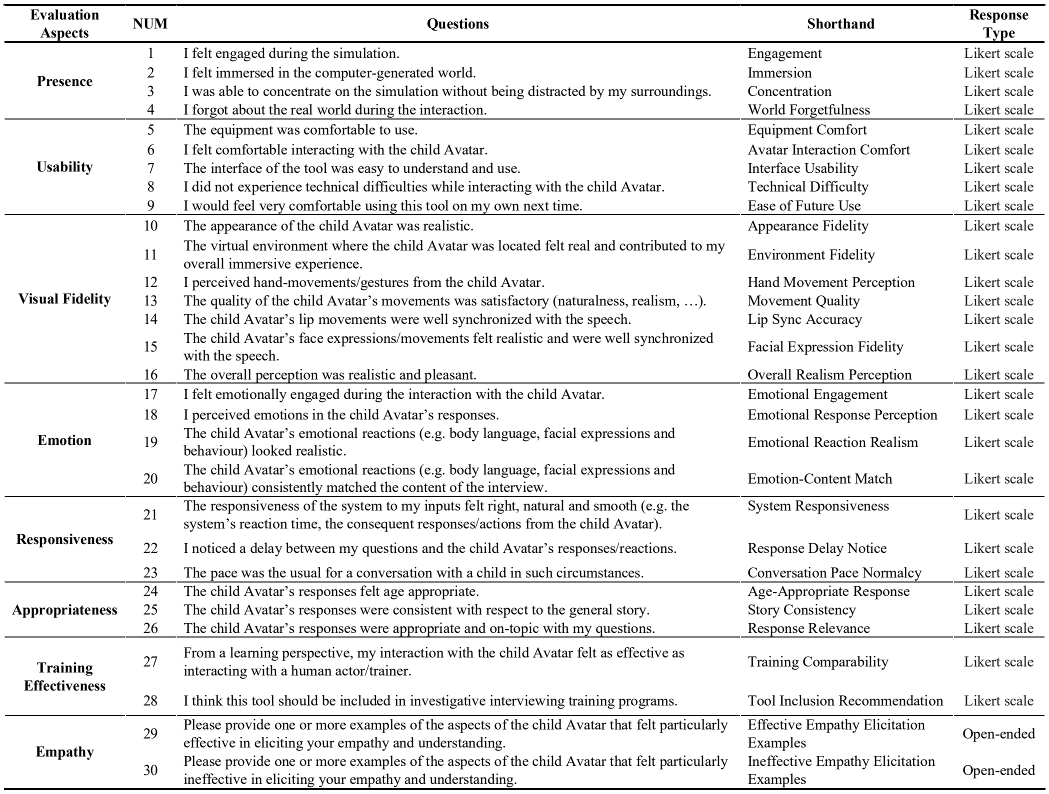
The questionnaire comprised 30 items, including 28 questions on a 5-point Likert scale and two open-ended questions. These questions were categorized into eight different evaluation keys: *presence* [16], *usability* [5], *visual Fidelity* [3, 43], *emotion* [3], *responsiveness* [33], *appropriateness* [45], *training effectiveness*, and *empathy*. Each question was designed to provide specific insights relevant to its category, ensuring comprehensive coverage of the user experience. For the full text of these questions, please refer to Figure 2.

5 RESULTS

In this section, we analyze the subjective feedback of participants for both 2D and 3D environments. The results are organized into three segments: Descriptive Analysis, Comparative Analysis, and Qualitative Feedback.

5.1 Descriptive Analysis

We began our analysis by calculating the central tendencies and dispersion of each evaluation aspect. In Figure 3, we present a visual comparison of mean scores and standard deviations between 2D and 3D environments. In the 2D environment, the *usability* scored the highest ($Mean$ = 4.09, SD = 0.67), and in the 3D environment, *presence* scored the highest ($Mean$ = 4.12, SD = 0.72).



Evaluation Aspects	NUM	Questions	Shorthand	Response Type
Presence	1	I felt engaged during the simulation.	Engagement	Likert scale
	2	I felt immersed in the computer-generated world.	Immersion	Likert scale
	3	I was able to concentrate on the simulation without being distracted by my surroundings.	Concentration	Likert scale
	4	I forgot about the real world during the interaction.	World Forgetfulness	Likert scale
Usability	5	The equipment was comfortable to use.	Equipment Comfort	Likert scale
	6	I felt comfortable interacting with the child Avatar.	Avatar Interaction Comfort	Likert scale
	7	The interface of the tool was easy to understand and use.	Interface Usability	Likert scale
	8	I did not experience technical difficulties while interacting with the child Avatar.	Technical Difficulty	Likert scale
	9	I would feel very comfortable using this tool on my own next time.	Ease of Future Use	Likert scale
Visual Fidelity	10	The appearance of the child Avatar was realistic.	Appearance Fidelity	Likert scale
	11	The virtual environment where the child Avatar was located felt real and contributed to my overall immersive experience.	Environment Fidelity	Likert scale
	12	I perceived hand-movements/gestures from the child Avatar.	Hand Movement Perception	Likert scale
	13	The quality of the child Avatar's movements was satisfactory (naturalness, realism, …).	Movement Quality	Likert scale
	14	The child Avatar's lip movements were well synchronized with the speech.	Lip Sync Accuracy	Likert scale
	15	The child Avatar's face expressions/movements felt realistic and were well synchronized with the speech.	Facial Expression Fidelity	Likert scale
	16	The overall perception was realistic and pleasant.	Overall Realism Perception	Likert scale
Emotion	17	I felt emotionally engaged during the interaction with the child Avatar.	Emotional Engagement	Likert scale
	18	I perceived emotions in the child Avatar's responses.	Emotional Response Perception	Likert scale
	19	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) looked realistic.	Emotional Reaction Realism	Likert scale
	20	The child Avatar's emotional reactions (e.g. body language, facial expressions and behaviour) consistently matched the content of the interview.	Emotion-Content Match	Likert scale
Responsiveness	21	The responsiveness of the system to my inputs felt right, natural and smooth (e.g. the system's reaction time, the consequent responses/actions from the child Avatar).	System Responsiveness	Likert scale
	22	I noticed a delay between my questions and the child Avatar's responses/reactions.	Response Delay Notice	Likert scale
	23	The pace was the usual for a conversation with a child in such circumstances.	Conversation Pace Normalcy	Likert scale
Appropriateness	24	The child Avatar's responses felt age appropriate.	Age-Appropriate Response	Likert scale
	25	The child Avatar's responses were consistent with respect to the general story.	Story Consistency	Likert scale
	26	The child Avatar's responses were appropriate and on-topic with my questions.	Response Relevance	Likert scale
Training Effectiveness	27	From a learning perspective, my interaction with the child Avatar felt as effective as interacting with a human actor/trainer.	Training Comparability	Likert scale
	28	I think this tool should be included in investigative interviewing training programs.	Tool Inclusion Recommendation	Likert scale
Empathy	29	Please provide one or more examples of the aspects of the child Avatar that felt particularly effective in eliciting your empathy and understanding.	Effective Empathy Elicitation Examples	Open-ended
	30	Please provide one or more examples of the aspects of the child Avatar that felt particularly ineffective in eliciting your empathy and understanding.	Ineffective Empathy Elicitation Examples	Open-ended

Figure 2: Categorization of questionnaire items according to Evaluation Aspects.

Conversely, the *responsiveness* scored the lowest in both the 2D and 3D environments, indicating concerns regarding interaction delays. Specifically, the mean score was 2.61 ($SD = 0.61$) in the 2D environment and 2.75 ($SD = 0.71$) in the 3D environment.

5.2 Comparative Analysis

In this section, we investigate the differential impact of the 2D versus 3D environment on QoE. This investigation utilized a paired sample T-test to analyze the variations between the two environments. Additionally, we calculated Cohen's d to assess the effect sizes. The results of these analyses are presented in Table 1.

In terms of *presence*, the results reveal that the 3D environment enhances the sense of presence more significantly than a 2D environment due to its sensory-rich and interactive dynamics. Specifically, participants experienced a higher sense of 'immersion', 'engagement' and 'world forgetfulness' in a 3D environment. However, 'concentration' levels did not show a significant difference, suggesting a complex relationship between immersive features and user focus. It is possible that the use of HMDs captures and diverts the participant's attention, thus impacting their concentration levels.

From the perspective of *usability*, this study claims that due to the absence of physically demanding equipment such as HMD, the 2D environment offers a more user-friendly experience. Specifically, participants experienced higher 'equipment comfort' in the 2D environment, suggesting that the absence of additional gear may contribute to increased comfort. Similarly, 'ease of future use' was

Table 1: Paired Sample T-Test and Cohen's d Effect Size for Learning Experience Variables. The gray areas show significant differences noted at p < .05.

	T-Statistic	P-Value	Cohen's d Effect Size
Engagement	2.798	0.008	0.466
Immersion	5.940	0.000	0.990
Concentration	2.023	0.050	0.337
World Forgetfulness	2.704	0.010	0.450
Equipment Comfort	-4.159	0.000	-0.693
Avatar Interaction Comfort	1.000	0.324	0.166
Interface Usability	-0.867	0.391	-0.144
Technical Difficulty	0.000	1.000	0.000
Ease of Future Use	-2.256	0.030	-0.376
Appearance Fidelity	1.190	0.242	0.198
Environment Fidelity	3.944	0.000	0.657
Hand Movement Perception	2.841	0.007	0.473
Movement Quality	1.847	0.073	0.307
Lip Sync Accuracy	0.849	0.401	0.141
Facial Expression Fidelity	2.142	0.039	0.357
Overall Realism Perception	1.661	0.105	0.276
Emotional Engagement	1.454	0.154	0.242
Emotion Response Perception	1.000	0.324	0.166
Emotional Reaction Realism	1.357	0.183	0.226
Emotion-Content Match	0.452	0.653	0.075
System Responsiveness	0.122	0.903	0.020
Response Delay Notice	1.183	0.244	0.197
Conversation Pace Normalcy	1.540	0.132	0.256
Age-Appropriate Response	1.152	0.257	0.192
Story Consistency	0.273	0.785	0.045
Response Relevance	-1.177	0.246	-0.196
Training Comparability	3.365	0.001	0.560
Tool Inclusion Recommendation	1.715	0.095	0.285

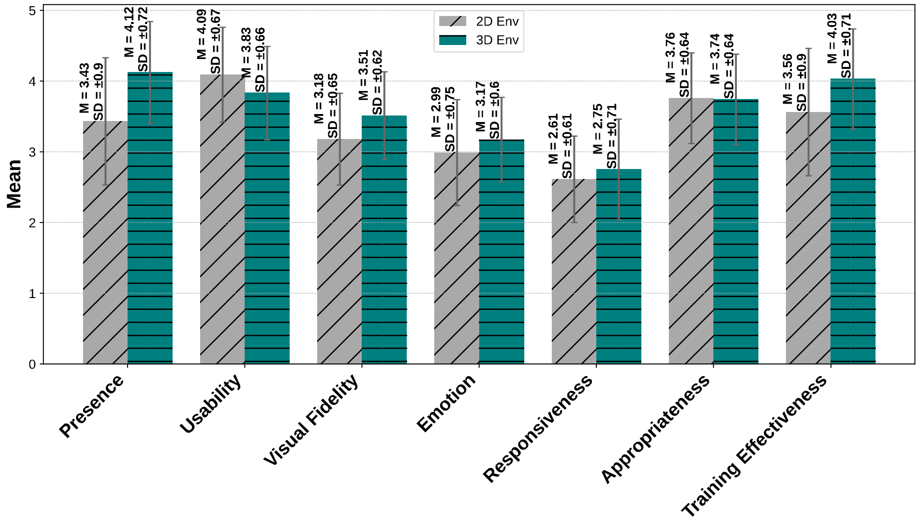


Figure 3: Bar-plot (95% confidence interval) of mean scores across the evaluative aspects in 2D and 3D environments.

significantly better in the 2D environment, which implies that users might prefer the 2D environment for future interactions, possibly due to its less physically exhausting nature. Conversely, 'avatar interaction comfort', 'interface usability', and 'technical difficulty' showed no significant differences, indicating that these aspects of *usability* are perceived similarly across both 2D and 3D environments.

Regarding *visual fidelity*, the study examined several key aspects, which indicated that there is an improved quality of visual experience in the 3D environment. In particular, 'environment fidelity' shows a substantial increase in 3D environment, suggesting that participants perceived a 3D environment as more lifelike. Similarly, 'hand movement perception' is significantly enhanced in 3D, indicating that the fluidity of VR technology provides a more lifelike representation of hand movement. Further, 'facial expression fidelity' is rated significantly higher in the 3D environment, reflecting that the immersive nature of the 3D environment may inherently strengthen the user's perception of facial expression naturalness.

However, 'movement quality', while not reaching statistical significance, shows a trend towards enhanced perception in 3D, implying a possible positive impact of VR technology on the naturalness and consistency of avatar movement. The 'overall realism perception' also trended towards a better experience in a 3D environment, but did not reach conventional significance levels. In addition, according to our expectations, there are no significant differences between 2D and 3D environments regarding 'avatar appearance fidelity' and 'lip sync accuracy'.

In terms of *emotion*, the study investigated how users perceive emotional dynamics with an avatar lacking emotional expressions in 2D and 3D environments. Contrary to the belief that a 3D immersive environment may compensate for the absence of emotional expressions [23, 37], the results indicate that there are no statistically significant differences between the 2D and 3D environments. This finding highlights the importance of explicit emotional signals in improving emotional dynamics during avatar-user interactions. However, a trend is observed to favor 'emotional engagement' and 'emotional reaction realism' in the 3D environment, although this trend is not statistically significant, which implies that immersion provided by the 3D environment leads the user to perceive certain behaviors and facial expressions of the avatar as an emotional reaction.

Regarding two conceptually close aspects, *responsiveness* and *appropriateness*, the results show that there is no significant difference in user experience between the 2D and 3D environments.

Furthermore, from the perspective of *training effectivenes*, we examined participants' perceptions of the training with the child avatar relative to human actors ('training comparability') and their approval of incorporating this tool into investigative interview training programs ('tool inclusion recommendation'). The results demonstrate a statistically significant preference for 3D over 2D in terms of 'training comparability', suggesting that participants found the 3D interaction to be nearly as effective as a real-life training experience. Additionally, there is a trend of preferring 'tool inclusion recommendation' in the 3D environment, although not statistically significant, suggesting that immersion provided by the 3D environment enhances the effectiveness of training by simulating real-world interactions more closely than in the 2D environment.

5.3 Qualitative Feedback

In the open-ended section of the questionnaire survey, participants were asked to clarify the aspects that contributed to the effective and ineffective elicitation of empathy within two simulated environments, as presented in Table 2. This table quantifies the feedback, offering a clear count of positive and negative comments for each evaluated aspect of the avatar's performance.

Notably, the 3D environment was dependent on its *visual fidelity*, with participants expressing mixed reviews, one stating: *'The hand*

Table 2: Distribution of Participant Comments on Effective and Ineffective Aspects of Empathy Elicitation in 2D and 3D Environments.

	Effective Aspects		Ineffective Aspect	
	2D	3D	2D	3D
Presence	0	1	2	0
Visual Fidelity	8	19	9	15
Emotion	2	0	4	5
Voice	6	7	4	5
Responsiveness	16	10	13	9
Response Speed	0	0	7	5

movements and eye contact were great!' In contrast, another participant stated: *'The avatar's body was not that of a child but more of a woman, which was distracting'*, underscoring the importance of appropriate visual cues in engendering empathy.

On the contrary, the participant responses indicated that the efficacy of the 2D environment was largely dependent on the avatar's *responsiveness*. One participant highlighted this as a strength, stating, *'When she spoke of harm perpetrated by her father, that tugged the heartstrings.'* Nonetheless, certain comments, like the one regarding the avatar's limited response of *'YEAH'*, underscore the necessity for enhanced emotional expressiveness and sophisticated interaction within virtual training platforms.

Furthermore, the participant feedback in both environments revealed aspects that hindered empathy engagement, indicating potential areas for improvement. A repeated critique related to the delay in the avatar's reactions; as one respondent explicitly observed, *'There was a large delay in her responses that made it hard to engage with the avatar and affected rapport building significantly.'* Emotional reaction was another aspect that received critical attention, with comments such as *'No emotion shown when disclosing sexual abuse'*, implying that more emotional expression could be crucial in augmenting the avatar's realism.

6 DISCUSSION AND FUTURE WORK

We explored the comparative effectiveness of 2D and 3D training environments in the context of CPS interview training. Differing from our earlier research [14], where VR's novelty significantly influenced participants, the current study provided a more comprehensive understanding of VR's practical utility in training scenarios. A noteworthy finding is the pronounced preference for the 3D environment in most of the questionnaire responses, suggesting that immersive experiences are important to further investigate in relation to improving training efficacy. This aligns with previous research suggesting that immersive learning contexts can substantially improve skill acquisition by providing more realistic and engaging experiences, as they closely mimic real-world interactions and scenarios [17, 20, 24]. This is demonstrated by a significant difference in 'facial expression fidelity', a crucial aspect of non-verbal communication, even without explicit emotion settings in the system. Participant feedback such as *'In the 2D version non-verbal behavior of the avatar was more difficult to perceive than in the VR version of the program'* supports this observation. Recognizing the importance of emotional dynamics in virtual environments, future iterations of the system will integrate Unreal Engine technology with NVIDIA Omniverse[1] to enhance emotional reactions. This integration is expected to result in a more sophisticated simulation of human-like emotional responses, thereby enriching the user's immersive experience by providing a nuanced and context-sensitive rendering of non-verbal cues. These advancements will likely contribute to the development of more empathetic and interactive virtual environments.

Another significant outcome is the advantage of visual interaction over text-based methods in developing effective communication skills. User comments like *'making eye contact but then also looking away when talking about negative moments (assault) seemed like a normal response in a human interview'* highlight the system's potential in simulating realistic human interactions. Interestingly, the immersive nature of the visual environment sometimes leads to overlooked responses, highlighting its realism. This was demonstrated in some participant's observations. For instance, one noted, *'When the child paused and then said the age of her brother very quietly. It seemed like a realistic moment'*, suggesting that in the visual environment, users think the delay is part of the child's hesitation. However, there is still a concern raised regarding the delay observed in the avatar's responses, primarily due to the current necessity for manual input to trigger these responses. Recognizing the criticality of fluid interaction in the context of child investigative interview interviews, future iterations will focus on automating the response mechanism, thereby eliminating the need for manual intervention.

Despite these advantages in VR technology, no significant differences were observed between the 2D and 3D environments regarding the 'tool inclusion recommendation'. This lack of distinction could be partly attributed to the discomfort associated with using HMDs, as indicated by participant comments such as *'Using headphones did not cause me pain, (...) with the VR I was in pain.'*

Additionally, it should be noted that the 3D environment requires additional equipment and a special setup, which is intrusive and uncomfortable for some individuals. In contrast, the 2D environment offers greater ease of use, which makes it ideal for 'anytime, anywhere' training sessions, making it a more accessible option for a broader audience.

7 CONCLUSION

This paper has presented a comprehensive analysis contrasting 2D and 3D virtual environments in their application to child interview training for CPS and law enforcement personnel. The findings reveal that the immersive 3D environment significantly enhanced aspects such as *presence*, *visual fidelity*, and *training effectiveness*, providing a more realistic and engaging interaction that greatly benefits training scenarios. Conversely, the 2D environment, while less immersive, was favored for its higher *usability* due to its fewer equipment requirements. However, this research also identifies critical challenges within these virtual training environments, particularly the need for emotional expressiveness and improved system responsiveness. Based on these findings, it is evident that future developments in virtual training tools must adopt a balanced approach. Prioritizing the refinement of the 3D environment is crucial, with a specific focus on integrating emotional expressiveness in avatars and improving avatar response speed. These improvements are essential for increasing the efficacy and accessibility of these virtual training environments for CPS and law enforcement personnel.

[1] https://www.nvidia.com/en-us/omniverse/apps/audio2face/

REFERENCES

[1] Gunn Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied cognitive psychology* 34, 3 (2020), 654–663.

[2] Gunn Astrid Baugerud, Ragnhild Klingenberg Røed, Helle BG Hansen, Julie Schøning Poulsen, and Miriam S Johnson. 2023. Evaluating child interviews conducted by child protective services workers and police investigators. *The British Journal of Social Work* (2023), bcac245.

[3] Frank Biocca and Mark R Levy. 2013. *Communication in the age of virtual reality.* Routledge.

[4] C.J. Brainerd and V.F. Reyna. 2012. Reliability of children's testimony in the era of developmental reversals. *Developmental review* 32, 3 (2012), 224–267.

[5] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Sonja P. Brubacher, Mairi S. Benson, Martine B. Powell, Jane Goodman-Delahunty, and Nina J. Westera. 2020. Chapter Twenty-Two - An overview of best practice investigative interviewing of child witnesses of sexual assault. In *Child Sexual Abuse*, India Bryce and Wayne Petherick (Eds.). Academic Press, 445–466.

[8] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. 2000. Investigative interviews of child witnesses in Sweden. *Child abuse & neglect* 24, 10 (2000), 1355–1361.

[9] Kevin Charles Dalli. 2021. Technological Acceptance of an Avatar Based Interview Training Application: The development and technological acceptance study of the AvBIT application. *Digitala Vetenskapliga Arkivet* (2021).

[10] Alex Dan and Miriam Reiner. 2017. EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays. *International Journal of Psychophysiology* 122 (2017), 75–84.

[11] Mark D. Everson and Jose Miguel Sandoval. 2011. Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect* 35, 4 (2011), 287–298.

[12] Chris Fowler. 2015. Virtual reality and learning: Where is the pedagogy? *British journal of educational technology* 46, 2 (2015), 412–422.

[13] Stephanie G Fussell and Dothang Truong. 2023. Accepting virtual reality for dynamic learning: An extension of the technology acceptance model. *Interactive Learning Environments* 31, 9 (2023), 5442–5459.

[14] Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. 2023. Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments. *Scientific Reports* 13, 1 (2023), 20403.

[15] Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children. In *Proceedings of the ACM Workshop on Games Systems (GameSys)* (Athlone, Ireland). 9–15.

[16] Igroup Presence Questionnaire (IPQ) Overview. . A Multi-Disciplinary Project Consortium Addressing New Interfaces Between Humans and the Real and Virtual Environment. Available at: http://www.igroup.org/pq/ipq/index.php.

[17] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 23 (2018), 1515–1529.

[18] Miriam Johnson, Svein Magnussen, Christian Thoresen, Kyrre Lønnum, Lisa Victoria Burrell, and Annika Melinder. 2015. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Applied Cognitive Psychology* 29, 5 (2015), 661–668.

[19] Maria Kandaurova and Seung Hwan Mark Lee. 2019. The effects of Virtual Reality (VR) on charitable giving: The role of empathy, guilt, responsibility, and social exclusion. *Journal of Business Research* 100 (2019), 571–580.

[20] Eric Krokos, Catherine Plaisant, and Amitabh Varshney. 2019. Virtual memory palaces: immersion aids recall. *Virtual reality* 23, 1 (2019), 1–15.

[21] Kuo-Wei Kyle Lai and Hao-Jan Howard Chen. 2021. A comparative study on the effects of a VR and PC visual novel game on vocabulary learning. *Computer Assisted Language Learning* 0, 0 (2021), 1–34.

[22] J Madden, S Pandita, JP Schuldt, B Kim, A S. Won, and NG Holmes. 2020. Ready student one: Exploring the predictors of student learning in virtual reality. *PloS one* 15, 3 (2020), e0229788.

[23] Valentina Mancuso, Francesca Bruni, Chiara Stramba-Badiale, Giuseppe Riva, Pietro Cipresso, and Elisa Pedroli. 2023. How do emotions elicited in virtual reality affect our memory? A systematic review. *Computers in Human Behavior* (2023), 107812.

[24] Thomas K Metzinger. 2018. Why is virtual reality interesting for philosophers? *Frontiers in Robotics and AI* 5 (2018), 101.

[25] Kathy A Mills and Alinta Brown. 2022. Immersive virtual reality (VR) for digital media making: transmediation is key. *Learning, Media and Technology* 47, 2 (2022), 179–200.

[26] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law* 21, 1 (2015), 28–52.

[27] Martine B Powell, Ronald P Fisher, and Carolyn H Hughes-Scholes. 2008. The effect of intra-versus post-interview feedback during simulated practice interviews about child abuse. *Child Abuse & Neglect* 32, 2 (2008), 213–227.

[28] Martine B Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26, 4 (2016), 365–374.

[29] Jaziar Radianti, Tim A Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. 2020. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education* 147 (2020), 103778.

[30] Fabin Rasheed, Prasad Onkar, and Marisha Narula. 2015. Immersive virtual reality to enhance the spatial awareness of students. In *Proceedings of the 7th Indian Conference on Human-Computer Interaction.* 154–160.

[31] Pegah Salehi, Syed Zohaib Hassan, Myrthe Lammerse, Saeed Shafiee Sabet, Ingvild Riiser, Ragnhild Klingenberg Røed, Miriam S Johnson, Vajira Thambawita, Steven A Hicks, Martine Powell, et al. 2022. Synthesizing a talking child avatar to train interviewers working with maltreated children. *Big Data and Cognitive Computing* 6, 2 (2022), 62.

[32] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A. Riegler. 2022. Is More Realistic Better? A Comparison of Game Engine and GAN-Based Avatars for Investigative Interviews of Children. In *Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR)* (Newark, NJ, USA). 41–49.

[33] Steven Schmidt. 2022. *Assessing the quality of experience of cloud gaming services.* Springer Nature.

[34] Donghee Shin. 2018. Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in human behavior* 78 (2018), 64–73.

[35] Semyon M Slobounov, William Ray, Brian Johnson, Elena Slobounov, and Karl M Newell. 2015. Modulation of cortical activity in 2D versus 3D virtual reality environments: an EEG study. *International Journal of Psychophysiology* 95, 3 (2015), 254–260.

[36] Marije Stoltenborgh, Marian J. Bakermans-Kranenburg, Marinus H. van IJzendoorn, and Lenneke R. Alink. 2013. Cultural–geographical differences in the occurrence of child physical abuse? A meta-analysis of global prevalence. *International Journal of Psychology* 48, 2 (2013), 81–94.

[37] Feng Tian, Minlei Hua, Wenrui Zhang, Yingjie Li, and Xiaoli Yang. 2021. Emotional arousal in 2D versus 3D virtual reality environments. *PloS one* 16, 9 (2021), e0256211.

[38] Feng Tian, Xuefei Wang, Wanqiu Cheng, Mingxuan Lee, and Yuanyuan Jin. 2022. A Comparative Study on the Temporal Effects of 2D and VR Emotional Arousal. *Sensors* 22, 21 (2022), 8491.

[39] UMA Steering Group. 2022. UMA Repository. https://github.com/umasteeringgroup/UMA. Online repository.

[40] Unity Asset Store. 2022. SALSA LipSync Suite. https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442. Software package.

[41] Pınar Üstel, Matthew J Smith, Shannon Blajeski, Jeffery M Johnson, Valerie G Butler, Johanna Nicolia-Adkins, Monica J Ortquist, Lisa A Razzano, and Adrienne Lapidos. 2021. Acceptability and feasibility of peer specialist-delivered virtual reality job interview training for individuals with serious mental illness: A qualitative study. *Journal of technology in human services* 39, 3 (2021), 219–231.

[42] Wendy A Walsh, Lisa M Jones, Theodore P Cross, and Tonya Lippert. 2010. Prosecuting child sexual abuse: The importance of evidence type. *Crime & Delinquency* 56, 3 (2010), 436–454.

[43] Erin Wilson, David G Hewett, Brian C Jolly, Sarah Janssens, and Michael M Beckmann. 2018. Is that realistic? The development of a realism assessment questionnaire and its application in appraising three simulators for a gynaecology procedure. *Advances in Simulation* 3, 1 (2018), 1–7.

[44] Misun Yi, Eunkyung Jo, and Michael E Lamb. 2016. Effects of the NICHD protocol training on child investigative interview quality in Korean police officers. *Journal of police and criminal psychology* 31, 2 (2016), 155–163.

[45] Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. 2018. Retrieval-enhanced adversarial training for neural response generation. *arXiv preprint arXiv:1809.04276* (2018).

A3Cplus - Efficient Anatomically Accurate Avatar Creation

Johannes Günter Herforth
University of Luxembourg
Esch-sur-Alzette, Luxembourg
johannes.herforth@uni.lu

Jean Botev
University of Luxembourg
Esch-sur-Alzette, Luxembourg
jean.botev@uni.lu

ABSTRACT

Virtual reality applications are witnessing increased adoption in mental and physical health fields, from rehabilitation therapies to psychological studies. The more advanced the application, the greater the demand to incorporate realistic, custom avatars based on the participant's physical characteristics to enhance embodiment. Current solutions focus on creating such avatars by using expensive camera arrays to capture a 3D representation, which requires technical skills and actively involves the participant in the process. However, equipment and space requirements, setup complexity for non-technical operators, and physical challenges for participants often lead to difficulties and high costs for consistent adherence. This paper presents A3Cplus, a tool to efficiently generate anatomically accurate avatars based solely on a small amount of participant phenotypic data. An optimized processing pipeline uses this data to manipulate specialized blend shapes automatically and mold a generic model into the correct dimensions. We provide illustrative examples of using our tool and discuss its general applicability to immersive avatar-based virtual environments that require a high degree of accuracy and embodiment.

CCS CONCEPTS

• **Computing methodologies** → **Shape modeling**; **Virtual reality**; • **Human-centered computing** → *Accessibility systems and tools*; • **Applied computing** → *Psychology*.

KEYWORDS

Avatar Creation, Blend Shapes, Clinical and Therapeutic Applications, Virtual Reality, Immersion

ACM Reference Format:
Johannes Günter Herforth and Jean Botev. 2024. A3Cplus - Efficient Anatomically Accurate Avatar Creation. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652215

1 INTRODUCTION

With the increasing availability and capabilities of Virtual Reality (VR) technologies, their significance has surged, ushering in new possibilities for immersive experiences and research. The unique properties of VR systems that enhance these experiences are best showcased when we can control the entire environment and provide a full-body immersion that engages the participant's body within a virtual environment. We are able to view objects and environments from multiple perspectives, observe changes in real time, and visualize how they would appear with alterations. The capability enables us to gain insights into aspects that are typically challenging to perceive directly. More formally, a full-body immersion within a virtual environment allows the user to experience a sense of embodiment [5], which combines the feelings of being present, owning their physical form, and exerting agency control over it. By optimizing these three elements, more convincing body illusions that enhance exploration and interaction with the virtual world as if one is truly present can be established.

These illusions are particularly useful when considering research combining VR and clinical psychological and physical treatments. VR has already been successfully applied across a wide range of issues, including pain management [9], anxiety [8] rehabilitation of gait [2], and even learning to use new tools like wheelchairs [11]. While these studies have provided valuable insights into the effectiveness of these novel treatments, they are almost exclusively conducted in isolation from a technical front, making it challenging to compare results across various studies.

According to a meta-analysis published in 2018 [3], several obstacles remain in VR applications for clinical settings and psychological experiments. These challenges include limited trials in non-lab settings, the technical expertise required, lack of standardization across studies, and overall costs. As these applications of VR are still in their infancy, it is not difficult to see that there are no trivial problems to solve. In fact, a further analysis from 2022 [4] confirms that many of the same challenges continue to persist in this field.

A major aspect that poses a significant barrier to entry for an effective full-body illusion, is the requirement for a realistic 3D avatar. Studies have demonstrated that realistic avatars can improve the sense of body ownership [12] and enhance the subjective experience for the participant [6]. For an effective experience, customizing the avatar to align with body shape, dimensions, and other phenotypic properties is crucial for achieving a high-quality embodiment. Unfortunately, the most commonly used solution is high-quality scanner-like systems, which introduce a substantial ongoing cost for the end user. These costs are associated with obtaining and maintaining the scanning equipment, in addition to the ongoing learning process. Furthermore, unique challenges also present themselves due to the highly sensitive nature of the data in a clinical or psychological application. These contexts demand strict adherence to privacy and data protection regulations to protect patient information. Another challenge is that not all users may have the ability to enter the machine to get an optimal scan, limiting their ability to receive treatment. Looking beyond stationary care, chronic therapies also pose a challenge when the health of the

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652215

body recovers (or worsens) at a faster rate, requiring more frequent rescan sessions to keep up with the progress.

These problems can be summarized by three requirements for this problem space. The first factor is the need for a *dynamic and consistent* system that can adapt to changes in the user's appearance over time. This includes the ability to accommodate alterations such as aging, body fitness or new features such as scars or a tan. These changes should also be reversible and applicable live within the VR application, enabling spontaneous, on-the-spot adjustments to maximize embodiment in as many situations as possible.

Secondly, we need to focus on *inclusive* solutions that cater to all participants involved in the process. This includes designing user-friendly systems for operators, requiring minimal cost, space, setup, maintenance, and reusability. Additionally, it should be optimized for accessibility for the users, accommodating individuals with physical impairments or forms of anxiety such as claustrophobia. Furthermore, it should be inclusive to developers of VR applications, allowing them to easily integrate it into their own projects, ensuring interoperability, compatibility, and consistency across different therapies and experiments.

Lastly, we need to focus on user *privacy* by minimizing the use of sensitive data for avatar creation. While creating the most accurate 3D representations is ideal, this usually requires using highly sensitive personal information, such as body scans or images. Doing so adds an extreme burden on operators to enact additional data management, security, and dealings with liability and potentially online 3rd parties. Scanning processes also affect participants by requiring them to wear specialized slim-fit clothes, which can be embarrassing and uncomfortable in front of others. It is crucial for users to have access to realistic avatars without sacrificing their privacy. Otherwise, it may deter and prevent them from receiving the treatments they need.

This paper presents A3Cplus, aimed at creating a dynamic, inclusive, and privacy-preserving avatar creation process. In Section 2, we provide an overview of existing tools used to create humanoid models, including commercial options and an in-depth analysis of two representative academic papers on their strengths and weaknesses. Throughout Section 3, we discuss in more detail the A3Cplus software, key features, and rationales that inspired its creation. Additionally, we introduce the process workflow, demonstrating how each feature contributes to simplifying its usage for the operator. We conclude the paper by discussing the broader implications of our tool along with areas for improvement, emphasizing the significance of general accessibility and applicability to immersive avatar-based virtual environments.

2 3D CHARACTER CREATION

This section examines various commercial and open-source tools on the market, and academic research conducted concerning 3D character creation, discussing their features to provide the relevant background behind the development of A3Cplus.

2.1 Commercial and Open-Source Software

Character creation tools are not limited to realistic avatars designed specifically for VR applications. Many established software packages offer robust solutions for character design in various contexts, including video games as prominent examples. Avatar-based video game genres, such as role-playing games, provide a vast array of character creation options for users to customize their characters. However, if these models are not limited to a single game, they can only be used in a narrow ecosystem of other games. Therefore, we limit the scope of software to those capable of integrating with any application by, e.g., exporting to 3D file formats. Overall, the software availability can be differentiated between *manual* and *automatic* creators.

A manual avatar creator is a software package focused on enhancing the experience for designers to configure different options to do with avatar creation manually. Examples of such applications include *Reallusion Character Creator* [1], *DAZ3D* [2] and *MakeHuman* [3]. All are intended to be used throughout the application development cycle before the application is shipped to the customer. This is because the tools have a high learning curve and have features to automatically import them into various game engine editors or export them to various 3D model data types.

While creating realistic, fully rigged, clothed, and textured avatars using these tools is simple, it is challenging to customize them based on accurate phenotypic measurements. However, this is not the goal of most of these tools, which were created for designers to make good-looking assets that fit into the specific design language of their applications.

The largest benefit to using manual avatar creators is that they operate offline and do not require any sensitive data to function, which helps maintain the user's privacy. As anatomical data cannot be incorporated due to missing measures, designers must rely on their artistic skills and knowledge of human anatomy to create a realistic avatar. This shows that while they are very inclusive to users, application developers, and operators who must prepare everything in advance, they do not provide a consistent and easy-to-use system. Looking at the output's dynamics, each application can export a functional humanoid rig. However, they each have their differences in the types of blend shapes that are available to export. DAZ3D stands out as the most dynamic manual creator as it enables the operator to export any available blend shape, with the added benefit of adding the body blend shape to attached clothes. Reallusion also allows for the export of blend shapes, but only when they pertain to facial expressions or are custom-made. However, the custom blend shapes only work correctly within the Reallusion editor itself, limiting their purpose outside the program. MakeHuman does not export any blend shapes at all, providing the least dynamic avatar export.

Automatic creators, on the other hand, rely on input images and apply algorithms to generate an avatar that can be used similarly to the manually generated avatar. Examples of such applications include *Meshcapade Me* [4] and *RealityScan* [5]. Both tools can produce lifelike representations of a human subject given a series of input images. Meshcapade Me is based on the skinned multi-person linear model (SMPL) [7], which operates via adjusting pose blend

[1] https://www.reallusion.com/character-creator/
[2] https://www.daz3d.com/
[3] http://www.makehumancommunity.org/
[4] https://www.meshcapade.com/
[5] https://www.unrealengine.com/en-US/realityscan/

shapes learned from thousands of 3D body scans. The user interface guides the operator through the process, allowing additional manual adjustments via sliders altering different phenotypic body measurements. RealityScan, like many other applications, applies photogrammetry techniques to images captured on a mobile phone to generate an output mesh automatically. This output mesh can then be used with services such as Mixamo [6] to automatically generate a rig for the human mesh. Meshcapade Me's avatars offer the greatest flexibility among the commercial alternatives due to using blend shapes as their building block. These shapes can be utilized throughout the model, providing dynamic animations inside the VR application. Note that these blend shapes may not have a clear definition from a human perspective as they were trained on data rather than being created by artists. On the other hand, applications such as RealityScan offer simpler functionality by providing an exact mesh that can be rigged for animation. They cannot dynamically alter body shapes without requiring manual creation using separate tools for each participant.

When looking at inclusivity from an operator's perspective, these tools provide a process of skipping the designing stage and going straight for reality. Still, the process on either application introduces opportunities for human error and requirements to rescan if something goes wrong. In addition, since there is no straightforward method for importing the generated models directly into VR applications, they must either be manually imported into an editor or imported fully at runtime. In terms of privacy, both programs require users to sign up for accounts and upload their input images for the service's hardware to apply the relevant algorithms to generate an output mesh. This process may raise concerns about data security and potential misuse of sensitive information. This feature alone makes it problematic or even impossible to include these options in clinical applications.

Overall, commercial applications create realistic-looking avatars and offer various customization options to adapt the avatar to the user. They typically export their models to common formats such as FBX, glTF, or obj, necessitating careful data management by the user and requiring developers to implement this functionality in their programs. However, these tools' significant limitations are either the absence of exact body dimension measurements or reliance on third-party platforms for generating anatomically accurate avatars.

2.2 Scientific Projects

3D character modeling encompasses a wide range of disciplines and technologies to achieve high-quality results.

One of the projects best resembling the goals of our study aims at creating a "Virtual Caliper" [10] for measuring the correct body dimensions using an HTC Vive headset and two SteamVR Lighthouse base stations. The authors present an application that allows for the creation of metrically accurate body shapes by utilizing the VR controllers as measuring points rather than relying on physically based methods. They employed user studies to identify the most effective measurement points for accurately capturing body dimensions in VR environments. By reducing the number of optimal measurements, they further refined the results by optimizing SMPL-based regressors based on these measures, ultimately landing on a few that became their user input in their process.

The result entails a process engaging the participant in following a guided walkthrough in VR. It is important to note that the tasks are entirely delegated to the participant rather than the experimenter. The measurements gathered from the six placements, as well as weight, are fed into custom SMPL linear regressors using least-squared computation to produce an avatar.

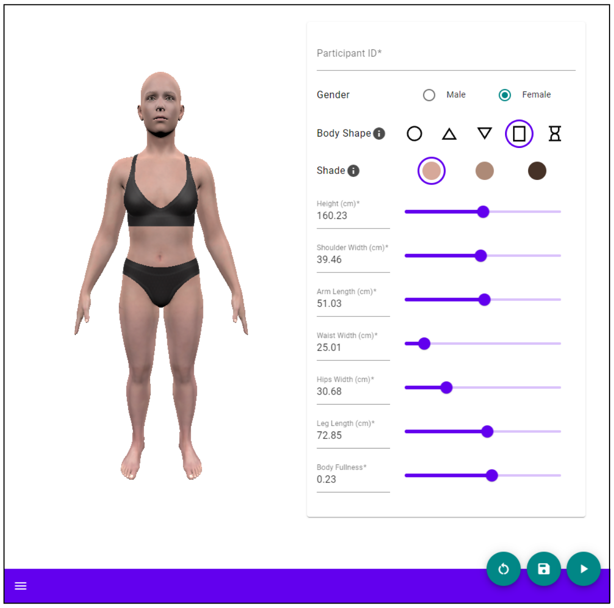
The study presents a model generation process that offers fast, guided, accurate, and privacy-aware body dimension measurement within VR environments. Although the resulting anatomy appears visually plausible based on input measurements, it lacks methods for adding textures such as skin color, clothing, or facial features. The application's reliance on the HTC Vive system limits its compatibility with other controllers and necessitates porting to different systems. Additionally, the user-guided process only applies to users who can stand up and go through the process, which may not be feasible for all individuals. The study showcases a desktop tool that allows for adjusting and exporting the model into FBX format, circumventing the virtual caliper process. Given that their results show subpar performance with real-life measurements and require post-processing adjustments to make the model more plausible, it indicates that it may not be well-suited for this specific task.

Another related study [1] focuses primarily on advancements in reconstructing human meshes out of a generic base model. To customize the base avatar model, the user must go through two separate 3D scanning steps. The first step consists of 40 DSLR cameras which capture the full body from a standing A-Position. In the second phase, a setup of 8 DSLR cameras captures the user from a sitting position, resulting in consistent facial scan information. For each scan, a point cloud is generated, where the goal is to align the base model with the new point set. After manually selecting nine landmarks on both the scan and the base model to wrap it around the point cloud.

The objective is to position the base model within the generated point cloud by automatically aligning nine key points from the base model to the scan. This is completed via a pose-optimizing pipeline, refining closest point correspondences and performing a fine-scale deformation to the initial point set. Textures are then computed based on camera images and refined or adjusted according to the presence of artifacts and the effectiveness of capturing details in unseen regions like under the arms. The facial reconstruction pipeline is similar. However, more features are transferred from the base model, such as its facial blend shapes. In addition, specific facial details such as teeth and eyes are retained from the original textures.

The finished product consists of the base mesh fitted and optimized to the structure given in the scan's point cloud. The authors showcase the flexibility of their approach by demonstrating how the textures of the newly created avatar can be effortlessly swapped between different scans having undergone the same treatment. This allows visual modifications to be even faster if only the texture needs to be altered. Since the procedure involved minimal manual intervention, consisting of selecting reference points and transferring images, the authors assert that it can be completed rapidly within 10 minutes.

[6] https://www.mixamo.com/



Participant ID*

Gender ○ Male ● Female

Body Shape ○ △ ▽ [✓ ▯] ⧗

Shade [✓ ●] ● ●

Height (cm)* 160.23

Shoulder Width (cm)* 39.46

Arm Length (cm)* 51.03

Waist Width (cm)* 25.01

Hips Width (cm)* 30.68

Leg Length (cm)* 72.85

Body Fullness* 0.23

Figure 1: A3Cplus user interface.

Overall, both papers present distinct methods for generating realistic-looking avatars using unique and specialized techniques. Since both rely on base avatar models as their foundation, they are both very dynamic for use in VR applications. Despite this, they share limitations that may affect their inclusivity, such as reliance on specific hardware, substantial space requirements, and an assumption of a fully capable user to participate in the avatar creation process.

3 THE A3CPLUS AVATAR CREATOR

A3Cplus is a tool for efficiently creating anatomically accurate avatars designed to mitigate the limitations of existing solutions, as discussed in Section 2. We placed particular emphasis on simplifying the process and making it as non-technical as possible.

3.1 Tool Architecture

To achieve our goal of facilitating the avatar creation process, we opted for the *Keep it simple, Stupid!* (KISS) and *What you see is what you get* (WYSIWYG) philosophies. The KISS principle promotes developing simple systems as they work best in contrast to more complex systems. In the context of our tool, this means keeping the number of variables low, letting the operator see the entire program in one window without any drop-down menus or hidden options. On the other hand, WYSIWYG techniques center on allowing the resulting output to be seen directly within the editing window. Within our tool, this provides a sanity check for operators and a guarantee that the output model resembles the participant. Combining both, we can ensure a straightforward and intuitive user experience in our user interface, as shown in Figure 1.

The user interface is organized between two sections, featuring a spacious and bifurcated layout consisting of an output view and a controller view.

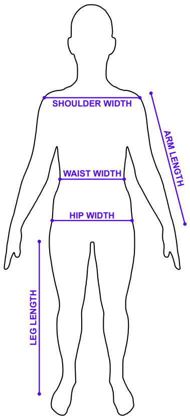


Figure 2: Main body measurements.

On the left side, the output view displays a live 3D-rendered scene with a uniformly lit humanoid model we use to project our measurements onto. The operator is free to adjust the position and rotation of the camera view to inspect the model from all perspectives to ensure it is correct. While the operator may interact and change the view on the model, they cannot alter any settings to do with changing the model directly.

The right side provides the operator an interactive space to seamlessly manipulate parameters that affect the model in real time on the output view. Here, the individual measurements can be applied, and non-visual metadata, such as participant ID, can be altered.

Combining both sides, the operator is guided linearly from the top to the bottom of the interface, ensuring a chronological and straightforward interaction from input to output.

3.2 Base Avatar

Due to the many advantages of the model outlined in Section 2, we chose to utilize DAZ3D's Genesis 9 model as a starting point for avatar creation. Compared to other options, by allowing full blend shape export and automatic generation of blend shapes for clothes, the avatars allow for simply and accurately incorporating extra features. While we utilized the available model for our initial implementation, our tool can be easily adapted to work with other base models. Utilizing a custom avatar could yield even more benefits and tailored results.

3.3 Blend Shapes and Optimization

Rather than wrapping the base model around a scan or basing the output on generalized blend shapes, we focus on modifying pre-designed, highly specialized blend shapes that carry human meaning, such as "Hip Size" or "Leg Length". This approach offers two advantages. First, by constraining changes to shapes that are plausible as human bodies, we ensure that our models remain realistic and grounded in human anatomy. Second, this enables developers to easily incorporate these shapes into their applications, allowing for live experiences where body shape can change dynamically.

To measure and adjust blend shape values based on phenotypic values, we first established methods of measuring the points of interest using our base model. Our aim was to create intuitive, systematic measurement processes that are easy for operators to understand, apply to a user, and implement into their procedures. We accomplished this by using the height and identifying key points of interest on the base mesh, labeled as indicated in Figure 2. These measurements represent specific aspects of human anatomy and could be easily measured by operators and generalized into combinations of blend shapes to represent an accurate body shape. The simplicity of the measurement process enables accessibility options as an avatar can be created while users are lying down, providing avatar options that include users with physical limitations.

With the measurements established and obtained, they are transferred to the base model. Some measurements, such as shoulder, leg, and arm length, are relatively simple to adjust since they operate independently of other measurements and can be achieved by utilizing blend shapes that proportionally adjust the relevant body parts. By making minute adjustments to these blend shapes, we can ensure that the measurements between the key points reflect the desired proportions.

The remaining measurements are more complex, as they involve working with multiple blend shapes simultaneously. As leg length is directly related to the height of the avatar, adjusting the height parameter only affects the upper body by stretching it until the height is reached, ensuring that each constraint is met. If both height and leg length are not set correctly, this can result in an unnatural body structure. On the other hand, when given valid values (where height is at least greater than leg length), the tool produces accurate and visually appealing height proportions.

The challenge becomes more complex when dealing with the body core. Humans come in various shapes and sizes, making it challenging to create a single set of blend shapes that accurately represents all bodies. To address this issue, we leverage the principles of *body shapes*, which are well-established in fashion design, as a starting reference to optimize clothes to a specific type of body shape. Adapting this concept to our models allows us to tailor how we utilize blend shapes based on the input body dimensions.

Using the measurements from Figure 1 and adapting the body shape parameter results in the differences shown in Figure 3. As only the body shape differs, the change leads to a more substantial contrast in the chest region, accompanied by a smaller variation in the pelvic area due to the reduced span between the waist and hips. For instance, the avatar in Figure 3b shows how the chest region narrows, whereas the pelvic area widens, leading to its triangular shape. Comparing the hourglass avatar in Figure 3e to the rectangular shape in Figure 3d, the hourglass-shaped avatar possesses both a wider chest and pelvic area, creating a distinct sharp and straight gradient extending from the waist. The optimizations for each body shape type are designed by utilizing blend shapes that impact the entire structure of the core body, ranging from the chest to the hips and thighs.

There are also differences in weight and fitness between various body shapes, with weight gain or loss having unique effects depending on the general body shape.

Individuals with a circular body shape will experience more weight gain around their waist. In contrast, those with a triangular body shape will notice greater increases in their lower waist and hip area. These differences are reflected by incorporating a final fullness slider that adjusts the body shape, ranging from a more fit representation to a more curved and full-body type. This enables operators to create models that accurately reflect the individual's desired appearance based on their specific body shape and weight distribution.

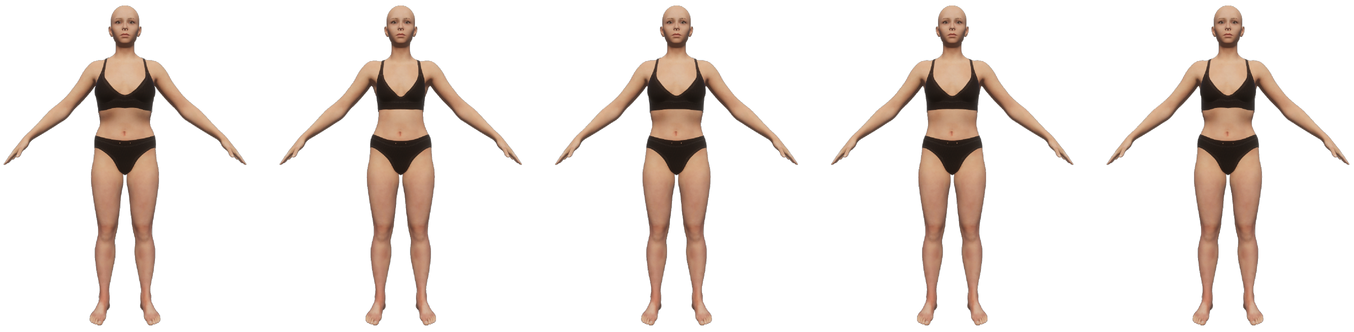
A3Cplus offers comprehensive customization options for the most prevalent body shapes, including circle, triangle, inverted triangle, rectangle, and hourglass. These shapes can be selected through a shape selection interface, as shown in Figure 1. In cases where an operator is unsure which body shape to select, they can cycle through the options until they find a visually suiting match.

3.4 Model Export

A3Cplus's export process is another major aspect, setting it apart from the existing tools. Conventionally, as explored in Section 2, tools require an export of the entire 3D model to a file to then import them into another application. While it is possible to export a complete model with applied blend shapes to the GL Transmission Format (glTF) file type in A3Cplus, it is not the recommended method as it requires the export of large files and an overall fragmented user experience. As the base model is generic and already contains all the potential positions that can be generated in the tool, exporting entire models would be space inefficient and more challenging to deal with in other software. Since the base model is designed to handle a wide range of blend shape possibilities, exporting the entire model is unnecessary or inefficient when only blend shape values need to be modified.

Since we are only modifying blend shape values, we can avoid the model exporting process altogether and save the blend shape values in more widely used interchange formats like JSON. This approach also makes it easier for users to work with the output in their 3D applications, as they can import and optimize the same model as seen in our tool into their editors. Editing their environments using the same model enables the developer to test their 3D environment with any potential blend shape combinations, guaranteeing that their program will always work. In addition, game engines such as Unity or Unreal do not support importing model files during runtime natively as their importers are part of their Editor code base, having to then rely on 3rd party importers.

Building on this point, exporting and manually managing the file still adds friction and the potential for human errors. As we simply want to pass the values into a different program, we do not have the requirement of storing any of the resulting outputs. To streamline the blend shape value use with A3Cplus, we developed a feature that enables the direct transfer of blend shape values into a custom target binary. This eliminates the need for manual file management, significantly reducing the potential for human error and further enhancing the user experience.



(a) Circular. **(b) Triangular.** **(c) Inverted triangular.** **(d) Rectangular.** **(e) Hourglass.**

Figure 3: A3Cplus output examples of different body shape types based on identical measurements.

4 DISCUSSION AND FUTURE WORK

The proposed tool's novel structural components leverage an artistically designed off-the-shelf base model to efficiently create anatomically accurate avatars based on phenotypic measurements.

A3Cplus fulfills the initial requirement by being capable of dynamically adapting avatars to the user's body in real-time, both during creation and runtime. In addition to automatic rigging, it enables the possibility of making real-time adjustments to its blend shapes in the editor and within game engines. Results are deterministic within a specified tolerance, i.e., by inserting the same input values, the output model remains consistent throughout each run.

A3Cplus also fulfills the inclusivity requirement for all involved parties, from developer to operator and user. Firstly, by constituting a low cost and a minimal space requirement, the tool provides a low barrier to entry for realistic full-body illusions. Furthermore, we have shown its advantages in development by incorporating an easy-to-use interface and simple connectors into game engines using command-line arguments. This makes it effortless for developers to create applications that require full-body illusions. The intuitive user interface simplifies the process of creating avatars, allowing users to customize their full-body illusions easily in a matter of seconds. Whether the user is used to the tool or just starting out, A3Cplus makes creating realistic and engaging full-body illusions for various applications effortless. Lastly, by prioritizing accessibility, our tool accommodates individuals who would not be able to take part in active scanning processes due to physical limitations or anxiety around scanners or confined spaces, providing full-body illusions accessible to anyone.

Regarding the final privacy requirement, we prioritized minimizing data use throughout the entire life cycle of the process. We mitigate contemporary privacy concerns by not utilizing photographic information and external online platforms for our tool's operation. If the final VR application does not store blend shape values, our tool allows privacy-conscious individuals to utilize it directly without disclosing their body measurement data to the operator, as the tool does not store but passes on the blend shape values to the application.

Although A3Cplus already aligns with our introduced primary requirements, several areas remain that require further attention, particularly for aspects such as the quality of the output. In its current state, the tool operates under various assumptions that do not accurately reflect reality, such as generating symmetrical bodies and focusing uniquely on adults. Additionally, currently predefined skin textures are used, i.e., visual features such as differing pigments, scars, and birthmarks are not accounted for, which can influence realism.

Introducing these elements into the user interface is complex, as it could overwhelm an average user. Another limitation is the inability to modify the head or facial features of the avatar, restricting its applicability to an egocentric perspective where users cannot view themselves in a mirror above the neckline. Lastly, as our blend shape modifications are primarily based on common artistic or fashion-based interpretations rather than scientific data, we cannot be certain that every type of body shape is represented authentically. The limitations conflict with creating a simple user experience to create avatars for full-body illusions. We will address these issues in future work and explore further user interface options that allow more complex behavior to be seamlessly integrated.

5 CONCLUSION

This paper presents A3Cplus, a tool for efficiently creating anatomically accurate avatars for clinical and therapeutic VR applications and avatar-based immersive applications in general. In particular, A3Cplus streamlines the creation process and provides an easy-to-use, secure, and offline workflow that helps generate realistic avatars without the need for complex scans and measurements. During the process, a base avatar is adapted with blend shapes by phenotypic measurements with a set of fundamental body types. The results can be easily exported and integrated with other tools, engines, or content creation pipelines.

A3Cplus is already employed in ongoing VR-based psychological studies to help experimenters quickly create realistic, morphable participant representations. Although the tool already satisfies the core requirements and produces sufficiently detailed avatars for VR applications in therapeutic or clinical contexts, several limitations and potential improvements in terms of user experience remain. We plan to address these and further refine the software to soon make it available as a free and open-source resource for researchers.

Generally, A3Cplus is not limited to VR scenarios but can also be used in less critical, non-therapeutic contexts where simple yet anatomically accurate avatar creation is desired, such as in games, digital fashion stores, or computer-aided design.

ACKNOWLEDGMENTS

This study has been supported by the Luxembourg National Research Fund (FNR) under grant number C22/BM/17121206 (VR BIAS).

REFERENCES

[1] Jascha Achenbach, Thomas Waltemate, Marc Erich Latoschik, and Mario Botsch. 2017. Fast generation of realistic virtual humans. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST ’17)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3139131.3139154

[2] Desiderio Cano Porras, Petra Siemonsma, Rivka Inzelberg, Gabriel Zeilig, and Meir Plotnik. 2018. Advantages of virtual reality in the rehabilitation of balance and gait. *Neurology* 90, 22 (May 2018), 1017–1025. https://doi.org/10.1212/WNL.0000000000005603 Publisher: Wolters Kluwer.

[3] Bernie Garrett, Tarnia Taverner, Diane Gromala, Gordon Tao, Elliott Cordingley, and Crystal Sun. 2018. Virtual Reality Clinical Research: Promises and Challenges. *JMIR Serious Games* 6, 4 (Oct. 2018), e10839. https://doi.org/10.2196/10839

[4] Huifang Guan, Yan Xu, and Dexi Zhao. 2022. Application of Virtual Reality Technology in Clinical Practice, Teaching, and Research in Complementary and Alternative Medicine. *Evidence-based Complementary and Alternative Medicine : eCAM* 2022 (Aug. 2022), 1373170. https://doi.org/10.1155/2022/1373170

[5] Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The Sense of Embodiment in Virtual Reality. *Presence: Teleoperators and Virtual Environments* 21, 4 (Nov. 2012), 373–387. https://doi.org/10.1162/PRES_a_00124

[6] So-Yeon Kim, Hyojin Park, Myeongul Jung, and Kwanguk (Kenny) Kim. 2020. Impact of Body Size Match to an Avatar on the Body Ownership Illusion and User’s Subjective Experience. *Cyberpsychology, Behavior, and Social Networking* 23, 4 (April 2020), 234–241. https://doi.org/10.1089/cyber.2019.0136 Publisher: Mary Ann Liebert, Inc., publishers.

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (Oct. 2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

[8] Jessica L. Maples-Keller, Brian E. Bunnell, Sae-Jin Kim, and Barbara O. Rothbaum. 2017. The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. *Harvard review of psychiatry* 25, 3 (2017), 103–113. https://doi.org/10.1097/HRP.0000000000000138

[9] Ali Pourmand, Steven Davis, Alex Marchak, Tess Whiteside, and Neal Sikka. 2018. Virtual Reality as a Clinical Tool for Pain Management. *Current Pain and Headache Reports* 22, 8 (June 2018), 53. https://doi.org/10.1007/s11916-018-0708-2

[10] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H. Bülthoff, and Michael J. Black. 2019. The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (May 2019), 1887–1897. https://doi.org/10.1109/TVCG.2019.2898748 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[11] Guillaume Vailland, Louise Devigne, François Pasteau, Florian Nouviale, Bastien Fraudet, Émilie Leblong, Marie Babel, and Valérie Gouranton. 2021. VR based Power Wheelchair Simulator: Usability Evaluation through a Clinically Validated Task with Regular Users. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 420–427. https://doi.org/10.1109/VR50410.2021.00065 ISSN: 2642-5254.

[12] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. 2018. The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (April 2018), 1643–1652. https://doi.org/10.1109/TVCG.2018.2794629 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Spatial Publish/Subscribe - Decoupling Game State Dissemination from State Computation for Massive Multiplayer Online Games

PJ Smit
Department of E&E Engineering
Stellenbosch University
South Africa
22573216@sun.ac.za

HA Engelbrecht
Deptartment of E&E Engineering
Stellenbosch University
South Africa
hebrecht@sun.ac.za

ABSTRACT

The rapid expansion of virtual environments, particularly Massively Multi-user Virtual Environments (MMVEs), presents significant challenges in scalability and performance. Traditional Client/Server and Client/Multi-Server architectures often encounter limitations such as server overload, which can lead to lag and reduced user capacity, negatively affecting the user experience. Interest management is an important mechanism in online games for improving scalability, yet it typically involves the server sending duplicate game state update messages for each client impacted by a game state update. This paper introduces an approach that decouples state update dissemination from state computation, enabling the server to focus on state computation while a dedicated server manages the dissemination of state updates to affected clients. Using the VAST architecture, a Spatial Publish/Subscribe (SPS) library, allows for a single update message per game state update to be sent by the server, thereby replacing the traditional interest management scheme. The effectiveness of this approach is experimentally verified through the implementation of SPS on an open-source, high-performance Minecraft server, SpigotMC. Initial implementation and evaluation demonstrate that the VAST architecture effectively reduces computational load and memory usage, while optimising network traffic and latency. For up to six clients, the Minecraft server utilizing SPS exhibits a six-fold decrease in the number of transmitted update messages, marking a substantial reduction in packet transmission compared to traditional methods. This research underscores the potential of Spatial Publish/Subscribe systems in creating more scalable and efficient virtual environments, addressing the evolving demands of virtual world interactions.

CCS CONCEPTS

• **Information systems → Data management systems**; **Publish-subscribe / event-based architectures**; • **Human-centered computing** → *Visualization*; • **Theory of computation** → *Online algorithms*.

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652216

KEYWORDS

Spatial Publish/Subscribe, Virtual Environments, Scalability, Multi-server Architecture, Distributed Systems, Minecraft, Network Traffic Optimisation, Latency Reduction

ACM Reference Format:
PJ Smit and HA Engelbrecht. 2024. Spatial Publish/Subscribe - Decoupling Game State Dissemination from State Computation for Massive Multiplayer Online Games. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652216

1 INTRODUCTION

The rapid increase of virtual environments, notably in Massively Multi-user Virtual Environments (MMVEs), large-scale simulations, and Massive Multiplayer Online Games (MMOGs), underscores the need for enhanced scalability and performance [9, 35]. These environments, facilitating real-time interaction among widespread users, face challenges in network performance and scalability, particularly when supporting thousands of simultaneous users in expansive worlds.

Traditionally, MMOGs employ centralised Client/Server (C/S) or Client/Multi-Server (C/MS) architectures, handling game state computations and user interactions centrally [35]. While relatively simple, these architectures struggle with scalability, leading to server overload, communication latency, and restricted user capacity, negatively impacting user experience. To mitigate this, Interest Management (IM) [3, 24] systems limit the game state accessible to a user, confined to their Area of Interest (AOI) [17, 30], but this still binds the server capacity to the number of users within an AOI.

Addressing these constraints, we propose a distributed architecture, separating game state update dissemination from computation. We integrate Spatial Publish/Subscribe (SPS) [14], an extension of the Publish/Subscribe model with a spatial dimension, to replace traditional IM. This approach of employing a single state update message for each state change, which is disseminated by an SPS broker, significantly improves scalability.

This paper explores the feasibility of substituting a commercial MMOG's IM with an SPS-based system, specifically in Minecraft, by decoupling the game state update dissemination from the game state computation. Using VAST, an open-source SPS communication library [26], we replaced Minecraft's AOI system with a SPS-based IM system. The concept is experimentally verified using Koekepan [10, 11], a research platform that extends SpigotMC (an open-source Minecraft server clone) [32] to support server clusters.

Our contributions are twofold: introducing the SPS architecture to improve MMOG scalability and performance, demonstrated using

Minecraft, and providing experimental evidence of its efficiency in state update dissemination with comparable latency to existing C/S models.

2 BACKGROUND

2.1 Massively Multi-user Virtual Environments

MMVEs, particularly MMOGs, support interaction among thousands to tens of thousands of users through avatars in a shared virtual world [35]. MMOGs can be seen as state machines, where game state changes (such as entity property changes), are updated through state update messages sent to clients. Changes in game state are caused by internal logic or user interactions, and are managed through a cycle of *event*, *processing*, and *update*. This cycle involves state update messages being disseminated to clients after server processing ensures game logic compliance, known as state computation. Packets, the primary data transmission units in networks, are crucial for communicating state changes. Efficient packet management is key for real-time interaction in VEs, as they carry essential metadata for routing and processing within the network.

The MMOG networking architecture must scale to handle dynamic user volumes. The prevalent Client/Server (C/S) architecture features a central server maintaining the Global State of the game world, with clients' Local State being updated through state dissemination. However, this architecture faces scalability limitations due to the high cost and limitations on upgrading server processing and network communication capacity.

Client/MultiServer (C/MS) architectures distributes the Global State across multiple servers, whereas the Peer-to-Peer (P2P) model utilizes each node as both client and server, sharing server computations. Hybrid architectures combine elements of these systems, like using a central server for critical functions and P2P for less crucial data, achieving a balance of control and scalability. Examples such as the Distributed Scene Graph [18] demonstrate the effectiveness of decoupling system components in hybrid models.

2.2 Interest Management

Interest Management (IM) is essential for scalability in multiplayer games, acting as a data filtration system to manage state dissemination by limiting server-to-client updates, thus reducing network traffic and packet transmission [13, 20, 24]. This restriction on data access is based on the concept that players have limited sensing, i.e. limited movement and vision based on their Area-Of-Interest (AOI), a spatial area surrounding them. This means that game data has both spatial and temporal locality [19], and leads to restricted interaction capabilities outside a players immediate vicinity.

A common IM strategy is *zoning* or *spatial partitioning*, dividing the game world into zones [6]. The player's AOI can be the entire zone the player is located in (i.e. the zone is the entire VE available to the player) or a subarea within it, often visualised as a circle or sphere centred on the player avatar's location. The AOI adjusts when the player moves, with the server updating the game state relevant to a player's AOI.

In C/MS architectures, a common issue is that players on the borders of zones must still be able to see and interact with objects that are just across the zone boundary in a neighbouring zone. This requires more advanced interest management schemes that allow the AOI to overlap multiple zones. Managing AOIs across zone borders presents challenges like object replication [35], necessitating solutions like *mirroring*, where servers replicate game object states across neighboring zones [8, 21].

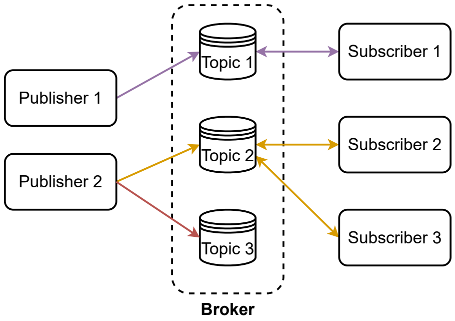


Figure 1: A Publish/Subscribe (Pub/Sub) network, where publishers and subscribers interact through topics in a decoupled system.

Despite its widespread adoption in MMOGs since its inception [13], IM has seen little innovation until 2009, when Hu [14] made a proposal to replace AOI-based IM with Spatial Publish/Subscribe (SPS). SPS, an extension to the Publish/Subscribe messaging paradigm, allows for the decoupling of game state update dissemination from game state computation, effectively creating two distinct layers for IM and the VE hosting. SPS facilitates dynamic resource allocation and removes IM overhead from VE hosting. This paper validates the SPS-based IM scheme first proposed by [14].

3 RELATED WORK

Yahyavi and Kemme [35] presented a comprehensive review of interest management (IM) in MMOGs, highlighting the need for efficient state synchronization in distributed virtual environments. Their work underscored the significance of spatially-aware Pub/Sub systems for enhancing scalability and reducing network traffic. More recently [29] presented an survey of AOI management in MMOGs, highlighting that IM is a core activity. Bharambe presents Donnybrook [4], a P2P system that enables MMOGs with hundreds of simultaneous players. Donnybrook reduces server bandwidth demands by using "interest sets" to model players' limited attention and disseminate frequent updates to only a small set of peers each player is focused on. It handles heterogeneity in peer capacity by using a dynamic forwarding pool where high-capacity peers assist lower-capacity peers.

The idea of using Publish/Subscribe messaging for IM was first suggested by Morgan [23]. However, its topic-based approach falls short for spatially-driven applications like MMOGs. Addressing this, Hu [14] proposed Spatial Publish/Subscribe (SPS) as a foundational element for Virtual Environment (VE) systems. SPS enables nodes in VEs to subscribe and broadcast within specific areas, refining message dissemination based on spatial relevance.

Buyukkaya and Abdallah [5] introduced Voronoi-based spatial partitioning in a fully distributed P2P architecture while also dealing with data management for mutable and immutable objects. In 2016 Abdulazeez proposed a static AOI management scheme and evaluated using simulation in OPNET Modeler 18.0 to simulate up to 1000 nodes [2] and in 2017 the static interest management scheme was made dynamic [1].

Perhaps the most similar recent work is Cloud Imperium Games' Star Citizen. It aims to enable massive-scale multiplayer spaceship combat through what they call Server Meshing. This interconnects game servers to distribute load by making use of a Replication Layer which players are connected to and allows seamless transition between server instances and entity syncing. All data gets passed through the replication layer to appropriate servers and clients. Recent improvements have introduced Persistent Entity Streaming (PES) across servers (via the replication layer) and allow servers to unload data of neighbouring servers that are not applicable or within view of the server's geographic area. The replication layer makes use of Object Container Streaming, which only transmits the necessary subset of game data between servers. Although promising, server meshing currently only supports basic static meshing and the technology is still in the early stages[27, 28].

To our knowledge, none of the related work addresses the issue that update dissemination using AOI management still scales linearly with the number of affected clients.

4 SPATIAL PUBLISH/SUBSCRIBE ARCHITECTURE (SPS)

The Publish/Subscribe (Pub/Sub) architecture, crucial to SPS, provides a flexible alternative to traditional network architectures like the C/S model for distributing information [12, 31]. In the basic Pub/Sub system, illustrated in Fig. 1, publishers send messages to a broker with a specific topic, and subscribers receive messages of their subscribed topics, maintaining anonymity between them. This model offers network scalability and supports event-driven architectures through time and space decoupling between publishers and subscribers.

However, the traditional Pub/Sub architecture, despite being beneficial in IoT applications such as MQTT [25], falls short in MMOGs due to its topic-based messaging. It does not accommodate the spatial and temporal locality inherent in MMOG packets. Some Pub/Sub protocols do offer geo-support but are restricted to specific locations and don't acommodate generic spatial locations [26]. Additionally, Pub/Sub architectures typically don't natively support multiple brokers, and often lack efficient message forwarding algorithms or a partitioning overlay network for load management.

To overcome these spatial limitations, Hu introduced SPS, which integrates spatial information into the Pub/Sub paradigm [14, 16], making it suitable for spatially dynamic applications such as MMOGs. As shown in Fig. 2, participants (publishers, subscribers and brokers) define a Area Subscription or Point Subscription, with spatial messages routed to clients whose Subscriptions intersect with a Area Publication or Point Publication. SPS supports four operations: (1) PUBLISH messages to a spatial area, (2) SUBSCRIBE to messages in a spatial area, (3) UNSUBSCRIBE from a spatial area, and (4) MOVE the participant's location. Hu [14] also provides a brief overview of how SPS can be used in the context of MMOGs.

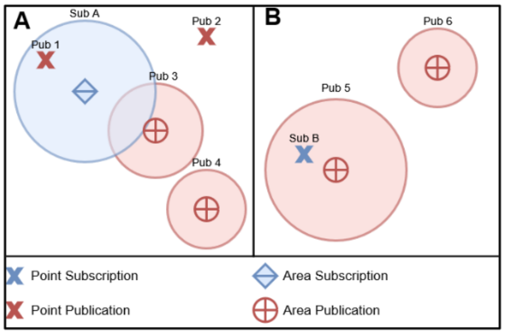


Figure 2: The operational framework of Spatial Publish/Subscribe (SPS) within a virtual environment.

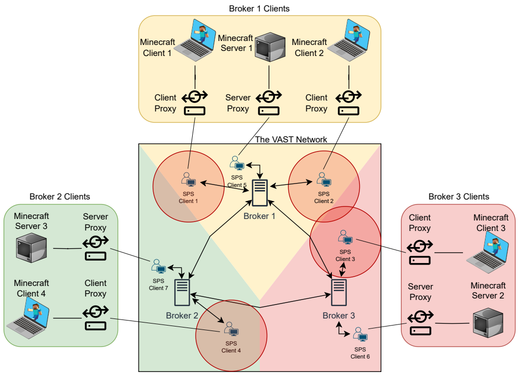


Figure 3: The VAST network implementation in Minecraft.

4.1 VAST

The VAST network library, developed in Javascript, is an implementation of the SPS paradigm, designed to enhance communication in virtual environments [26]. VAST introduces multiple brokers, termed Interest Matchers, each overseeing a specific region in the Virtual Environment (VE). These regions are organised through Voronoi Partitioning within the Voronoi Overlay Network (VON), optimising load distribution and preventing node overload [15, 33]. The VoroCast algorithm is employed for efficient message forwarding to the appropriate broker when publications fall outside a broker's region, creating a spanning tree for seamless inter-broker communication [7].

VAST operates on a peer-to-peer (P2P) framework, allowing nodes to dynamically join, leave, or change their positions within the network. Its communication is based on spatial relevance, ensuring that only relevant brokers and clients process and communicate pertinent events. The library supports the fundamental SPS operations: publications (transmission of spatially relevant information) and subscriptions (receiving updates on publications within a specific spatial area), with brokers accomodating these roles based on client locations and interests.

Fig. 3 demonstrates VAST's application in a C/MS MMOG architecture, using Minecraft as an example. Minecraft clients and servers correspond to SPS clients on the network. Each SPS client interacts only with the broker in their area, with Minecraft client subscriptions (their AOI) depicted as red circles around SPS Clients. Minecraft servers subscribe to the entire area under their broker's responsibility. Spatial messages published by SPS clients are forwarded by the relevant broker to the appropriate client, possibly via other brokers. Further details about VAST and its use of multiple SPS brokers are available in [26].

5 METHODOLOGY

In this section we introduce the popular commercial MMOG Minecraft, which is used to experimentally evaluate the proposed SPS-based IM scheme. After discussing the technical details of Minecraft, we delve into the practical implementation of the VAST SPS library within a Minecraft multiplayer server environment. We briefly discuss the system architecture and the modifications made to the standard Minecraft C/S model for VAST integration.

5.1 Minecraft

Minecraft is a sandbox-style MMVE game developed by Mojang Studios [22], which operates within a procedurally generated 3D environment comprised of block elements. Its VE is divided into "chunks" (16x16 block areas, extending vertically up to 265 blocks), forming the basic unit of the game's spatial structure.

In multiplayer mode, Minecraft adopts a C/S model [34]. The server holds the *global game state*, managing all aspects of the game world, including chunks, block types, entities, and other elements. Clients connect to this server, receiving data and state changes necessary to render their *local game state* – essentially, the immediate surroundings they interact with.

Minecraft's IM system is designed to optimize network traffic and server load by transmitting information based on players' locations and actions, essentially an AOI system. For example, only chunks within a player's view are sent to their client, minimizing network data transmission.

Minecraft's architecture mirrors that of typical MMOGs, featuring expansive, interactive worlds. It, however, uniquely allows extensive player-driven modification of both game environment and mechanics. This flexibility makes Minecraft a suitable platform for networking research in MMVEs. However, its non-open-source nature poses challenges for modifying core game components.

5.1.1 Bukkit and Koekepan. To enhance Minecraft's functionality, the community has developed tools like Bukkit, an open-source API framework, and CraftBukkit, an open-source server clone. Bukkit offers event-based software hooks into the Native Minecraft Server (NMS) code, enabling third-party 'mods' to interact with events processed by the server. SpigotMC, a high-performance version of CraftBukkit [32], optimizes NMS server code by improving entity handling, chunk loading, and packet transmission.

In previous work we proposed Koekepan [10, 11], an extension to SpigotMC, enhancing the Minecraft server's networking for distributed server architectures. It uses spatial partitioning with Voronoi diagrams to allocate server regions and introduces a proxy between Minecraft clients and servers, facilitating entity migration and load distribution.

5.2 System Architecture and Design

Integrating the VAST library into Minecraft's multiplayer environment necessitated adapting the existing C/S model to accommodate SPS mechanisms while preserving game functionality.

Integration into Minecraft: The integration involved two key steps: removing Minecraft's existing IM and implementing an SPS Broker for packet management. Minecraft uses TCP connections for C/S communication. We broadly catagorise the packets as Spatial, Player Specific, and Global based on their spatial relevance [34].

We modified the Koekepan architecture, dividing the single proxy into separate server and client proxies, connected via an SPS broker, as shown in Fig. 3. The server proxy manages TCP connections to the server for each client, forwarding packets without direct client-server connections. The client proxy similarly manages server message transmission to clients. This setup allows for the indirect exchange of packets via the SPS broker.

In the original architecture, the server broadcasts game state updates to all relevant clients, identified by their AOI. With SPS-based IM, the server publishes a single update, offloading IM from the NMS code to the SPS Broker, thus requiring modifications to the SpigotMC and Koekepan server applications. Additionally, we introduced a dedicated TCP connection between the Minecraft server and server proxy for non-client-specific SPS Publications, such as spatial packets.

The resultant SPS-Koekepan server architecture, depicted in Fig. 3, integrates SPS-based IM within Minecraft's architecture.

System Components: Key components in this architecture (Fig. 3) include the SPS broker (VAST Matcher), handling subscriptions and publications based on spatial dynamics, and corresponding server and client proxies for Minecraft. The server proxy translates between Minecraft and SPS packets, managing spatial packet publication. It also maintains a TCP connection with the server for transmitting spatial packets.

Client proxies serve dual roles: interfacing between the Minecraft client and the SPS network, and managing client-specific SPS interactions. Each client proxy connects to an individual SPS client, representing the Minecraft client in the SPS network, and interfaces with the Broker for publications and subscriptions.

6 EXPERIMENTAL VALIDATION

6.1 Experimental Setup

Our experimental setup includes a single Minecraft server and up to six clients, all running on a host equipped with a 12-core Intel i7-12700 (4.8GHz) CPU and 32GB DDR4 memory, using POP!_OS based on Debian 22.04.

We use two baseline systems: the Java native Minecraft server (NMS) version 1.11.2 from Mojang Studios and the open-source, high-performance SpigotMC server version 1.11.2, which includes the NMS code. Both baselines implement AOI Management within the NMS code. The version number refers to the Minecraft communication protocol for server-client information exchange.

Our system, termed SPS-Koekepan, is a modified version 1.11.2 SpigotMC server integrating the SPS interest management scheme. We employ a single SPS broker, although VAST supports multiple brokers. The evaluation involves three server configurations: NMS, SpigotMC, and SPS-Koekepan. The simulation procedure for each configuration is as follows:

(1) Initialize all system components: server, clients, proxies, and broker.
(2) Connect from 1 to a maximum of 6 clients, which then navigates the environment for 120 seconds. All clients stay predominantly within each others' AOI.
(3) Disconnect all clients from the server.
(4) Continue system operation for an additional 120 seconds to collect post-operation performance data.
(5) Terminate all system components.

The "Superflat" Minecraft world is used, which does not influence the size of the chunk data sent to the clients, with non-player entities roaming. The presented results are the average of 10 simulation runs for each experimental setup. A client emulator replicates avatar movements and packet transmissions for consistency. We use a client emulator that implements the version 1.11.2 communication protocol, allowing each client to be programmed to repeat the same avatar movement in each simulation. The client emulator sends the same packets during each simulation.

The evaluation focusses on update dissemination efficiency, measuring total messages sent from the server to clients. Fewer transmissions indicate better network efficiency and reduced duplication of state updates. Additionally, we assess the latency added by SPS, with less than 100 ms being acceptable for MMOGs, and evaluate the SPS broker's CPU and memory usage using Linux's process status (*ps)* application, providing insights into scalability.

6.2 Experimental Results

6.2.1 Update Dissemination Efficiency. Analysis of the total server packet transmissions for the two baseline configurations, NMS and SpigotMC, reveals a proportional increase in transmissions with the number of connected clients. Notably, the NMS configuration consistently sends more transmissions than SpigotMC. For example, with six clients connected, the NMS server sends approximately 300,000 packets, while the SpigotMC server transmits around 100,000, indicating that SpigotMC requires roughly one third the number of packets compared to NMS. This reduction is attributed to SpigotMC's efficient handling of entity movement, minimising redundant packet transmissions.

On the contrary, the SPS-Koekepan configuration demonstrates even greater efficiency. With six clients, it averages around 50,000 packets, six times less than NMS and half of what SpigotMC transmits. This efficiency becomes more pronounced with an increasing number of clients, suggesting the superiority of the SPS scheme in update dissemination. This is starkly illustrated when the server packet transmission is compared in Fig. 4. SPS-Koekepan achieves this by sending a single packet per state update, compared to the multiple duplicate packets necessary for NMS and SpigotMC.

Interestingly, for a single connected client, both SPS-Koekepan and SpigotMC transmit a similar number of packets, aligning with expectations since SPS-Koekepan is an extension of SpigotMC and thus should transmit the same number of server packets when only a single client is connected. However, as client numbers increase to six, SPS-Koekepan maintains its efficiency, transmitting approximately half the packets of SpigotMC and one-sixth of NMS. This pattern, representing the SPS scheme's update dissemination efficiency, is highlighted in Fig. 4, which summarizes the server packet transmissions across the different server configurations.

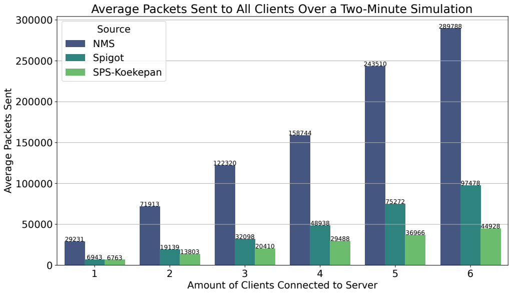


Figure 4: Comparison of Total Server Packet Transmissions for NMS, SpigotMC and SPS-Koekepan configurations

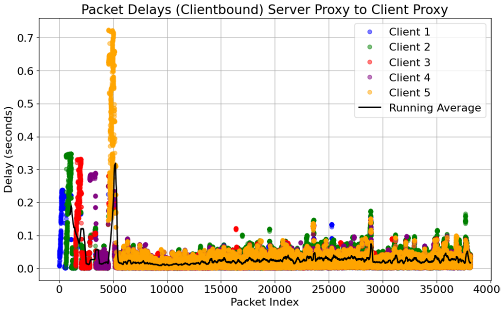


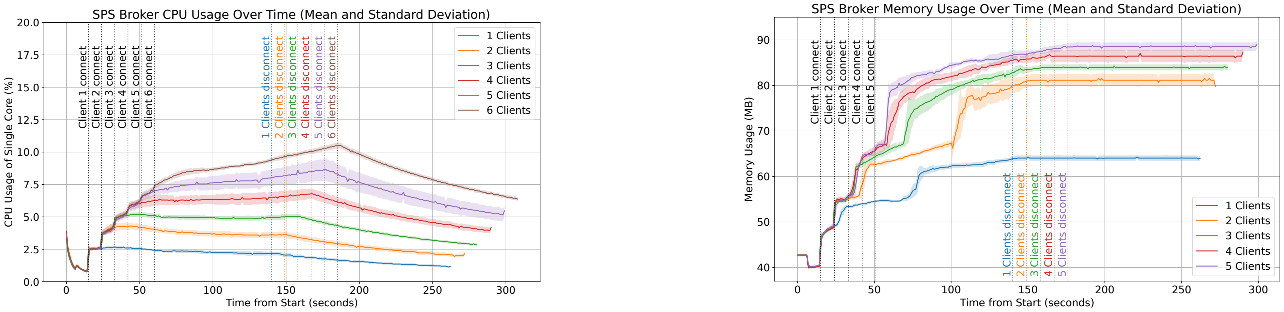
Figure 5: Latency variations with packet delivery for five clients connected to a server.

The data clearly show that while an increase in connected clients results in more server packet transmissions for all configurations, the rate of increase is substantially lower for SPS-Koekepan. This trend underscores the enhanced efficiency of the SPS-based scheme, particularly as the client count rises.

6.2.2 Latency Introduced by SPS scheme. We assess the communication latency added by the SPS scheme to server packets sent to clients. Latency is measured by timestamping network packets, with all simulations on the same host for accurate latency measurements. The latency, from server packet transmission to client reception, is reported for up to five clients connected to a single SPS-Koekepan server.

Fig. 5 shows the average latency for five clients (Client-1 to Client-5) during simulations. Each data point reflects the additional latency from the SPS broker, server and client proxies, and SPS clients, as network packets are transmitted from the SPS-Koekepan server to the clients. The solid black line indicates the running average latency, representing the typical additional latency introduced by the SPS scheme. The packet index is increased each time the server transmits a packet to a client, thus the packet index represents the ordered sequence of sent packets to an individual client.

An initial latency 'spike' is noted for each client at the start of their connection session, particularly within the first 5000 packets. This spike results from the server sending large chunk packets



(a) SPS Broker CPU usage

(b) SPS Broker Memory usage

Figure 6: Average resource utilization of the SPS broker for each client scenario, after 5 simulation runs. Standard deviation is indicated in solid colours.

containing game state data for Minecraft blocks when clients first connect. Each chunk packet consists of the game state data for a column of 16x16x256 Minecraft blocks, and the server sends about 441 chunk packets to each of the clients when they connect. The chunk packets are needed so the client has a local copy of the game state of the Minecraft virtual world. The transmission of these chunk packets explains the increased latency initially observed of up to 700ms. Once the chunk packet transmission is completed, the latency introduced by SPS decreases and stabilizes under 100ms. Occasional peaks in latency are likely due to player movements and requests for more chunk data. A running average latency of around 20ms suggests that, even with five connected clients, the additional latency from SPS is negligible and well below the 100ms MMOG threshold. Considering separate host deployment, an additional 80ms is available for network-induced latency.

6.2.3 Computational and Storage Demand of SPS Broker. The SPS-based scheme's decoupling of game state update computation from dissemination assigns the latter responsibility to the SPS broker. We examined the computational and memory demands of the SPS broker, focusing on CPU and memory usage. Figs. 6a and 6b display the broker's average CPU and memory usage over time, respectively, for one to six clients across five simulations.

Fig. 6a illustrates the CPU usage spike upon each client's connection, indicating an approximate 2% CPU usage increase per client for our test hardware. When clients disconnect, a corresponding decrease in CPU usage is noticeable. Fig. 6b highlights the broker's memory usage, which also rises with client numbers. However, the incremental memory increase per client diminishes with more connections, hinting at a potential plateau in memory demand, however an experiment with significantly more connected clients is needed to verify this conclusion.

The data indicate that update dissemination burden placed on the SPS broker is not significant, with a maximum of 10% CPU usage for one core and no more than 90MB memory for six clients. Assuming linear growth in CPU demand, as Fig. 6a implies, the broker could theoretically support up to 60 clients on a single core.

7 CONCLUSION AND FUTURE WORK

In this paper, we have presented the Spatial Publish/Subscribe (SPS) interest management scheme, which allows us to decouple the game state dissemination from the game state computation performed by a server within the context of Massively Multiplayer Online Games (MMOGs). We highlighted the limitations of traditional area-of-interest management in commercial MMOGs and introduced SPS, integrating spatial information into state updates. This paper, building on Hu's proposal of 2009 [14], provides the first experimental validation of SPS using Minecraft, a widely played commercial MMOG.

Our experiments demonstrated that for six connected clients, SPS significantly reduces server packet transmission by up to six times compared to the native Minecraft server. Moreover, the additional latency introduced by the SPS broker remained below the critical 100ms threshold, averaging around 20ms, thus preserving user experience. The computational and storage demands of the SPS broker were also found to be moderate, with a maximum of 10% CPU usage of a single core and 90MB memory usage, indicating the capacity to handle up to 600 connected clients.

While promising, these preliminary results were obtained with only six clients and should be validated with a larger number of clients running on separate hosts, to also include the effects of network latency. Nevertheless, we are of the opinion that the results do indicate that the SPS-scheme can be used as a replacement for the traditional area-of-interest management scheme.

Future work will focus on validating SPS's performance with a significantly larger number of clients on separate hosts and assessing the impact of a real local area network on latency. Additionally, we plan to test the SPS scheme in the Koekepan architecture, which supports hosting a Minecraft virtual world on a server cluster with up to 120 nodes, using Voronoi-based spatial partitioning. This setup, allowing avatar migration between server nodes without the need for *mirroring* [8, 21], will be a crucial test for SPS's effectiveness in a distributed server environment.

Furthermore, we aim to explore the scalability implications of using multiple SPS brokers, as supported by the VAST implementation. This includes examining the potential increase in transmission latency and, if necessary, considering performance enhancements by reimplementing the VAST library in a more efficient programming language. These steps will be critical in fully understanding and leveraging the benefits of SPS in large-scale virtual environments.

REFERENCES

[1] Sarmad A. Abdulazeez, Abdennour El Rhalibi, and Dhiya Al-Jumeily. 2017. Dynamic Area of Interest Management for Massively Multiplayer Online Games Using OPNET. In *2017 10th International Conference on Developments in eSystems Engineering (DeSE)*. 50–55. https://doi.org/10.1109/DeSE.2017.19

[2] Sarmad A. Abdulazeez, Abdennour El Rhalibi, and Dhiya Al-Jumeily. 2016. Simulation of Area of Interest Management for Massively Multiplayer Online Games Using OPNET. In *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*. 163–168. https://doi.org/10.1109/DeSE.2016.28

[3] Steve Benford and Lennart Fahlén. 1993. *A Spatial Model of Interaction in Large Virtual Environments*. Springer Netherlands, Dordrecht, 109–124. https://doi.org/10.1007/978-94-011-2094-4_8

[4] Ashwin Bharambe, John R. Douceur, Jacob R. Lorch, Thomas Moscibroda, Jeffrey Pang, Srinivasan Seshan, and Xinyu Zhuang. 2008. Donnybrook: Enabling Large-Scale, High-Speed, Peer-to-Peer Games, In Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication (Seattle, WA, USA). *SIGCOMM Comput. Commun. Rev.* 38, 4, 389–400. https://doi.org/10.1145/1402958.1403002

[5] Eliya Buyukkaya and Maha Abdallah. 2008. Data Management in Voronoi-Based P2P Gaming. In *2008 5th IEEE Consumer Communications and Networking Conference*. 1050–1053. https://doi.org/10.1109/ccnc08.2007.239

[6] Wentong Cai, P. Xavier, S.J. Turner, and Bu-Sung Lee. 2002. A scalable architecture for supporting interactive games on the internet. In *Proceedings 16th Workshop on Parallel and Distributed Simulation*. 54–61. https://doi.org/10.1109/PADS.2002.1004201

[7] Jui-Fa Chen, Wei-Chuan Lin, Tsu-Han Chen, and Shun-Yun Hu. 2007. A forwarding model for Voronoi-based Overlay Network. *2007 International Conference on Parallel and Distributed Systems*, 1–7. https://doi.org/10.1109/ICPADS.2007.4447818

[8] Eric Cronin, Anthony R Kurc, Burton Filstrup, and Sugih Jamin. 2004. An Efficient Synchronization Mechanism for Mirrored Game Architectures. *Multimedia Tools and Applications* 23, 1 (2004), 7–30. https://doi.org/10.1023/B:MTAP.0000026839.31028.9f

[9] Herman A. Engelbrecht and John S. Gilmore. 2017. Pithos: Distributed Storage for Massive Multi-User Virtual Environments. *ACM Transactions on Multimedia Computing, Communications and Applications* 13, 3, Article 31 (6 2017), 33 pages. Issue 3. https://doi.org/10.1145/3105577

[10] Herman A. Engelbrecht and Gregor Schiele. 2013. Koekepan: Minecraft as a research platform, In 2013 12th Annual Workshop on Network and Systems Support for Games (NetGames). *Annual Workshop on Network and Systems Support for Games*, 1–3. https://doi.org/10.1109/NetGames.2013.6820615

[11] Herman A .Engelbrecht and Gregor Schiele. 2014. Transforming Minecraft into a research platform. *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, 257–262. https://doi.org/10.1109/CCNC.2014.6866580

[12] Patrick Th. Eugster, Pascal A. Felber, Rachid Guerraoui, and Anne-Marie Kermarrec. 2003. The Many Faces of Publish/Subscribe. *ACM Comput. Surv.* 35, 2 (jun 2003), 114–131. https://doi.org/10.1145/857076.857078

[13] Lennart Fahlen, The Swedish Institute of Computer Science (SICS), and Sweden. 1993. A Spatial Model of Interaction in Large Virtual Environments Steve Benford The University of Nottingham, UK. In *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13-17 September, 1993, Milan, Italy G De Michelis, C Simone and K. Schmidt (Editors)*.

[14] Shun-Yun Hu. 2009. Spatial Publish Subscribe. *2nd International Workshop on Massively Multiuser Virtual Environments* (9 2009). http://www.badumna.com/

[15] Shun-Yun Hu, Jui-Fa Chen, and Tsu-Han Chen. 2006. VON: a scalable peer-to-peer network for virtual environments. *IEEE Network* 20 (2006), 22–31. https://ieeexplore.ieee.org/document/1668400

[16] Shun-Yun Hu, Chuan Wu, Eliya Buyukkaya, Chien-Hao Chien, Tzu-Hao Lin, Maha Abdallah, Jehn-Ruey Jiang, and Kuan-Ta Chen. 2010. A spatial publish subscribe overlay for massively multiuser virtual environments. *2010 International Conference on Electronics and Information Engineering* 2, V2–314–V2–318. https://doi.org/10.1109/ICEIE.2010.5559789

[17] Björn Knutsson, Honghui Lu, Wei Xu, and Bryan Hopkins. 2004. Peer-to-peer support for massively multiplayer games. In *23rd Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM)*, Vol. 1. 107.

[18] Huaiyu Liu, Mic Bowman, Robert Adams, John Hurliman, and Dan Lake. 2010. Scaling virtual worlds: Simulation requirements and challenges, In Proceedings of the 2010 Winter Simulation Conference. *Proceedings of the 2010 Winter Simulation Conference*, 778–790. https://doi.org/10.1109/WSC.2010.5679112

[19] Yohai Makbily, Craig Gotsman, and Reuven Bar-Yehuda. 1999. Geometric algorithms for message filtering in decentralized virtual environments. 39–46. https://doi.org/10.1145/300523.300527

[20] Tielman Francois Septimus Malherbe. 2016. *A Comparative Study of Interest Management Schemes in Peer-to-Peer Massively Multiuser Networked Virtual Environments*. mathesis. Stellenbosch University, www.eng.sun.ac.za.

[21] Martin Mauve, Stefan Fischer, and Jörg Widmer. 2002. A generic proxy system for networked computer games. In *Proceedings of the 1st workshop on Network and system support for games - NETGAMES '02*. ACM Press, New York, New York, USA, 25–28. https://doi.org/10.1145/566500.566504

[22] Mojang. 2011. What is Minecraft? build, discover realms & more. https://www.minecraft.net/en-us/about-minecraft

[23] Graham Morgan, Fengyun Lu, and Kier Storey. 2005. Interest Management Middleware for Networked Games. In *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games* (Washington, District of Columbia) *(I3D '05)*. Association for Computing Machinery, New York, NY, USA, 57–64. https://doi.org/10.1145/1053427.1053436

[24] Katherine L. Morse. 1996. Interest Management in Large-Scale Distributed Simulations. In *Technical Report 96-27*.

[25] OASIS. 1999. The standard for IOT messaging. https://mqtt.org/

[26] Victory Opeolu, Herman Engelbrecht, Shun-Yun Hu, and Charl Marais. 2023. VAST: A Decentralized Open-Source Publish/Subscribe Architecture. *Proceedings of the 14th Conference on ACM Multimedia Systems*, 423–429. https://doi.org/10.1145/3587819.3592554

[27] Paul Reindell. 2023. *CitizenCon 2953 Highlight | Server Meshing, PES & Replication Layer*. Youtube | PES. https://www.youtube.com/watch?v=fAbcr35_Teg

[28] Paul Reindell, Benoit Beausejour, Roger Godfrey, and Clive Johnson. 2023. Server Meshing and Persistent Streaming Q&A. https://robertsspaceindustries.com/comm-link/transmission/18397-Server-Meshing-And-Persistent-Streaming-Q-A

[29] Laura Ricci and Emanuele Carlini. 2018. *Area of Interest Management in Massively Multiplayer Online Games*. Springer International Publishing, Cham, 1–3. https://doi.org/10.1007/978-3-319-08234-9_239-1

[30] Arne Schmieg, Michael Stieler, Sebastian Jeckel, Patric Kabus, Bettina Kemme, and Alejandro Buchmann. 2008. pSense - Maintaining a Dynamic Localized Peer-to-Peer Structure for Position Based Multicast in Games. In *2008 Eighth International Conference on Peer-to-Peer Computing*. 247–256. https://doi.org/10.1109/P2P.2008.20

[31] Tarek R Sheltami, Anas A Al-Roubaiey, and Ashraf S Hasan Mahmoud. 2016. A survey on developing publish/subscribe middleware over wireless sensor/actuator networks. *Wireless Networks* 22 (2016), 2049–2070. Issue 6. https://doi.org/10.1007/s11276-015-1075-0

[32] SPIGOTMC. 2012. SPIGOTMC - High Performance Minecraft. https://www.spigotmc.org/

[33] Manrich van Greunen and Herman A. Engelbrecht. 2014. A comparison of Quad-tree and Voronoi-based spatial partitioning for dynamic load balancing. *Proceedings of the first International Conference on the use of Mobile Informations and Communication Technology (ICT) in Africa UMICTA 2014*. http://hdl.handle.net/10019.1/96157

[34] wiki.vg contributors. 2019. Protocol: 1.11.2 Minecraft Protocol Documentation. https://wiki.vg/index.php?title=Protocol&oldid=8543 Online; accessed September 27, 2023, version with oldid=8543.

[35] Amir Yahyavi and Bettina Kemme. 2013. Peer-to-Peer Architectures for Massively Multiplayer Online Games: A Survey. *ACM Computing Surveys (CSUR)* 46 (7 2013). https://doi.org/10.1145/2522968.2522977

Influence of Gameplay Duration, Hand Tracking, and Controller Based Control Methods on UX in VR

Tanja Kojić
Quality and Usability Lab
Technical University of Berlin

Maurizio Vergari
Quality and Usability Lab
Technical University of Berlin

Simon Knuth
Quality and Usability Lab
Technical University of Berlin

Maximilian Warsinke
Quality and Usability Lab
Technical University of Berlin

Sebastian Möller
Quality and Usability Lab
Technical University of Berlin and
erman Research Center for Artificial
Intelligence (DFKI)

Jan-Niklas Voigt-Antons
Immersive Reality Lab
Hamm-Lippstadt University of
Applied Sciences

ABSTRACT

Inside-out tracking is growing popular in consumer VR, enhancing accessibility. It uses HMD camera data and neural networks for effective hand tracking. However, limited user experience studies have compared this method to traditional controllers, with no consensus on the optimal control technique. This paper investigates the impact of control methods and gaming duration on VR user experience, hypothesizing hand tracking might be preferred for short sessions and by users new to VR due to its simplicity. Through a lab study with twenty participants, evaluating presence, emotional response, UX quality, and flow, findings revealed control type and session length affect user experience without significant interaction. Controllers were generally superior, attributed to their reliability, and longer sessions increased presence and realism. The study found that individuals with more VR experience were more inclined to recommend hand tracking to others, which contradicted predictions.

CCS CONCEPTS

• **Human-centered computing → Virtual reality**; **Interaction paradigms**.

KEYWORDS

Gameplay Duration, Hand Tracking

ACM Reference Format:
Tanja Kojić, Maurizio Vergari, Simon Knuth, Maximilian Warsinke, Sebastian Möller, and Jan-Niklas Voigt-Antons. 2024. Influence of Gameplay Duration, Hand Tracking, and Controller Based Control Methods on UX in VR. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652222

ACM ISBN 979-8-4007-0618-9/24/04
https://doi.org/10.1145/3652212.3652222

1 INTRODUCTION

Virtual Reality (VR) headsets are rapidly gaining popularity, with sales expected to triple in three years by 2023 [18]. The development of standalone head-mounted displays (HMDs) with inside-out tracking has surely contributed to their success. This device type opens up the platform to a wider audience by eliminating the need for advanced technical abilities, high-performance PCs, and sensor setups. VR headsets, like game consoles, are typically controlled using specialised handheld controllers. These controllers allow for low-latency interaction with 3D material by tracking them in space. Immersion in virtual environments relies heavily on input, with more natural-looking input leading to higher levels of immersion [12]. VR systems try to simulate real-world interactions as precisely as feasible. The next step for input is to eliminate controllers and use only hand tracking. Modern headsets include integrated cameras for inside-out tracking, which can be used to track hands and fingers with great precision [6]. The Meta Quest platform demonstrates this capability. Using hand tracking instead of a controller can lower the barrier of entry for those unfamiliar with VR, eliminating the need for button mappings. Previous evaluations of these control systems in user experience (UX) have shown inconsistent results [5, 8, 19], suggesting another element may be at play. This paper aims to explore the impact of gameplay time in relationship with the control method, as well as assessing users' willingness to interact with technology to see if those who are more open to new systems [4] are more likely to be convinced by hand tracking, given the limitations of current methods.

1.1 Gameplay Duration and Technology Affinity

In the field of user study design, it is usual practice to keep the duration of the experience uniform throughout the experiment. However, there have been cases where researchers deviated from this pattern, undertaking studies that investigated the impact of different experience durations on user satisfaction and engagement. For example, in one study, participants engaged with a VR game for 2 and 5 minutes, providing insight into the possible implications of time on user experience. Intriguingly, the data revealed a potential link between longer durations and increased flow experiences, though it should be noted that this study did not provide thorough insights into other critical aspects of user experience [20].

In order to explore more into the world of user adaptation to innovative technologies, it becomes clear that the process frequently necessitates a time and effort investment on the part of the user. This investment might vary greatly depending on things such as previous experiences and personal character features. Some users are naturally hesitant to face the obstacles of unknown technology, but others are eager to embrace and study new systems and functionalities in order to solve problems more efficiently. The Affinity for Technology Interaction (ATI) questionnaire accurately captures and models these two conflicting preferences [4]. Interestingly, while the ATI questionnaire is a repeating component of user studies involving participants' interactions with technology, it is rarely used as an independent variable in research designs.

1.2 Hand Tracking as Control Method

Inside-out tracking technology is widely used in consumer VR systems, with recent examples including the Meta Quest and Pico. This system uses data from a series of integrated cameras to detect complex hand motions in three-dimensional space in real time. Hand tracking's exceptional accuracy makes it a very practical way to navigate some of the different virtual environments. However, despite many advantages, hand tracking technology is not without disadvantages.

One prominent challenge is the complexity of hand-to-hand interactions and the tracking of unusual hand positions [6]. The headset's camera system has a limited field of view, which is a major concern. While many hand activities occur directly in front of the user's face, because they are naturally focused on these actions, some tasks, particularly those that replicate natural movements, take place in the peripheral and lower fields of vision. This offers an important challenge because gestures conducted in these locations may not be efficiently caught by as many cameras, resulting in a decrease in tracking accuracy. In rare situations, specific movements may fall totally outside the scope of the cameras, resulting in a complete loss of tracking functionality [3]. As a result, while inside-out tracking technology has transformed VR interaction, overcoming the issues associated with field of view constraints and guaranteeing consistent tracking precision in all hand positions remains an important focus of research and development within the VR industry.

While hand tracking technology is not without its limitations, it presents an promising potential for enhancing immersion within virtual reality environments. One of its main advantages is its ability to improve the user experience by expanding the range of "natural sensorimotor contingencies for perception" offered by VR systems [16]. Modern VR controllers, while effective in many respects, are limited by their design. They can only track certain parts of the hands, such as individual fingers, and restrict the range of hand poses that users can perform while holding them. This constraint is especially apparent when considering scenarios involving complicated 3D manipulation activities, in which the VR system must collect and duplicate details of these manipulations [14]. In these cases, hand tracking technology appears as a more appealing option than controllers, at least when physical feedback is not the major concern. Hand tracking's capacity to accurately simulate natural hand movements and gestures has the potential to provide users with a more intuitive and immersive virtual experience.

Several studies have investigated the potential of hand tracking technologies in virtual reality (VR), resulting in a complex findings and suggestions. In one such study, which intended to determine the comparative usefulness and satisfaction levels of VR controllers and hand tracking within a medical training simulation, no significant differences were discovered [8]. In contrast, another study investigated the effectiveness of these two control approaches in carrying out simple reach-pick-place tasks. Surprisingly, the results favoured controllers, both in terms of objective performance measures and participant subjective ratings [5]. This finding highlights the complex character of the hand tracking vs. controller argument, implying that the choice between different control systems may be determined by the unique environment and job at hand.

Furthermore, a third study added another degree of complexity to this discussion by concluding that, while hand tracking technology resulted in a more positive overall user experience, it appeared to come at the expense of lower perceived dominance in contrast to controllers [19]. This intriguing contradiction highlights the complex nature of the comparison between these two control modalities, implying that factors other than usability may influence the final preference for one over the other. It is important to note that, while these studies provide useful insights, they only partially overlap in their judgements, and they do not give a clear consensus on how hand tracking technology compares to traditional controllers in the VR landscape. As a result, additional study and a thorough examination of the unique situations and user preferences will be needed to untangle the details of this ongoing research and reach more definitive findings.

1.3 Objectives

In terms of VR games, hand tracking and controllers both offer advantages and disadvantages. Previously mentioned studies have been conducted to determine how they affect the user experience, but few have examined how the handling mechanism and game length interact.

While hand tracking technology has huge potential, it has yet to become widely used in consumer products. This raises interesting questions about the factors influencing its adoption. Specifically, it is important to explore the effect of users' ATI in shaping users willingness to embrace hand tracking. A higher ATI score may indicate an increased interest to investigate and experiment with this technology, regardless of its occasional technical difficulties. As a result, understanding how ATI interacts with the popularity and use of hand tracking in VR experiences aims to give insight into its future direction in the constantly evolving arena of VR technology.

Therefore two research questions have been created for this paper as:

- How do control method and gameplay duration influence user experience for virtual reality gaming, and is there an interaction effect between the two?
- Is hand tracking more popular with people who are less experienced with virtual reality, and does this preference vary over different gameplay durations?

2 METHODOLOGY

The study was mainly aimed to evaluate the user experience in VR and to collect additional data as needed to answer our research questions. To ensure consistency, we carried out the experiment in a controlled laboratory setting on the university campus. For the practical VR component of the research, we chose two different gaming durations and two separate control approaches. Each period related to a separate VR game.

The shorter duration lasted three minutes and included the puzzle game Cubism. We restart the game for each player, beginning with the first levels. Each level introduced a new three-dimensional geometric form that players had to "assemble" using a predetermined set of smaller pieces. The solution options were restricted, but both the goal form and the smaller shapes could be freely rotated and shifted, expanding the number of potential locations. All of the forms floated inside the available area, and players were free to shift their viewpoint. Interactions focused mostly on grabbing moving, rotating, and releasing shapes. The longer duration lasted nine minutes and included the interactive narrative game Vacation Simulator. Like the shorter game, we reset it for each player, beginning with a special lesson. Participants played the game's "Back to Job" option, which provided an infinite simulation of a receptionist's job at a holiday resort. The space was restricted to a main workstation and a kitchen area. Players were required to complete the resort's visitors' basic demands by discovering and interacting with the appropriate items in their surroundings. The game has an episodic format, with each episode comprising one or more main tasks. A virtual screen showed visuals of what participants should be looking for and what actions they needed to do with the items. Interactions mostly consisted on grabbing, moving, rotating, and releasing items, with rare interactions with virtual buttons.

The average session time for VR headsets, excluding those that rely on mobile phones, is about 46 minutes [17]. To avoid possible VR-induced symptoms and consequences, it is recommended that session lengths be limited to 55 to 70 minutes while conducting user research in VR [9]. Given that many participants may be experiencing VR for the first time, it is best to aim for somewhat shorter periods to avoid problems. Given the necessity for participants to complete surveys and follow instructions, the overall study time was set at around 60 minutes. A bit less than half of this time was spent within the VR headset, comfortably falling below of the limit.

The study's control techniques included two options: controllers and hand tracking. This research used the Meta Quest 2 (previously known as Oculus Quest 2), a commercial VR headset with six degrees of freedom. This independent headset removes the need for any attached connections, giving users a great deal of mobility. With a per-eye resolution of 1832x1920 and built-in audio output speakers, the device provides an immersive experience. Participants simply had to set up the room setup once, enabling them to get started right away. The only extra step was to adjust the head strap for comfort. Both games featured Hand Tracking 2.0, the most advanced hand tracking technology available from Meta at the time of the research. This system performed well in reliably tracking hand motions, especially in difficult situations like fast gestures or short hand-to-hand contact.

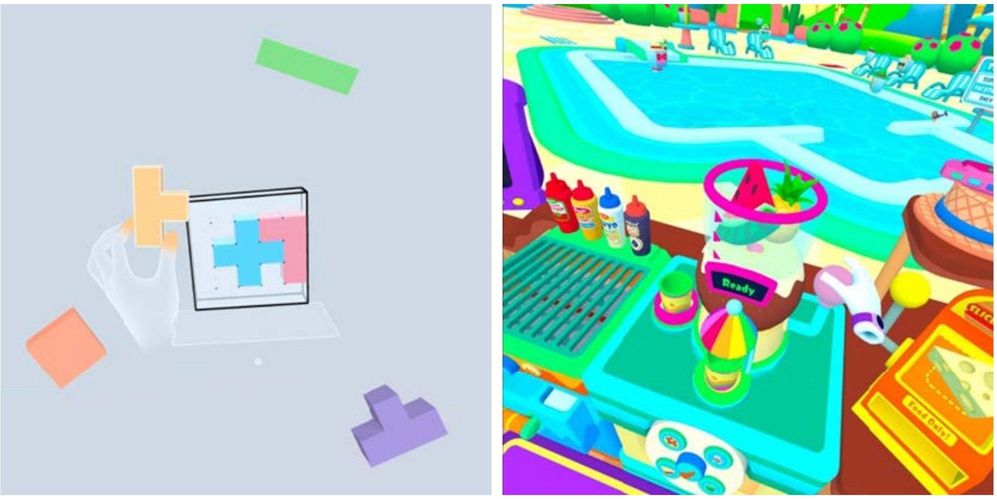
The research used a within-subjects design, which ensured that every participant encountered each condition. The combination of differed gaming durations, including both short and long sessions, and two control modalities (controllers and hand tracking), resulted in four separate experimental conditions with the order of conditions set by a Latin square layout. The study's procedure included a number of brief introduction sessions. Whenever participants began a new gaming session, they were given a short description of the setting and the tasks they were given, supported by illustrations such as screenshots. Similarly, control method adjustments were accompanied by short instructions that included controller use demos, button functionality, and explanations of hand tracking movements. Following each gaming session, participants were asked to answer a series of UX questions, finishing in a final post-questionnaire at the end of the research.

The demographics part of pre-questionnaire asked participants about their gender, age, employment, and self-assessment of their VR experience. Responses were scored on a scale of 1 ("not at all") to 5 ("very experienced"). The part of pre-questionnaire was the standardised ATI (Affinity for Technology Interaction) questionnaire. This questionnaire had nine questions, each with a 6-point Likert scale ranging from 1 ("completely disagree") to 6 ("completely agree"). The ATI score was calculated by taking the average of all item scores and adjusting for three items that were reverse-worded in compared to the others [4].

Following, several UX questionnaires were used to measure different ascepts of UX for each condition.

- The igroup presence questionnaire (IPQ) was used to assess presence, with questions scored on a 7-point Likert scale and altering anchor points [15].
- In addition, the Self-Assessment Manikin (SAM) was used to assess different emotional responses. SAM used a 5-point Likert scale and had a nonverbal, picture-oriented design, that included several sorted variants of an image reflecting different aspects. This questionnaire, well-established and notably concise, has found application across diverse contexts, offering an advantage when presented alongside a variety of questionnaires.
- The final UX questionnaire used was the Short User Experience Questionnaire (UEQ-S), which is a simplified version of the User Experience Questionnaire (UEQ). This condensed version reduced the original 26-item collection to only 8, while additionally lowering the number of categories from six to two: pragmatic and hedonic quality. The average score from both of these categories might be viewed as an overall assessment [11].
- Lastly, the Flow State Scale (FSS) was used to assess the state of flow, representing a comprehensive questionnaire with 36 individual items distributed across the 9 dimensions of flow [7].

In a subsequent post-questionnaire, participants were given the opportunity to express their willingness to recommend VR controllers and hand tracking to others. They could also provide reasoning for their decisions and offer feedback on the overall study experience.



(a) Cubism **(b) Vacation Simulator**

Figure 1: Screenshots of games: a) A scene from Cubism depicting a partially solved puzzle with some shapes inserted, one in the player’s left hand and additional shapes floating in the play space, b) A scene from Vacation Simulator: Back to Job depicting the player pulling a lever on a stylized blender that’s filled with fruits while standing in a kitchen with a pool in the background.

2.1 Participants

The study included a total of 20 participants, recruited via the institution’s online portal for test subjects. Within the sample of 20 individuals, 7 identified as female and 13 as male. In particular, 65% of the participants identified as students, which matches our predictions given the recruiting methods. The participants’ average age was 28.65 years, with the youngest being 20 years old and the oldest being 57 years old, for a standard deviation of 9.25. Additionally, the mean level of VR experience, measured on a scale of 1 ("not at all") to 5 ("very experienced"), was 2.50, with a standard deviation of 0.95.

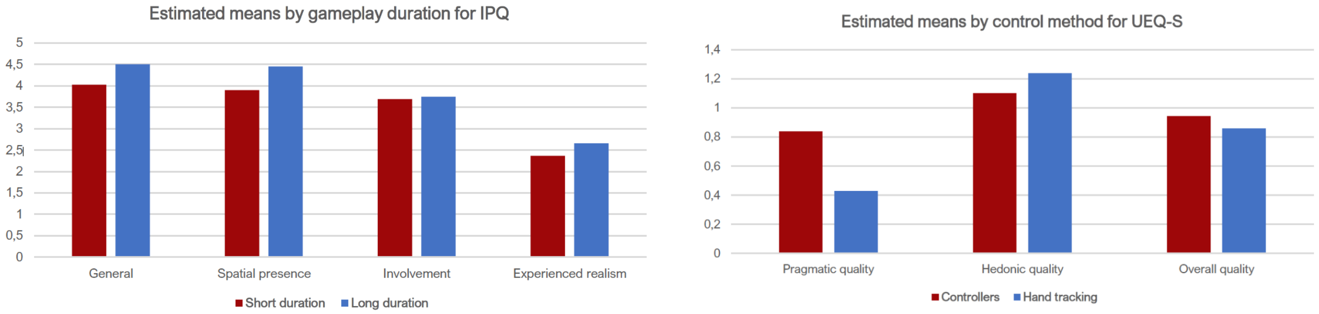
3 RESULTS

To analyze the collected questionnaire data, two separate statistical approaches were used. The results from the UX questionnaires (IPQ, SAM, UEQ-S, FSS) were analysed using a two-way repeated measures analysis of variance (ANOVA). This statistical test determines if two variables have a statistically significant interaction impact on a continuous dependent variable. To reduce the probability of type I errors, Bonferroni correction was used. The ANOVA’s assumption of normality was tested using the Shapiro-Wilk test, which is considered a more trustworthy technique than utilising raw data [10]. It is worth noting that non-normal data was obtained; yet, ANOVA is often recognised as resilient in the face of departures from the normality assumption. The research results show that there are no significant effects on type I error rates [2], confirming the validity of using ANOVA even in the absence of strict normality [13]. For the remaining dataset, including the ATI score, the VR experience rating, and the hand tracking recommendation, a binomial logistic regression analysis was performed.

3.1 Presence

Significant differences were revealed across all dimensions of the IPQ questionnaire when comparing different gameplay duration conditions. Before proceeding with the analysis, the original 1/7 scale used in the study was converted to a 0/6. Figure 2 provides a complete overview of how gaming duration affects the IPQ dimensions.

In terms of general presence, gaming time had an effect, with a significant difference between short and long durations (F(1,19) = 7.006, p =.016, partial η^2 =.269). Overall, the long duration condition (4.500±0.185) was reported to make users feeling more in presence compared to the short duration condition (4.025±0.225), with a mean difference of 0.475 (95% CI, 0.099 to 0.851). In the context of spatial presence, a statistically significant main effect of gameplay duration was found as well (F(1,19) = 19.413, p = .001, partial η^2 = .505). Spatial presence increased significantly in the long duration condition (4.450±0.142) compared to the short duration condition (3.895±0.184), with a mean difference of 0.555 (95% CI, 0.291–0.819). Involvement was significantly higher when using controllers (3.925±0.249) than hand tracking (3.450±0.300) for short durations (F(1,19) = 7.228, p =.015), with a mean difference of 0.475 (95% CI, 0.105 to 0.845). However, long-term participation with controllers (3.738±0.280) did not show a statistically significant difference from hand tracking (3.750±0.282) (F(1,19) = 0.005, p =.944). In the scope of experienced realism, the main effect of gameplay duration produced statistical significance (F(1,19) = 4.367, p = .050, partial η^2 = .187). Experienced realism increased significantly in the long length condition (2.656±0.237) compared to the short duration condition (2.363±0.237) where users due to less time to play have felt environment is less realistic, with a mean difference of 0.294 (95% CI, 0 to 0.588).



(a) IPQ and Gameplay Duration

(b) UEQ-S and Control Methods

Figure 2: Overview of results: a) Chart depicting estimated marginal means by gameplay duration for IPQ dimensions, b) Chart depicting estimated means by control method for UEQ-S dimensions.

3.2 Pragmatic Quality

Pragmatic quality, one of the dimensions of the UEQ, showed significant differences in this research, resulting in it being the only dimension within the UEQ to show statistical significance. Pragmatic quality evaluates a system or interface's usefulness, efficiency, and usability from the standpoint of the user. The study's 1/7 scale was transformed into a −3/+3 scale prior to analysis.

The main effect of control method showed a statistically significant difference between controllers and hand tracking (F(1,19) = 6.252, p =.022, partial η^2 =.248). The study found that controllers (0.938±0.156) outperformed hand tracking (0.569±0.130) in terms of pragmatic quality, with a mean difference of 0.369 (95% CI, 0.060 to 0.677). Figure 2 shows an overview of how control methods effected the UEQ-S dimensions.

3.3 Clear Goals, Concentration and Sense of Control

The study of the Flow State Scale (FSS) data revealed significant results in important elements of the flow experience. Particularly, participants' assessments for specific goals, attention, and sense of control changed significantly throughout the analysed activities, giving insight on these important features of the flow state.

The study revealed a statistically significant difference in the main impact of gameplay length (F(1,19) = 7.404, p =.014, partial η^2 =.280) on the dimension associated with specific goals. Clear goals were significantly greater during short length sessions (4.488±0.109) than long duration sessions (4.019±0.163), with a mean difference of 0.469 (95% CI, 0.108 to 0.829). When it comes to the focus on the task, the main effect of control method was statistically significant (F(1,19) = 5.000, p =.038, partial η^2 =.208), indicating that controllers (4.475±0.109) were significantly more focused on the task at hand than hand tracking (4.350±0.124), with a mean difference of 0.125 (95% CI, 0.008 to 0.242). The main effect of gaming time was statistically significant (F(1,19) = 6.491, p =.020, partial η^2 =.255) in terms of the feeling of control dimension. Short length sessions resulted in a stronger sense of control (4.238±0.102) compared to long duration sessions (3.813±0.196), with a significant mean difference of 0.425 (95% CI, 0.076-0.774). The main effect of control method was statistically significant (F(1,19) = 15.073, p =.001, partial η^2 =.442). Additionally, controllers were reported to have a significantly higher sense of control (4.281±0.121) than hand tracking (3.769±0.170), with a mean difference of 0.513 (95% CI, 0.236 to 0.789).

3.4 Previous VR experience

In terms of participants' past VR experience, a binomial logistic regression was used to determine its impact on the probability of recommending hand tracking technology. In this respect, the model explained 26.1% of the variation in hand tracking suggestions while correctly classifying 74.1% of instances. It is worth mentioning that the predictor variable, VR experience, was statistically significant (p =.042). This finding points out to a relationship: as individuals' levels of VR experience increased, their tendency to recommend hand tracking to others showed a significant rise.

4 DISCUSSION

The study's findings provide intriguing insights into how gaming time and control approaches affect VR user experience. While controllers received recognition for their precision and ease of use, hand tracking received mixed reviews, with some appreciating its inventive potential but others criticising its current technological limits.

4.1 Feedback on control methods

Participants provided their opinions on the control methods and hand tracking through a concluding survey, after testing each for approximately 12 minutes. The controllers received equally positive feedback, including recognition for their precision and reliability. Participants rated it easier to use, with many noticing that grabbing items in VR felt very comparable to real-world interactions. In contrast, evaluations on hand tracking differed. It was recognised as an innovative and exciting technology that provided a better level of immersion and realism by allowing users to see their hands in the virtual world. However, it was stated that the technique required additional refinement. There were issues with its accuracy and the strange feeling caused by delay. The need to keep hands inside the camera's view was criticised, and the lack of actual sensation when grabbing digital objects. The tactile sense provided by controller buttons was preferred to the absence of feedback in hand tracking. Observations throughout the study revealed challenges with hand-to-hand interactions and the cameras' narrow field of vision, resulting in unpredictable motions when hands went out and then back into the monitored region. Hand tracking also did not work with precision activities like turning items, which significantly impacted the gameplay experience.

4.2 User Experience Insights

The data mainly showed the independent effects of gaming duration and control method on several metrics, with significant effects detected. Longer gaming durations increased measures of presence and realism, indicating that prolonged VR experiences improve the perception of being in a real environment. Surprisingly, despite user feedback, hand tracking had no equivalent effect on perceived realism. The study noted differences in clarity of objectives and control sensation between short and long gameplay sessions, potentially influenced by the nature of the games used. Controllers were shown to considerably improve task attention, probably due to experience with comparable gaming gadgets and their inherent reliability. The lack of tactile feedback in hand tracking has a negative effect on user experience, highlighting the advantages of controllers for replicating realistic interactions. Gameplay duration also played a significant role in immersion and presence, with longer sessions resulting in better outcomes.

General presence, spatial presence, and experienced realism, all of which were measured using the IPQ, were statistically significantly higher for the long gameplay duration. This could indicate that spending more time in a VR experience helps to convince the player of being a part of a real environment, instead of just playing a game. Interestingly, hand tracking did not show a comparable effect on the experienced realism, even though multiple participants explained that the control method felt more real in the post-questionnaire. However, the goals seemed clearer, and the sense of control was statistically significantly improved for the short gameplay duration, as measured using the FSS. This could be ascribed to the nature of the two games. Vacation Simulator is a bit more expansive compared to Cubism, which may have had an effect on the Clear goals dimension. Additionally, Vacation Simulator requires the use of two hands for some scenarios, which perhaps impacted the sense of control due to players typically using just a single hand for Cubism as observed during the study. This is not in line with another study that indicated higher flow for the longer duration. However, that experiment also labelled 2 minutes as the short duration and 5 minutes as the long duration, as opposed to 3 and 9 minutes here [20].

Hedonic quality however was not impacted by gameplay duration or control method, even though controllers were objectively and subjectively worse. Both were measured using the UEQ-S. Sense of control was statistically significantly higher for controllers as well, in addition to the effect caused by the gameplay duration, which probably also stems from the reliability discrepancy and was measured using the FSS. Participants commented that hand tracking does not feel natural at all due to the delay between moving one's hands and seeing the result in VR. The average temporal delay for hand tracking is a significant 38.0 milliseconds [1]. Measured using the IPQ, involvement showed a two-way interaction effect and was higher for controllers for the short gameplay duration. This could again have been influenced by the controller's superior reliability.

4.3 VR Experience Level and ATI Score

The study also aimed to investigate the relationship between participants' VR experience and their willingness to suggest hand tracking. Interestingly, people with greater VR expertise were more likely to recommend hand tracking, contrary to our hypothesis. This might indicate that experienced users are more willing to accept the limits of existing VR technology. Participants with the least experience reported regular problems with hand tracking, but those with the most experience did not, indicating a better tolerance or acceptance. The technology interaction affinity (ATI) score had no significant effect on hand tracking suggestions, indicating that other characteristics were not evaluated.

4.4 User Study Limitations

A critical limitation of the study was the use of different games for varying gameplay durations, driven by the impracticality of developing a custom VR application. The games were chosen based on compatibility with both control techniques, ethical acceptability, and beginning accessibility, resulting in the choice of using different games for short and long sessions. This provided a variable that might influence the perception of findings, particularly regarding feeling of presence and realism due to different game design. The short-duration game Cubism and the long-duration game Vacation Simulator presented distinct experiences that might impact user feedback and performance measures, needing additional caution when using these outcomes.

5 CONCLUSION

In conclusion, our findings align with the expected results regarding the comparison between controllers and hand tracking within virtual reality (VR) environments. As assumed, the comparison of controllers and hand tracking revealed that controllers are the more rational choice in every way, at least for the specified VR games, corroborating previous findings [5]. This advantage is attributed to the controllers' more reliability, their perception of control, and users' increased task attention, all of which lead to a more immersive VR experience. However, it is important to note that previous study has shown that hand tracking can outperform controllers for some interactions [19], demonstrating a context-dependent preference that changes with the nature of the interaction and the users' experience with VR technology. The mixed findings point to a dynamic environment for VR interaction approaches, with the option between controllers and hand tracking potentially evolving as technology progresses and user experiences expand. Future research is encouraged to further explore these interactions and the potential shifts in user preference as the fidelity of hand tracking improves.

REFERENCES

[1] Diar Abdlkarim, Massimiliano Di Luca, Poppy Aves, Sang-Hoon Yeo, R Chris Miall, Peter Holland, and Joseph M Galea. 2022. A methodological framework to assess the accuracy of virtual reality hand-tracking systems: A case study with the oculus quest 2. *BioRxiv* (2022), 2022–02.

[2] M José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, and Rebecca Bendayan. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema, 2017, vol. 29, num. 4, p. 552-557* (2017).

[3] Gavin Buckingham. 2021. Hand tracking for immersive virtual reality: opportunities and challenges. *Frontiers in Virtual Reality* 2 (2021), 728461.

[4] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[5] Asim Hameed, Andrew Perkis, and Sebastian Möller. 2021. Evaluating hand-tracking interaction for performing motor-tasks in vr learning environments. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 219–224.

[6] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. 2020. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)* 39, 4 (2020), 87–1.

[7] Susan A Jackson and Herbert W Marsh. 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology* 18, 1 (1996), 17–35.

[8] Chaowanan Khundam, Varunyu Vorachart, Patibut Preeyawongsakul, Witthaya Hosap, and Frédéric Noël. 2021. A comparative study of interaction time and usability of using controllers and hand tracking in virtual reality training. In *Informatics*, Vol. 8. MDPI, 60.

[9] Panagiotis Kourtesis, Simona Collina, Leonidas AA Doumas, and Sarah E MacPherson. 2019. Validation of the virtual reality neuroscience questionnaire: maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Frontiers in human neuroscience* 13 (2019), 417.

[10] Marcin Kozak and H-P Piepho. 2018. What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *Journal of agronomy and crop science* 204, 1 (2018), 86–98.

[11] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*. Springer, 63–76.

[12] Rory McGloin, Kirstie Farrar, and Marina Krcmar. 2013. Video games, immersion, and cognitive aggression: does the controller matter? *Media psychology* 16, 1 (2013), 65–87.

[13] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15 (2010), 625–632.

[14] Bernhard E Riecke, Joseph J LaViola Jr, and Ernst Kruijff. 2018. 3D user interfaces for virtual reality and games: 3D selection, manipulation, and spatial navigation. In *ACM SIGGRAPH 2018 Courses*. 1–94.

[15] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 266–281.

[16] Mel Slater. 2018. Immersion and the illusion of presence in virtual reality. *British journal of psychology* 109, 3 (2018), 431–433.

[17] Statista. 2018. Virtual reality and augmented reality (VR and AR) devices average session time in the United States as of 2018. https://www.statista.com/statistics/831819/us-virtual-augmented-reality-device-average-session-time/. Accessed: 2024-1-30.

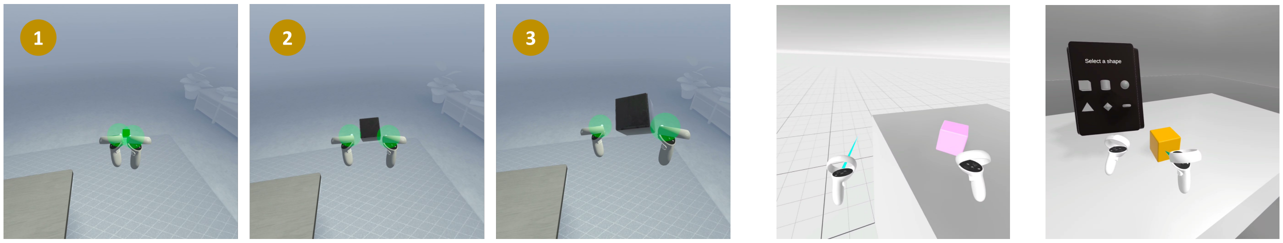
[18] Statista. 2024. Consumer and enterprise VR revenue worldwide 2026. https://www.statista.com/statistics/1221522/virtual-reality-market-size-worldwide/. Accessed: 2024-1-30.

[19] Jan-Niklas Voigt-Antons, Tanja Kojic, Danish Ali, and Sebastian Möller. 2020. Influence of hand tracking as a way of interaction in virtual reality on user experience. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–4.

[20] William G Volante, Jessica Cruit, James Tice, William Shugars, and Peter A Hancock. 2018. Time flies: Investigating duration judgments in virtual reality. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 1777–1781.

Designing with Two Hands in Mind?: A Review of Mainstream VR Applications with Upper-Limb Impairments in Mind

Caglar Yildirim
Northeastern University
Boston, MA, USA
c.yildirim@northeastern.edu



(a) Bimanual Manipulation

(b) Locomotion

(c) Menu

Figure 1: Common bimanual VR interaction tasks

ABSTRACT

Virtual reality (VR) has the potential to transform work, collaboration, and socialization in diverse settings. Nevertheless, most immersive interactions are inaccessible to individuals who use one hand, as their design assumes that VR users have simultaneous usage of two hands, excluding individuals who use hand from the VR community. It is also unclear the extent to which existing VR applications are accessible to individuals who use one hand. We, therefore, conducted a systematic review of mainstream VR applications for collaboration, productivity, and socialization to identify in what ways they support one-handed interactions. Our review showed that the assumption of bimanual input was pervasive in the design of VR tasks, that more than half the applications were inaccessible to individuals who use one hand, and that none of the applications supported customizations for physical disabilities. Our findings underscore the need for increasing access to VR by devising and supporting unimanual input paradigms for key VR tasks.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**.

KEYWORDS

VR accessibility, unimanual interaction, bimanual interaction, app review, upper-limb impairments, physical disability

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652224

ACM Reference Format:
Caglar Yildirim. 2024. Designing with Two Hands in Mind?: A Review of Mainstream VR Applications with Upper-Limb Impairments in Mind. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3652212.3652224

1 INTRODUCTION

Thanks to the advances in display and tracking technology, head-mounted display-based (HMD-based) virtual reality (VR) is becoming more and more immersive. With increased immersion comes increased reliance on the assumption that VR users have non-disabled bodies, however. The design of existing VR systems places an emphasis on non-disabled bodies and makes certain assumptions regarding hardware and interaction design from their perspective. For example, one such ableist assumption baked into the design of most VR interfaces and interactions is that all VR users have simultaneous usage of two hands and can complete VR interaction tasks requiring *bimanual input* (executed with two hands). This implicit assumption of the 'corporeal standard' leads to inaccessible VR systems and experiences that fail to cater to the needs of individuals with disabilities, rendering VR an ableist technology [4]. While certain VR experiences do have accessibility options (mostly addressing sensory disabilities), they are usually added after the experience has been designed. This accessibility-after-the-fact approach is antithetical to the basic tenets of human-centered design, since it fails to incorporate the needs and expectations of disabled users into the design of these interactive systems from the initial stages of hardware and software development.

Within the context of VR, bimanual interactions require users to use both controllers (or hands when hand-tracking is available) to perform canonical interaction tasks [9]. For instance, in existing VR experiences, object manipulation (scaling and positioning an object in 3D) is usually achieved using a bimanual metaphor, one example of which is illustrated in Figure 1a. As seen in the figure, the user begins by selecting an object with both controllers (image 1) and needs to move controllers apart to scale up the object (images 2-3). In this bimanual metaphor, the scale of the object is tied to the distance between the controllers (or tracked hands).Another example of a key task requiring bimanual interaction is locomotion (i.e., navigation in virtual environment) while interacting with an object. In most VR experiences, locomotion and object interaction tasks are assigned to different controllers, with one controller dedicated to locomotion and the other to object selection and manipulation. This is illustrated in Figure 1b, where the left controller is used for locomotion (with raycasting) and the right controller is used for interactions with virtual objects. Yet another example pertains to applications where menu access is essential and is performed in conjunction with object interaction. In this task, as illustrated in Figure 1c, the menu is tied to one controller, and users use the other controller to interact both with the menu (e.g., select a menu item) and with virtual objects. These examples illustrate common interaction tasks in VR and demonstrate the extent to which VR interactions rely on bimanual metaphors based on the assumption of VR users being able to use two hands. While these three examples are essential VR tasks and are pervasive in most VR experiences, it is not possible to perform these tasks for a user who can use one hand only.

In order for these users to have equitable access to VR experiences and their potential benefits, it is crucial for VR designers and developers to support unimanual (executed with one hand) interactions in VR applications. Yet, it is not clear the extent to which existing VR applications are accessible to individuals with mobility impairments. In this paper, we aimed to address this need by conducting a systematic review of popular VR applications, focusing on the extent to which they are usable by individuals who use one hand. We contribute the first review of mainstream VR applications (n = 16) from an accessibility lens for individuals with mobility impairments. Our review was guided by the following research questions:

(1) RQ1: To what extent do mainstream VR applications rely on the assumption of bimanual input?
(2) RQ2: To what extent is unimanual input supported in mainstream VR applications?

2 RELATED WORK

2.1 3D Interaction in VR

In virtual environments, users perform a variety of 3D tasks to interact with the VR system and to complete their desired goals. Broadly speaking, 3D interaction tasks can be categorized into object selection, object manipulation, navigation, and system control [1, 5]. Object selection involves identifying or acquiring an object or subset of objects among a larger set of objects. Common selection tasks include pointing to a target, pressing a button, and grabbing an object in the virtual environment. Object selection is usually achieved through directly pointing to and selecting an object in 3D space. It can also be achieved by indirect selection using raycasting, which involves casting a virtual ray to point to objects that are located beyond the area of reach. Regardless of which method is used, 3D selection tasks usually support unimanual input and do not require the simultaneous usage of two hands.

3D object manipulation refers to virtually handling objects, and common manipulation tasks include positioning, rotating, and scaling an object in 3D space [1, 5]. In most cases, it is possible to change the position of virtual objects using unimanual input. That said, rotation and scaling tasks are usually implemented using bimanual input and thus require the simultaneous usage of two hands.

3D navigation is an integral part of VR applications, as it is through locomotion that users explore the virtual environment and perform other 3D tasks. When real walking is not possible due to physical space constraints or desirable due to user preferences and needs, the teleportation metaphor is commonly used in existing VR applications, which instantly translates the user from one point to another by updating their 3D position vector to that of the destination often specified by raycasting. While teleportation can be completed using unimanual input, it often requires bimanual input when users are holding a virtual object in one hand, as illustrated in Figure 1b. In this usage scenario, users cannot perform navigation tasks in conjunction with other interaction tasks.

System control refers to the ways in which users communicate their intentions to the VR system [1, 5]. Common system control tasks include menu interactions and text entry. Most menu interactions in existing VR applications are designed with the bimanual input assumption in mind and cannot be completed using unimanual input alone. As for VR text entry, the majority of text entry tasks can be completed using unimanual input. However, based on the assumption of bimanual input, existing VR applications predominantly use a virtual keyboard using the QWERTY layout [12]. This typically translates into slower performance and increased physical demand if a user relies solely on unimanual input for these text entry tasks, as the design of the QWERTY layout assumes that users will provide bimanual input.

2.2 VR Accessibility for Individuals with Limited Mobility

Previous research has shown that individuals with limited mobility do have an interest in using VR applications for a variety of use cases, including entertainment, socialization, and productivity [6, 7, 10]. One example of these use cases is being able to experience physical world activities that are otherwise inaccessible to them (e.g., paragliding) [2]. Therefore, it is essential that VR applications be made accessible and usable for individuals with limited mobility. Addressing the lack of efforts to make VR gaming accessible for wheelchair users, Gerling et al. [3] developed three VR games based on a survey of needs and expectations of wheelchair users. In the design of the games, the researchers ensured wheelchair users could play the games while controlling their wheelchair. While wheelchair users were not directly involved in the design of these games, they were involved in the evaluation of the games. Gerling et al. found that wheelchair users enjoyed being able to play VR games while using their wheelchairs. Their findings also highlighted the importance of making controls flexible and adaptable so that wheelchair users can customize how they provide input based on their mobility limitation. Gerling et al.'s study [3] also demonstrated that by including disabled users in the design of VR games, these accessibility issues could be mitigated and that VR games that match the abilities of disabled users could be developed, enabling disabled users to take advantage of immersive capabilities of VR gaming.

Mott et al. [7] explored the challenges faced by individuals with mobility limitations when setting up and using VR systems, focusing on hardware-related challenges. They conducted semi-structured interviews with 16 individuals with mobility limitations and identified seven categories of accessibility issues. These include challenges associated with (1) setting up the VR system; (2) preparing VR peripherals; (3) donning and removing the VR headset; (4) managing cords; **(5) holding and using two motion controllers simultaneously**; (6) reaching and pressing the buttons on VR controllers; (7) keeping the VR controllers in view of the cameras on the VR headset. Of relevance to our proposed research plan, Mott et al.'s findings showed that the requirement of using two controllers at the same time would prevent individuals with mobility limitations from using these systems. This highlights the importance of moving away from the assumption of two-handed interactions in VR. While Mott et al. identified the hardware-related accessibility issues faced by individuals with limited mobility, their study provides no unimanual interaction solutions.

In the only study that focuses on the (in)accessibility of bimanual VR tasks for individuals with limited mobility, Yamagami et al. [11] proposed a taxonomy to facilitate the design of unimanual counterparts of bimanual interaction techniques. Their taxonomy categorized bimanual tasks into synchronous/asynchronous and coordinated/uncoordinated tasks. They did not provide any unimanual alternatives or involve users in the design of unimanual interaction techniques. That said, they conducted a video elicitation study with individuals with limited mobility to gather their feedback on some prototypes. Their findings highlighted the importance of providing customizable input techniques and underscored the need for devising unimanual interaction techniques for common VR tasks.

As the foregoing review indicates, there is a scarcity of research into accessibility challenges faced by individuals with limited mobility when using VR systems [4]. The few studies that addressed this growing need have shown that VR is largely inaccessible for individuals with limited mobility [3, 7]. It is clear that the simultaneous use of two controllers to interact with the immersive environment is a barrier to the physical accessibility of VR technology for individuals with limited mobility. Nevertheless, to what extent do existing VR applications consider this limitation and support unimanual input? That is precisely the question this review aims to answer, as outlined in the following sections.

3 METHOD

To better understand the current practices and to determine whether and how existing VR applications provide accessibility features supporting unimanual input, we conducted a systematic review of mainstream VR applications. We performed searches on Meta Quest Store [8], one of the mainstream stores for VR applications. We chose to focus on Meta Quest store, as it features VR applications that can be used on standalone, tetherless VR headsets. Prior research has shown that individuals with mobility impairments experience difficulty when setting up and using tethered VR systems [7]. We, thus, surmised that VR applications that run on standalone VR headsets would be more accessible to individuals with mobility impairments. In addition, we intentionally excluded VR games from this review because our focus was on VR applications for collaboration, productivity, and socialization. The exclusion of VR games from our review is a deliberate choice, grounded in our objective to concentrate on applications that foster collaboration, productivity, and socialization. While VR games are a prominent and popular segment of VR applications, their design and interaction paradigms often differ substantially from non-gaming applications. The insights gleaned from examining VR games may not seamlessly translate to applications centered on collaboration, productivity, and socialization, warranting their exclusion from this review.

On the Meta Quest store, we searched through all the applications. We filtered the search results by category and sorted the applications by popularity, which was supported when the review was conducted in Summer 2023. The initial filtering by category revealed a total of 55 VR applications for socialization (13), 3D design and collaboration (17), and office/work productivity (25). Given the scope of the review, we chose to focus on most popular VR applications. The popularity of these applications suggests a higher likelihood of their adoption and use, making them particularly relevant for assessing the state of one-handed accessibility in mainstream VR applications. The search revealed a total of 16 most popular VR applications for socialization (n = 8), 3D design and collaboration (n = 4), and office/work productivity and collaboration (n = 4), as listed in Table 1.

Table 1: The 16 Reviewed VR Applications

VR App	App Category
Horizon Worlds	Socialization
Horizon Workrooms	Office productivity and collaboration
RecRoom	Socialization
ShapesXR	3D design and collaboration
vSpatial	Office productivity and collaboration
Spatial	Socialization
Engage	Socialization
Noda	Office productivity and collaboration (ideation)
Vtime XR	Socialization
Gravity Sketch	3D design and collaboration
VRChat	Socialization
bigscreen	Socialization
MeetinVR	Socialization
Ribla Studio	3D design and collaboration
immersed	Office productivity and collaboration
Arkio	3D design and collaboration

Our review involved the author running the 16 identified applications on a Meta Quest 2 headset and going through the main tasks of a given application. While reviewing the applications, we paid particular attention to the previously mentioned four canonical tasks in VR: selection, manipulation, navigation, and system control (including UI and menu interaction and text entry). For each of these broader task categories, we identified whether bimanual input was necessary or whether the task could be completed using unimanual input. We also examined the accessibility settings available in the applications, focusing on customizations for accommodating physical disabilities by enabling unimanual input. Given our research objectives, our review did not focus on sensory disabilities (e.g., visual/hearing impairments). An application was deemed inaccessible to individuals with limited mobility if any of its key tasks could not be completed using unimanual input. For instance, Shapes XR required bimanual input for object manipulation and UI/menu interaction, with users having to use two controllers for scaling objects and to hold the menu in one hand and select items with another hand. Because these two tasks could not be completed with unimanual input, Shapes XR was marked as inaccessible to individuals who use one hand.

4 RESULTS

The results from the review of all 16 identified VR applications are summarized in Table 2. Of the 16 applications reviewed, only 5 (31.25%) were fully usable by individuals who use one hand, and 2 (12.5%) were partially usable (manipulation tasks could not be completed using unimanual input, but they were not the key tasks in the application). The remaining 9 applications (56.25%) were not usable without bimanual input, rendering them inaccessible to individuals who use one hand.

4.1 Prevalence of Bimanual Input Assumption

For selection tasks, the majority of the 16 reviewed applications (n = 14) supported selecting objects or menu options with either controller, whereas two applications, VRChat and Ribla Studio, designate the right controller for selection tasks.

In applications where 3D object manipulation (rotation and scaling) is applicable, namely ShapesXR, Spatial, Noda, Gravity Sketch, Ribla Studio, immersed, and Arkio, there is a tendency to default to bimanual input. In fact, all of these seven applications require the simultaneous use of two hands for 3D manipulation.

Of the nine applications where navigation is applicable, four (i.e., Horizon Worlds, RecRoom, Engage, and Arkio) assume bimanual input in that they designate different functionality to different controllers, with one controller being used for navigation and the other being used for interacting with objects and environment. This separation of functionality necessitates that VR users always use both controllers while using the application. If a user were to use only one controller, all the functionality assigned to the other controller would simply be unavailable. The remaining five applications (i.e., Spatial, Noda, VRChat, bigscreen, and MeetinVR) afford users the ability to use either controller for navigation tasks, because they tie navigation functionality to the same button(s) on both controllers.

When it comes to interacting with UIs and menus, bimanual input is commonly required, with half of the 16 applications necessitating it. The principle of separation of functionality is applied to these tasks, wherein the UI or menu is tied to one controller and users are expected to use the other controller to interact with the UI elements or menu items on the other controller, as illustrated in Figure 1c. For instance, popular 3D design and collaboration applications, such as ShapesXR and Gravity Sketch, operate based on this assumption, and it is not possible to change this control setup. In relation to text entry tasks, which represents another category of system control tasks, most applications use virtual keyboards supporting unimanual input.

4.2 Physical Accessibility Settings

The review revealed that none of the 16 reviewed applications provided any accessibility settings to support unimanual input for individuals with mobility impairments. While these applications featured some accessibility options for sensory disabilities, they failed to incorporate any options to accommodate the needs of individuals with mobility impairments. For instance, in applications where bimanual input was assumed by default, there was no option to switch to using one controller only. Neither was there a setting to change the coupling of key functionality to different controllers.

5 DISCUSSION

Our study addressed a gap in the VR accessibility literature by providing the first application review on the extent to which mainstream VR applications are accessible to individuals who use one hand. Results from our systematic review point to the importance of moving away from the assumption of bimanual input in VR applications and underscore the need for devising unimanual counterparts to key bimanual interaction metaphors. We discuss our findings in relation to the research questions that guided this review.

5.1 RQ1: To what extent do mainstream VR applications rely on the assumption of bimanual input?

Our review showed that more than half of the 16 reviewed applications (56.25%) relied on the assumption that all VR users can simultaneously use both controllers. Their interaction design reflected this implicit assumption in that most key VR tasks such as object manipulation, navigation, and UI/menu interaction required the use of both controllers. This directly translates into individuals who use one hand not being able to use these applications and benefit from what they have to offer. This points to the importance of incorporating the needs of individuals with mobility impairments into the design and development of VR interactions and of making VR interactions more accessible to individuals who have various mobility impairments [7, 11].

5.2 RQ2: To what extent is unimanual input supported in mainstream VR applications?

The review revealed that only five out of 16 VR applications (31.25%) were fully usable by individuals who use one hand. In these applications, one common pattern was that they were mostly stationary experiences. Another common aspect was that they supported some of the key interaction tasks on both controllers. Users could, for example, teleport in the virtual environment using either controller. That said, it should also be noted that most of these applications (four out of five) did not include any object manipulation tasks. The only application that included object manipulation tasks and was still rated as fully usable was Spatial, in which object manipulation was not a key aspect of the application itself (it was available for adding more interactivity to the application). Therefore, it is not

Table 2: Summary of VR Applications and Accessibility Features

VR App	Selection	Manipulation	Navigation	UI/Menu	Typing	A11y	Usable
Horizon Worlds	Either	NA	B	U	V	None	No
Horizon Workrooms	Either	NA	S	U	P	None	Yes
RecRoom	Either	NA	B	U	V	None	No
ShapesXR	Either	B	S	B	V	None	No
vSpatial	Either	NA	S	U	V	None	Yes
Spatial	Either	B	U	U	V	None	Yes
Engage	Either	NA	B	B	V	None	No
Noda	Either	B	U	U	V	None	Partially
Vtime XR	Either	NA	S	U	V	None	Yes
Gravity Sketch	Either	B	S	B	V	None	No
VRChat	Right	NA	U	B	V	None	No
bigscreen	Either	NA	U	U	V	None	Yes
MeetinVR	Either	NA	U	B	V	None	No
Ribla Studio	Right	B	S	B	V	None	No
immersed	Either	B	S	B	P	None	No
Arkio	Either	B	B	B	V	None	Partially

A11y refers to the presence of accessibility settings for physical disabilities.
B: bimanual input. **U:** unimanual input. **S:** stationary.
V: virtual QWERTY keyboard. **P:** Physical keyboard

possible to conclude that these applications truly supported unimanual input for all key interaction tasks. These findings highlight the need for devising unimanual counterparts of common bimanual VR tasks [11].

5.3 Design Implications for Accessible VR Interaction Design

Based on our review, we present the following design implications, in the hope that VR designers and developers will apply them to existing and new VR applications to increase their compatibility with the needs of individuals with mobility impairments. These design implications are drawn from the insights from the review and are in line with other valuable guidelines established in prior work by [3], [6] and [7].

Make it possible to customize physical accessibility settings. Our results showed that none of the 16 reviewed applications included a physical accessibility settings menu, where users could customize controller input options and choose unimanual input. VR designers and developers should urgently incorporate such settings to accommodate the needs of individuals with mobility impairments.

Provide bilateral support for control input. Our review showed that separation of functionality between the two controllers was a common design decision. While this may work for able-bodied users, it is inaccessible to individuals who use one hand, for it requires bimanual input. Therefore, VR designers and developers should provide the option to decouple the functionality from two controllers and combine them into one controller. This could be achieved by presenting this option in a dedicated physical accessibility settings menu.

One specific way in which bilateral control input can be accomplished is through assigning different tasks to different buttons on the same controller. Most modern HMD-based VR systems use controllers that have a joystick, a trigger button, a grip button, and two functionality buttons. One way in which VR designers and developers could take advantage of this universal control design is through designating the trigger button for selection task, the grip button for grabbing, the joystick for navigation tasks, the combination of grip and joystick for rotating objects, and the combination of primary button and joystick for scaling objects. This way a single controller would be sufficient to perform the canonical VR interaction tasks without the need for bimanual input. Regardless of which controller is used, the same functionality would be available.

Place contextual menus in the environment rather than on the controller. One common pattern in 3D design applications was the use of a contextual menu attached to the controller, which automatically requires bimanual input. To free up the controller and move away from the bimanual input requirement, VR designers and developers could instead leverage the spatial environment when placing the contextual menus. Rather than attach the contextual menu to the controller(s), VR designers and developers could integrate them spatially into the environment (e.g., the menu sits on a hovering platform in front of the user). This can also be presented as an option in physical accessibility settings so that they can choose how and where the contextual menu should appear.

Leverage toggling to switch between different interaction tasks. For VR applications where interaction tasks need not be performed simultaneously, it is possible to enable users to switch between different interaction tasks (navigation, manipulation, etc.) by toggling a designated option. For instance, for a 3D design application, a contextual menu tied to the controller could still be used when users are afforded the ability to press a designated button on the controller to switch to navigation mode. This would eliminate the need to require the use of two controllers, while at the same time ensuring that all functionality is still available to users.

5.4 Limitations and Future Work

In this review, we intentionally focused on a certain category of VR applications in the Meta Quest Store library. We identified the most popular applications for collaboration, productivity, and socialization and included them in the review. In a future review, we intend to include a wider selection of VR applications from multiple application stores (although most of the reviewed applications are available across all stores). Another point to consider is that we completely excluded VR games from the review. As noted before, there is a vast number of VR games available in the market, some of which may have interesting solutions for supporting unimanual input. In future work, we plan on expanding this review to include VR games, as well.

6 CONCLUSION

In this paper, we contributed the first systematic review of mainstream VR applications for collaboration, productivity, and socialization to ascertain the extent to which these applications are accessible to individuals who use one hand. Our review showed that the assumption of bimanual input was pervasive in the applications reviewed, that separation of functionality to different hands was common, and that accessibility options to accommodate physical disabilities were nonexistent in mainstream VR applications. Our findings highlight the urgent need for supporting unimanual interactions in VR. We hope that VR community will join the efforts to make VR more accessible and usable for individuals with mobility impairments.

REFERENCES

[1] Doug A Bowman, Jian Chen, Chadwick A Wingrave, John Lucas, Andrew Ray, Nicholas F Polys, Qing Li, Yonca Haciahmetoglu, Ji-Sun Kim, Seonho Kim, et al. 2006. New directions in 3d user interfaces. *International Journal of Virtual Reality* 5, 2 (2006), 3–14. https://doi.org/10.20870/IJVR.2006.5.2.2683

[2] Lilian de Greef, Meredith Morris, and Kori Inkpen. 2016. TeleTourist: Immersive telepresence tourism for mobility-restricted participants. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 273–276. https://doi.org/10.1145/2818052.2869082

[3] Kathrin Gerling, Patrick Dickinson, Kieran Hicks, Liam Mason, Adalberto L Simeone, and Katta Spiel. 2020. Virtual reality games for people using wheelchairs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3313831.3376265

[4] Kathrin Gerling and Katta Spiel. 2021. A critical examination of virtual reality technology in the context of the minority body. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3411764.3445196

[5] Joseph J LaViola Jr, Ernst Kruijff, Ryan P McMahan, Doug Bowman, and Ivan P Poupyrev. 2017. *3D user interfaces: theory and practice*. Addison-Wesley Professional.

[6] Liam Mason, Kathrin Gerling, Patrick Dickinson, Jussi Holopainen, Lisa Jacobs, and Kieran Hicks. 2022. Including the Experiences of Physically Disabled Players in Mainstream Guidelines for Movement-Based Games. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[7] Martez Mott, John Tang, Shaun Kane, Edward Cutrell, and Meredith Ringel Morris. 2020. "I just went into it assuming that I wouldn't be able to have the full experience" Understanding the Accessibility of Virtual Reality for People with Limited Mobility. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13. https://doi.org/10.1145/3373625.3416998

[8] Meta Quest. 2023. *Oculus Store*. https://www.oculus.com/experiences/quest/.

[9] Sebastian Ullrich, Thomas Knott, Yuen C Law, Oliver Grottke, and Torsten Kuhlen. 2011. Influence of the bimanual frame of reference with haptics for unimanual interaction tasks in virtual environments. In *2011 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 39–46.

[10] Johann Wentzel, Sasa Junuzovic, James Devine, John Porter, and Martez Mott. 2022. Understanding How People with Limited Mobility Use Multi-Modal Input. In *CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3491102.3517458

[11] Momona Yamagami, Sasa Junuzovic, Mar Gonzalez-Franco, Eyal Ofek, Edward Cutrell, John R Porter, Andrew D Wilson, and Martez E Mott. 2022. Two-In-One: A Design Space for Mapping Unimanual Input into Bimanual Interactions in VR for Users with Limited Movement. *ACM Transactions on Accessible Computing* (2022). https://doi.org/10.1145/3510463

[12] Caglar Yildirim. 2023. Point and select: Effects of multimodal feedback on text entry performance in virtual reality. *International Journal of Human–Computer Interaction* 39, 19 (2023), 3815–3829. https://doi.org/10.1080/10447318.2022.2107330

An Extensible Architecture for Recognizing Sensory Effects in 360° Images

Raphael Abreu
MidiaCom Lab, Fluminense Federal University (UFF)
Niterói, Brazil
raphael.abreu@midiacom.uff.br

Joel Dos Santos
Multimedia Research Group, CEFET/RJ
Rio de Janeiro, Brazil
jsantos@eic.cefet-rj.br

Débora C. Muchaluat-Saade
MidiaCom Lab, Fluminense Federal University (UFF)
Niterói, Brazil
debora@midiacom.uff.br

ABSTRACT

The use of 360° content with sensory effects can enhance user immersion. However, creating such effects is complex and time-consuming as authors must annotate the spatial position (i.e.,´ ´origin of the effect”) in 360°. To tackle this mutimedia authoring issue, this paper presents an extensible architecture to automatically recognize sensory effects in 360° images. The architecture is based on a data treatment strategy that divides multimedia content into several manageable parts, operates on each part independently, and then joins the responses. The proposed architecture is capable of taking advantage of the diversity of recognition solutions and adapting to a possible author configuration. We also propose an implementation that provides three effect recognition modules, including a neural network for locating effects in equirectangular projections and a computer vision algorithm for sun localization. The results offer valuable insights into the effectiveness of the system and highlight areas for improvement.

CCS CONCEPTS

• **Applied computing** → **Hypertext / hypermedia creation**; • **Information systems** → **Multimedia information systems**; • **Computing methodologies** → Object recognition.

KEYWORDS

Sensory effects, automatic recognition, multisensory experiences, mulsemedia authoring

ACM Reference Format:
Raphael Abreu, Joel Dos Santos, and Débora C. Muchaluat-Saade. 2024. An Extensible Architecture for Recognizing Sensory Effects in 360° Images. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3652212.3652213

1 INTRODUCTION

The term “immersion” has been widely used to describe the quality of multimedia content such as movies, games, and presentations. Immersion refers to the user's perception of being physically present in a virtual world [13]. With the rise of head-mounted displays (HMDs), 360° multimedia content has become more prevalent, with a strong correlation between 360° experiences and the user's perception of immersion [18]. The industry has focused primarily on improving audiovisual quality to increase immersion. However, other senses beyond hearing and sight also play a role in perceiving and interacting with the world. 360° Multisensory or Mulsemedia (from *Multiple Sensorial Multimedia*) refers to 360° multimedia content combined with stimuli that engage additional senses such as wind, fog, heat, etc. Sensory effects have been shown to improve the quality of experience (QoE) [5] and sense of presence [10].

The authoring of 360° multimedia applications with sensory effects is a two-step endeavor where the mulsemedia author first identifies the sensory effects, i.e., identifies their spatio-temporal location in the audiovisual content and then annotates it with metadata describing their occurrence. This process, done on the mulsemedia authoring tool, can be challenging, costly, and prone to errors. In the context of addressing these limitations, the work of Amorim et al.[7] explores the use of crowdsourcing as a method to author coherent sensory effects associated with video content. To facilitate the authoring process, several studies [1, 17, 19] are focused in to create multimedia content analyzing algorithms into authoring tools to aid the identification of sensory effects, which we call sensory effect recognition in this paper. However, limitations such as the difficulty of annotating sensory effects and the subjectivity of the authoring process still exist [6].

The common approach for automatically recognizing sensory effects in multisensory media is to train a DNN (Deep Neural Network) to identify them [6]. However, this integration is not simple due to a few key challenges. Firstly, there is a wide variation in DNN architectures for content recognition, leading to different input modalities and outputs for the same content. The mulsemedia authoring tool must be able to handle these varying outputs, but this raises a problem of subjectivity, as the authoring of sensory effects is based on personal preferences and artistic decisions. Training a recognition method like a DNN can actually hinder this process, as it may not align with the author's preferences and retraining the network would be time-consuming.

Each DNN is is limited to the scenarios it was trained on, making it crucial for the network to return descriptive labels. However, standards for relating these labels to sensory effects are lacking, and label selection is influenced by various factors. In addition, deciding the the placement of sensory effects in 360° images is complex and requires the author's sense of combining sensory effects. Disambiguation is necessary in some cases, as sensory effects may overlap. For instance, if a wind effect is coming from the left and

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04
https://doi.org/10.1145/3652212.3652213

a flower scent from the right, the author must decide whether to render the aroma or not.

The challenges mentioned above motivates the integration of multiple DNNs into a single tool that would enable the author to not only specify the effects to be recognized but also determine how these effects should be combined to create a uniform description. This paper proposes an extensible 360° image processing architecture for recognition of sensory effects to address these issues. To perform this, The primary contribution of this research is an architecture that allows for interoperability with multiple recognition methods. Moreover, we propose the implementation of a mechanism to control and analyze recognition method results, aiming to produce a combination of the recognized sensory effects.

The rest of the paper is organized as follows. Section 2 presents concepts related to mulsemedia authoring, methods to perform sensory effect annotation using deep learning techniques, and related work that delve into sensory effect recognition. Section 3 presents an overview of the proposed architecture and each of its components. Section 4 implements this architecture in the task of locating sensory effects in 360° images and validates with a use case the integration of recognition modules. Section 5 concludes this paper by presenting future work and research directions.

2 RELATED WORK

In order to recognize sensory effects, there are two main approaches based on content analysis. The first approach [17] involves training an algorithm to specifically identify sensory effects in media content and return them as labels. For instance, the trained algorithm may associate the red color with the heat sensory effect, resulting in the "heat" label being returned as an annotation of sensory effects for that content. For 2D content, some studies have utilized DNN architectures for recognition. Siadari *et al.* [17] present a DNN framework for classifying sensory effects in videos, identifying activation moments of four effects: movement, vibration, wind, and flash. Zhou *et al.* [19] use a combination of DNN methods to detect sensory effect activation times and predict accompanying rendering attributes. Abreu *et al.* [1] build a DNN architecture that leverages both audio and video information to infer activation times of sensory effects, identifying effects such as explosions, wind, thunder, rain, and gunshots. All of these methods employ the first approach, which aims to completely identify sensory effects in the media content.

In previous work, we presented a second approach to recognize sensory effects that involves the use of an algorithm to output generic labels from the audiovisual scene, such as *sun*, *water*, and *trees*, and relate them to the activation of sensory effects [2, 6]. This current paper expands upon the existing approach to encompass the recognition of 360° images and to enable the utilization of multiple algorithms in combination.

In search of DNN for sensory effect recognition in 360°, we made a search in the Google Scholar database with the following string: *(sensory effects OR 4D effects) AND 360.* The first 100 returned studies were selected. In this set, there was no mention of automated authoring of sensory effects in 360° content. Thus we performed a broader search to find work that utilizes DNN content recognition with 360° in search for recognition tasks similar to the ones being presented in this paper. Query results preset DNN capable of locating objects in equirectangular images [4] or predicting where the user should be looking in the 360° content [14]. Therefore, it is clear that although there is no work with the recognition of sensory effects in 360° content, some recognition methods can already be used in this context. As far as we are concerned, this work presents the first proposal of an extensible architecture to recognize the location of sensory effects in 360° images.

Apart from 360° content, several models specifically tailored to detect objects are also present in the literature and can be integrated into our architecture. One prime example is the detection of weather events. Zhu et al. [20] presents a machine learning solution that can detect extreme weather divided into four classes (sunny, rainstorm, blizzard, and fog). Another context is the detection of specific types of plants and associating them with the corresponding aroma.In adittion, Dias et al. [8] presents several models capable of detecting fifty species of plants and cultivars commonly cultivated in Brazil and also worldwide.

3 PROPOSED ARCHITECTURE

As discussed in Section 2, there are several ways to perform content recognition and associate it with sensory effects. The core challenge we address with our proposed architecture is the need to harness the power of multiple object detection methodologies while respecting the diverse preferences of the mulsemedia author. However, achieving a one-size-fits-all solution for comprehensive detection is a formidable task, given the variations introduced by different image contexts, such as animations or low-light conditions. Consequently, we propose an innovative approach that prioritizes flexibility. Our architecture empowers authors to leverage a wide array of machine-learning methods and tools. This approach allows them to tailor the architecture to their specific requirements, mixing and matching techniques to achieve the desired results. Figure 1 presents a high-level view of the proposed architecture. Each component of this architecture will be explained in detail in the remainder of this section.

The proposed architecture is composed of three basic components: *Map*, *Inference Module* and *Filter and combine*. There may be an extensible number of inference modules that encapsulate a content recognition algorithm. Each inference module implements three inner components (*pre-process*, *post-process* and *labels2Effects*) responsible for handling the recognition algorithm and converting to sensory effects annotations. This coupling of modules allows integration with any method for media content recognition. Finally, the responses from the modules are passed to the *Filter and combine*. There the responses will be combined and possibly complemented for a better decision on the specification of the sensory effect type and rendering characteristics.

Map. Its purpose is to read the audiovisual content and distribute it to the modules, according to the input needed by each module. The Map component can run all the loaded modules and their respective inputs. Furthermore, natively it provides a series of possible transformations to be applied to the audiovisual content to suit the modules' input. We chose to leave these transformations in the Map component to avoid an unnecessary amount of conversions in each inference module. This is evident by assuming a 360°

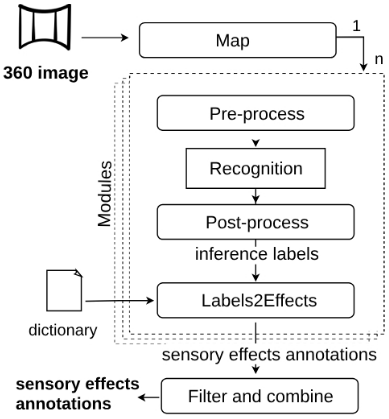


Figure 1: Extensible architecture for SE recognition

video input and several modules requiring only audio. To prevent each module from converting, the Map component performs only one conversion and provides the data for each module. The Map component is also responsible for passing configuration parameters to modules through a JSON file. For example, a module might need a specific API key to access a web service.

Pre-process. This component is the entry point for each module. Although the Map has already made a preliminary adaptation of the multimedia input, this step also aims to adapt the multimedia content received to the recognition algorithm used in the module. For example, in a module that uses a DNN for audio recognition, it may be necessary to separate the audio into 10s chunks for recognition, as done by [6, 15].

Recognition. This is the third-party component that performs the execution of the recognition algorithm. Therefore, its implementation details are outside the scope of the architecture. However, a requirement a third-party recognition component to be used in the proposed architecture is to return a set of labels that describe the multimedia content. As seen in Section 2, common strategies for the Recognition component are using deep learning methodologies or classical computer vision approaches.

Post-process. This component aims to adapt the output of the recognition module to a spatial representation consistent with the one needed for sensory effect annotation (e.g., spherical coordinates). It also applies necessary conversions and corrections in the returned data, besides its aggregation when several recognition module calls were made. In that case, individual call responses are aggregated to generate a single *label* response to describe the whole content.

Labels2Effects. As discussed in the earlier sections, it is essential to allow a configuration of which labels can represent which sensory effects. Since each recognition method may follow a different label nomenclature, one must know in advance the possible labels used by the specific recognition method. Therefore, this component receives a the set of labels recognized by the recognition algorithm, and correlates them to the chosen types of sensory effects. Each module must be accompanied by a dictionary of labels related to sensory effect types that will be passed to this *labels2effects*. The dictionary correlates labels and a sensory effect type and its presentation characteristics as illustrated in Listing 1. Output examples of *labels2effects* components are given in Listing 2. The dictionary approach solves two problems related to authoring sensory effects using content recognition. The first is to allow greater control over which labels represent the sensory effect types according to the author's possible preferences. The second is to enable extensibility and interoperability of the tool, as the file can be adapted to match any labels that are returned by the chosen recognition method.

Listing 1: Labels to effects dictionary for the scene understanding module

```
...
"Heat": {"sun": {"intensity": 40}, "summer": {
    "intensity": 30}},
"Aroma": {"flower": {"intensity": 20, "
    aromatype":"flower"}, "garden": {"
    intensity": 40,"aromatype":"flower"}, "
    tree": {"intensity": 40,"aromatype":"tree"
    }}
```

Filter and combine. Individually thee modules can be interesting on their own. However, they can be used to inform the construction of more sophisticated hierarchical rules. Such a strategy is based on the notion of cooperative data fusion, in which several independent modules recognize information to provide information that would not be available in the individual modules [9]. This component, therefore, completes the proposed architecture by enabling the author to define rules to fuse recognized information into the final sensory effect annotation for the multisensory application. This component oversees the sensory output of all modules, resolving conflicts and optimizing synergies at a higher abstraction level.

The application of fusion rules can follow the same principles of XSLT transformations [3], and more recently for JSON transformations[1]. The transformations can be a set of template rules that associate a data pattern of sensory effects coming from the modules. Then the application of those rules can transform this data based on the preferences of the author. As an output, it will construct a new set of sensory effects that may have been combined, filtered, or restructured according to those rules. The final output of the architecture comprises annotations of sensory effectsin the 360° image. The following section presents an implementation of this architecture for inferring sensory effects as to their placement, type, and intensity using 360° images.

4 PROPOSAL VALIDATION

To validate our proposal, we implemented an extensible architecture comprising three recognition modules for 360° image content. This discussion will explain how to construct such a module and present the three modules afterwards. Before constructing a module, a crucial preliminary step is to determine the content recognition algorithm that aligns with the desired functionality. This could involve deep learning models, classical computer vision techniques, or any other suitable method for the media context. Then the developer must construct the module's steps and integrate them into the

[1] https://github.com/bazaarvoice/jolt

architecture. The steps involved in building the module include: (i) Developing the Pre-process component to adapt the multimedia input for accurate recognition. (ii) Integrating the chosen recognition algorithm into the Recognition component to output meaningful labels. (iii) Designing the Post-process component to handle label conversion and rearrangement for spatial representation in 360° or address any labeling inconsistencies. (iv) Developing the Labels2Effects component to correlate recognized labels with predefined sensory effect types. Finally, (v) returning the recognized sensory effects to the map component.

In our implementation, each inference module encapsulates a specialized recognition component for different purposes. The first module is called *Effect localization* and uses a convolutional neural network for object detection. The second module is called *Scene understanding* and uses a public API capable of recognizing concepts in the visual content. The third module is called *Sun localization* and uses computer vision for localizing the sun in a 360° image.

Together with the three modules, we instantiate the Map component to have an equirectangular image as input. The Map component then converts this input to a cubic projection image for the *Effect localization* module and just relays the equirectangular projection image to the *Scene understanding* and *Sun localization* modules. The following sections explain in detail the process of creating and running each module. Finally, Section 4.4 describes the Filter and combine component instantiated to combine the three modules' output.

4.1 Effect localization module

The *Effect Localization* module instantiates the proposed module architecture to locate sensory effects in 360° images. Having received a cubic projection image as input (from the Map component), this module's pre-processing component calls the recognition component for each face of the cubic projection.

The recognition component uses YOLO V3, a CNN-based object localization network architecture, trained on the Google OpenImages dataset [12, 16]. This dataset includes 14.6 million bounding boxes for 600 objects in 1.74 million images, covering a wide range of object labels, such as animals, clothing, vehicles, food, and more. YOLO identifies objects in a projection and returns a set of labels and bounding boxes for those objects.

This module's post-processing component combines the responses for each face of the cube, converts these responses back to the equirectangular projection and then to latitude and longitude coordinates. That conversion is performed for visualization on an HMD. This representation follows the proposal of Josué et al. [11].

The post-processing component outputs a list of labels, detection reliability, and the four coordinates that specify the corners of the bounding box. The *Labels2Effects* component analyzes that list to decide which labels should become sensory effects and their rendering attributes. For example, a *plant* label may be converted to a tree aroma effect with 50% intensity. An example of output from the Effect localization module is shown in Listing 2.

Listing 2: Return from the effect localization module

```
{'effect': 'aroma','type': 'tree','intensity':
     50,'location': [(-93,-43),(-45,-34),(-93,
     11),(-45,8)]}
```

4.2 Scene understanding module

The *Scene understanding* module is a part of a system that identifies sensory effects in a 360° environment, without specifying their locations. It associates the concepts present in the visual content, represented by the labels from a neural network output, with sensory effects. Having received an equirectangular image as input, this module's pre-processing component decreases the image size and performs a call to the recognition component.

This module uses as its recognition component a cloud-based neural network API, specifically the *General* recognition model from Clarifai[2]. It returns more than 11,000 descriptive scene labels, including *travel*, *beach*, *architecture*, *tree*, *sky* and *sun*. No neural network training was required to use it.

Since the recognition component returns labels to the whole image, there is no post-process of labels. The *Labels2Effects* component converts the labels from Clarifai to sensory effects using a predefined dictionary (Listing 1), where labels are associated to sensory effect types and an initial intensity. The conversion also sets the specific type of the effect for those that have characteristics such as aroma. An example of the conversion of labels to aroma and heat effects.

4.3 Sun localization module

The Sun Localization module uses a classical computer vision algorithm to identify the sun in an image by finding the brightest pixels. Having received an equirectangular image as input, this module's pre-process component is simply a call to the recognition component.

The recognition component is implemented using the *OpenCV* library, which returns a bounding-box indicating the location of the sun. However, this approach has limitations, such as the inability to work when the sky is cloudy or there are multiple light sources, which could be addressed in future work by improving the recognition method.

The post-processing component converts the bounding-box tag to latitude and longitude coordinates that conform to the spherical coordinates system. Equation 1 is used to convert each of the (x,y) points to spherical coordinates based on the size of the input image.

$$lat = \frac{y * 180}{height - 90} \; lon \quad = \frac{x * 360}{width - 90} \tag{1}$$

Lastly, the *Labels2Effects* component converts the bounding-box information to the activation of the heat sensory effects. This phase was parameterized by the dictionary in 1 that sets a heat effect with the initial intensity of 20°C.

4.4 Filter and combine

Finally, considering the sensory effects identified by each individual module, this component filters and combines sensory effects to generate the final output. In our current implementation, a simple association rule was used to define that: if there is an ambient sensory effect (obtained by the *Scene understanding* module) and a localized effect (obtained by the *Effect localization* or the *Sun localization* modules, then the intensity of the localized effect should be the highest value between the ambient and the localized effect.

[2] https://clarifai.com

Thus, aroma effects and heat localized effects will have their intensities updated by the highest intensity of the same effects suggested by the modules. Listing 3 presents a snippet of the final list of sensory effects obtained after combining the output of the modules, including aroma and heat effects.

Listing 3: Example output of aroma and heat effects after filter-and-combine

```
{'effect': 'heat','intensity': 35, 'location':
     [(-59,-85.),(-59,-59),(-43,-85),(-43,-59)
    ]},
{'effect': 'aroma','type': 'tree','intensity':
     40,'location': [(-93,-43),(-45,-34),(-93,
    11),(-45,8)]}
```

4.5 Implementation Results

To showcase the performance of the recognition system, this section provides a visual representation of its ability to identify sensory effects in 360° images. A diverse set of real-world images was selected to demonstrate the system's capabilities, with the results displayed in Figure 2. The demonstration involved the application of the recognition architecture to identify tree aroma (in red) and heat effects from the sun (in purple). The results offer valuable insights into the system's effectiveness and highlight areas for improvement. In particular, it can be observed from Figure 2 (f) and (g) that the sun recognition module failed to fully recognize the sun. This limitation underscores the need for continued development and improvement of new modules based on novel vision-based recognition systems.

Figure 3 displays the runtime analysis of the effect recognition module's components across eight images. Tests were conducted on a laptop with an Intel i9 11900H processor, with all processing done on the CPU. Effect recognition was executed in each image ten times, and their average runtime was recorded. Notably, across all images the average pre-processing time, recognition time and post-processing time was ≈ 0.04s, ≈ 0.8s, and ≈ 0.4s, respectively. The "labels2effects" component's runtime, although not shown in the figure due to its minimal impact, averaged at 7µs.

5 CONCLUSION

This work presented an extensible architecture to automatically recognize sensory effects in 360° images. The architecture is capable of using different recognition modules, combining their results to provide both localization and presentation characteristics of sensory effects. The proposed architecture was instantiated with three recognition modules for 360° image content. The first used a DNN focused on identifying the position of sensory effects. The second used a DNN API focused on identifying the context of the scene as a whole. The third used classical computer vision algorithms to locate the sun and associate it with a heat effect.

As future work, an improvement of the sensory effect localization module is the most important step. There is a need to build DNN architectures to fully utilize equirectangular images and fasten the recognition process. Also in this module, the post-process phase can perform the union of multiple bounding-boxes of the same label that are close together, to represent a large and/or close object. Another important future work is to define a domain-specific language to apply JSON transformations on the filter-and-combine step. Another future work is to extend the proposed architecture for the recognition of sensory effects in 360° video content. Lastly, another future work is to develop an authoring tool to receive the output of recognition modules and evaluate sensory effect recognition with authors.

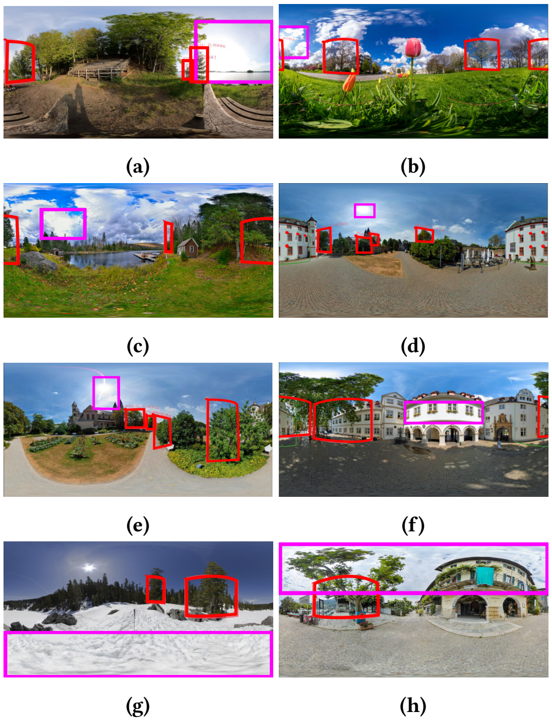


Figure 2: Sensory Effects Recognition in 360 Images

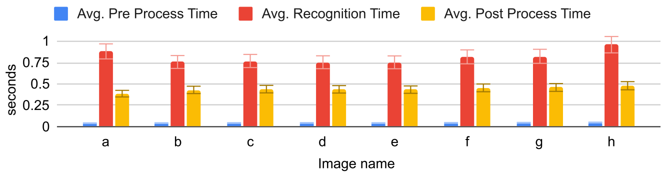


Figure 3: Average runtimes for each component of the effect localization module (with 90% confidence interval)

ACKNOWLEDGMENTS

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001. The authors also thank FAPERJ, CNPq, CAPES, CAPES Print and FINEP for partially funding this work.

REFERENCES

[1] Raphael Abreu, Joel dos Santos, and Eduardo Bezerra. 2018. A Bimodal Learning Approach to Assist Multi-sensory Effects Synchronization. In *IJCNN '18* (Rio de Janeiro, Brazil). IEEE.

[2] Raphael Abreu, Douglas Mattos, Joel Santos, George Guinea, and Débora C Muchaluat-Saade. 2023. Semi-automatic mulsemedia authoring analysis from

the user's perspective. In *Proceedings of the 14th Conference on ACM Multimedia Systems*. 249–256.

[3] James Clark et al. 1999. Xsl transformations (xslt). *World Wide Web Consortium (W3C). URL http://www. w3. org/TR/xslt* 103 (1999).

[4] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 518–533.

[5] Alexandra Covaci, Ramona Trestian, Estêvão Bissoli Saleme, Ioan-Sorin Comsa, Gebremariam Assres, Celso AS Santos, and Gheorghita Ghinea. 2019. 360° Mulsemedia: a way to improve subjective QoE in 360° videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2378–2386.

[6] Raphael Silva de Abreu, Douglas Mattos, Joel dos Santos, Gheorghita Ghinea, and Débora Christina Muchaluat-Saade. 2020. Toward content-driven intelligent authoring of mulsemedia applications. *IEEE MultiMedia* 28, 1 (2020), 7–16.

[7] Marcello Novaes de Amorim, Estêvão Bissoli Saleme, Fábio Ribeiro de Assis Neto, Celso AS Santos, and Gheorghita Ghinea. 2019. Crowdsourcing authoring of sensory effects on videos. *Multimedia Tools and Applications* 78 (2019), 19201–19227.

[8] René Octivio Queiroz Dias and Díbio Leandro Borges. 2016. Recognizing Plant Species in the Wild: Deep Learning Results and a New Database. In *2016 IEEE International Symposium on Multimedia (ISM)*. 197–202. https://doi.org/10.1109/ISM.2016.0047

[9] Wilfried Elmenreich. 2002. An introduction to sensor fusion. *Vienna University of Technology, Austria* 502 (2002), 1–28.

[10] Gabriel Giraldo, Myriam Servières, and Guillaume Moreau. 2020. Perception of multisensory wind representation in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 45–53.

[11] Marina Josué, Raphael Abreu, Fábio Barreto, Douglas Mattos, Glauco Amorim, Joel dos Santos, and Débora Muchaluat-Saade. 2018. Modeling sensory effects as first-class entities in multimedia applications. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 225–236.

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* (2020), 1–26.

[13] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring presence: the temple presence inventory. In *Proceedings of the 12th annual international workshop on presence*. 1–15.

[14] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. 2018. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM international conference on Multimedia*. 1190–1198.

[15] Karol J Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[17] Thomhert S Siadari, Mikyong Han, and Hyunjin Yoon. 2017. 4D Effect Video Classification with Shot-Aware Frame Selection and Deep Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1148–1155.

[18] Kristin Van Damme, Anissa All, Lieven De Marez, and Sarah Van Leuven. 2019. 360 video journalism: Experimental study on the effect of immersion on news experience and distant suffering. *Journalism Studies* 20, 14 (2019), 2053–2076.

[19] Yuhao Zhou, Makarand Tapaswi, and Sanja Fidler. 2018. Now You Shake Me: Towards Automatic 4D Cinema. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7425–7434.

[20] Ziqi Zhu, Li Zhuo, Panling Qu, Kailong Zhou, and Jing Zhang. 2016. Extreme Weather Recognition Using Convolutional Neural Networks. In *2016 IEEE International Symposium on Multimedia (ISM)*. 621–625. https://doi.org/10.1109/ISM.2016.0133

A Platform for Collecting User Behaviour Data during Social VR Experiments Using Mozilla Hubs

Thomas Röggla
t.roggla@cwi.nl
Centrum Wiskunde & Informatica
The Netherlands

David A. Shamma
aymans@acm.org
Toyota Research Institute
USA

Julie R. Williamson
julie.williamson@glasgow.ac.uk
University of Glasgow
United Kingdom

Irene Viola
i.viola@cwi.nl
Centrum Wiskunde & Informatica
The Netherlands

Silvia Rossi
s.rossi@cwi.nl
Centrum Wiskunde & Informatica
The Netherlands

Pablo Cesar
p.s.cesar@cwi.nl
Centrum Wiskunde & Informatica
TU Delft
The Netherlands

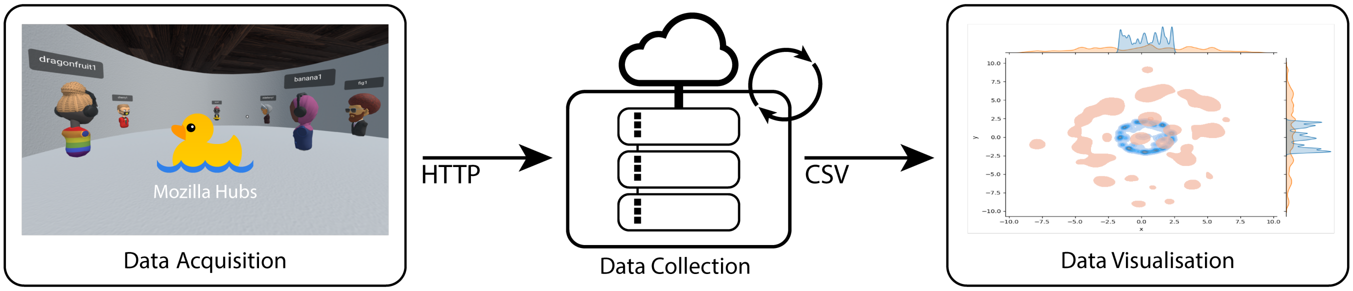


Figure 1: Workflow overview enabled by the proposed toolchain: from data gathering in Mozilla Hubs, collection and validation on a cloud machine to final data analysis.

ABSTRACT

In recent years, a large variety of online communication tools have emerged, including social Virtual Reality (VR) platforms for interacting in a virtual world with participants being represented as virtual avatars. Given their popularity, an active area of research focuses on improving the user experience in these virtual experiences. To enable experimentation at large scale on online platforms, it is however essential to collect behavioural data (e.g. movements and audio information). In this work, we present a toolchain that enables the running of experiments using a modified version of the social VR platform Mozilla Hubs. Specifically, our toolkit enables collection and tracking or user positions and movements at a central location, enabling fine-grained analysis of user behaviour during a social VR experience. The proposed tool is available at https://github.com/cwi-dis/mozillahubs-datalogger

CCS CONCEPTS

• **Human-centered computing → Visualization systems and tools**; • **Information systems → Web services**.

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04
https://doi.org/10.1145/3652212.3652225

KEYWORDS

Social VR, data collection, data visualisation, plugin, Mozilla Hubs

ACM Reference Format:
Thomas Röggla, David A. Shamma, Julie R. Williamson, Irene Viola, Silvia Rossi, and Pablo Cesar. 2024. A Platform for Collecting User Behaviour Data during Social VR Experiments Using Mozilla Hubs. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3652212.3652225

1 INTRODUCTION

Over the past few years, a new era for remote communication has begun thanks to technological advances and the introduction of novel online services in the realm of Virtual Reality (VR). VR goes beyond traditional remote communication technologies, putting the users at the center of the action and providing them with a sense of immersion and new interactive capabilities. Going one step further, social VR applications enable the virtual co-presence of multiple users within the same virtual environment, allowing body interactions similar to face-to-face communication and creating new possibilities not only for remote communication but also for collaboration and social presence, redefining the way individuals engage in virtual experiences [6, 17, 18, 23].

Social VR has been used by researchers and practitioners to enable collaborative work [4, 5, 7] and design experiences in areas such as health care [16], food [19], learning and training [1, 10, 15, 27], artistic design [22] and museum exploration [20]. A key aspect of social VR is its ability to enable embodied interactions so that users can navigate the space and interact with one another using body language and non-verbal cues as well as verbal communication. Thus, the spatial dynamics of social interaction, such as proxemics, play a significant role in understanding interpersonal relationships and communication patterns within virtual environments. Physical displacement and proxemic interactions have been analyzed to investigate which social cues are the most influential and relevant to ensure presence and immersion [14, 25]. To do so, however, researchers are in need of social VR platforms that enable the accurate logging of behavioural data (e.g. body position and rotation, interaction modalities or audio information) to analyse user behaviour [21]. Platforms such as Ubiq [9] and VR2Gather [24] were developed to enable such logging and support researchers in running user studies in controlled environments. However, such frameworks require Unity knowledge to design and develop VR experiences, which might prove difficult for designers and researchers with low technical skills in this software. Moreover, their deployment over the network needs to be orchestrated by the researchers, posing further challenges. Current commercial platforms such as Mozilla Hubs [8], Spatial.io [12], Rec Room [11] and VRChat [13] offer easy design and deployment, but do not offer data logging. This is the research gap we aim to address in this paper.

In this work, we present a toolkit that enables the running of experiments using a modified version of the popular social VR platform Mozilla Hubs [8]. The main advantage of this experimental platform relies on its versatility in working with immersive devices, such as head mounted displays, but also traditional web browsers on computers or mobile devices. Specifically, we further extend the toolkit presented in [25] such that position and movement of head and hands can be collected and tracked at a central location, enabling fine-grained analysis of user behaviour during a social VR session (Figure 1). In the following sections, we describe the implementation of the platform and we give some examples of deployment to demonstrate how it can be used to foster research in the field.

2 PLATFORM

Figure 1 outlines the workflow enabled by the system presented in this paper. The first step is data acquisition, which is realised by a plugin for the popular Social VR platform Mozilla Hubs. Then, the gathered data is streamed via HTTP to a data collection component, an optimised purpose-designed web server that validates and stores the data in a compressed format. Finally, the collected data can be visualised and checked for validity before being more deeply analysed. In following, we further describe these components.

Data Acquisition. Mozilla Hubs completely runs within a browser environment and is based on open-source libraries: ThreeJS [3] for WebGL support and AFrame [2] for integration with VR headsets. Our toolchain consists of a client-side, which runs in the user's browser and is responsible for collection and transmission of the data and a server-side, which validates the received data and stores it for further processing. Communication between the two is enabled through the HTTP protocol using POST requests.

The client-side consists of a JavaScript module which is registered in the global scope of the A-Frame library. This gives the module access to all objects within the virtual world and is executed by A-Frame on every tick of its event loop. After system initialisation, the tick function is activated and runs on every frame update of the system. The module extracts a variety of information from the browser's DOM tree and the virtual environment. In the current version of the client-side module, system-related metrics, such as timestamp and frame rate are collected from the browser environment while behavioural data is extracted from the DOM tree. A complete list of metrics collected by these two sources are given in Table 1a and 1b, respectively. All this data is updated and saved based on the frame rate of the user's headset. To minimise the number of requests sent to the data collection server, the data is buffered and a POST request containing all the data as a JSON-formatted payload is only sent every 4000 ticks.

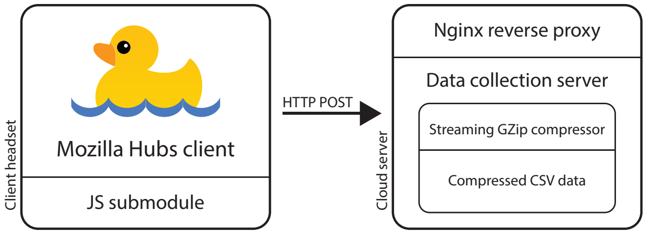


Figure 2: Architecture of the system, showing the relationship between Hubs and the data collection server.

Data Collection. The data collection is done on a separate cloud server which is responsible for validating and storing all session data. As shown in Figure 2, the cloud server with a single HTTP POST endpoint is implemented using the Go programming language to achieve adequate performance and keep system load to a minimum. Running as a background process, the server listens for incoming POST requests on TCP port 6000. Further, all requests are handed to the server through a Nginx reverse proxy, which takes care of CORS policy validation to allow the Mozilla Hubs clients to communicate directly with the data collection server via AJAX.

Upon reception of a request on the right endpoint, a streaming JSON decoder is instantiated, ready to receive the payload body of the POST request. Once the entire payload is received and validated, the program checks the presence of all required fields. If all required fields are present, the decoded payload is converted to a comma-separated format and appended to a compressed CSV file via a streaming GZip compressor. The server also adds a UNIX timestamp to each record, which can be used to correct possible time drift and/or inaccuracies in the timestamps received from the clients. To prevent file corruption through concurrent access, the write operation is guarded by a mutex. After a successful write, the request handler returns a message with HTTP status 200 to the client; if the submitted data did not pass validation, the server returns an error with HTTP status 400; and if the data could not be written, an error with HTTP status 500 is returned. Through the use of GZip compression, the data collected in a session, which typically amounts to about 2 GB, can be compressed to about 500 MB, keeping storage space use to a minimum. Further, by using a streaming compressor, the file handle can be held onto without having to close and reopen for every request.

timestamp	Device's UNIX timestamp
fps	Current frame rate
uuid	Random UUID
user_agent	Device user agent
isBrowser	Device type
isLandscape	Device orientation
isWebXRAvailable	VR availability
avatarID	Avatar ID
isHeadsetConnected	Headset connection status
isRecording	Recording status
pathname	Current URL
urlQuery	Query section of the URL

(a) Data collected from the browser environment

isLoaded	Has user finished loading
isEntered	Has user joined room
isFlying	Is user flying
isVisible	Is user visible
isSpeaking	Is user speaking
isMuted	Is user muted
volume	Current user volume
rigPos{X, Y, Z}	Avatar position
rigDir{X, Y, Z}	Avatar direction
rigQuat{X, Y, Z, W}	Avatar quaternion rotation
povPos{X, Y, Z}	POV position
povDir{X, Y, Z}	POV direction
povQuat{X, Y, Z, W}	POV quaternion rotation

(b) Data collected from the DOM tree

Table 1: User metrics collected by the toolchain

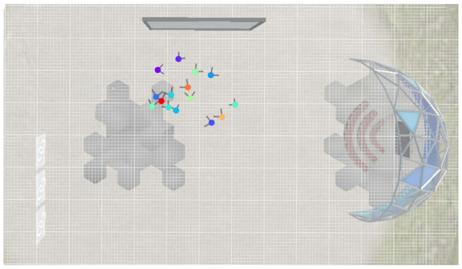


Figure 3: Use case of instrumented Mozilla Hubs in an academic workshop [25].

Data Visualisation. The presented toolchain also offers the possibility to perform a quick sanity check on the collected data by taking the data stream and replaying it using a crudely rendered model of a head and hands, placed into a 3D environment. This allows researches to quickly assess the viability of the collected data before analysing it. From this point onward, the data that consists of a CSV format with a column for all the properties in Table 1, indexed and sorted by timestamp. The file can be decompressed and analysed using conventional data analysis tools. For reference, Figure 3 shows an example of a visualisation generated from results obtained using the toolchain during a previous study.

3 DEPLOYMENT AND USE CASES

The deployment of the proposed toolchain involves a series of prerequisites. Chief among them is a private instance of Mozilla Hubs that can be achieved by using Mozilla's official AWS CloudFormation recipe. This recipe will deploy all the needed services on a selected AWS account and start them. After configuration of a custom domain, email settings for login and the admin panel, the custom client including the data collection JavaScript submodule can be checked out from Github and deployed to the running instance directly from the command line, following the guides found in the Hubs documentation. The second prerequisite to complete the system is the server to collect and validate the data. The Go-based server can be checked out from Github and, after compilation of the sources to a self-contained binary, can be launched and will start listening for incoming HTTP requests on port 6000. If the server runs on a domain different from the Mozilla Hubs instance, the clients will refuse to send AJAX requests to the data collection server because it would violate the *Cross-Origin Security Policy (CORS).* To address this, we encourage putting the data collection server behind a reverse-proxy such as Nginx and configure it to allow requests from the domain name of the Hubs instance. This way, Nginx handles CORS negotiation and hands off authorised requests to the server. From this point onward, the logger is automatically enabled for any client that joins a room with the parameter `?log=true` appended to the query string of its URL and will start streaming content to the data collection server. Arrival of data can be monitored on the standard output of the server.

Our proposed toolkit can be essential to investigate interactions in social VR. The system [25] on which we based our work has enabled investigation on digital proxemics, an emerging area focused on understanding the human use of space within virtual environments. Specifically, it has been used to analyse how people use space in a virtual academic workshop [25] and how personal displacement changes between VR and traditional desktop users [26]. In both cases, the system supported the gathering and collecting of data from participants allowing data visualisations as shown in Figure 3 and enabling new behavioural analysis. The use of our novel toolkit will enable the augmentation of collected data, paving the way to new investigations such as the impact of the design of virtual environments and how interaction unfolds in social VR.

It should be noted, however, that Mozilla recently announced the sunset of their AWS deployment recipe, complicating the use of their private instances. As alternative, Mozilla started offering a new professional plan, which outsources the hosting and management completely to Mozilla, while still allowing to deploy custom clients. This makes use of Hubs easier, as the management of the infrastructure is completely taken care of, albeit slightly changing the deployment process of our toolchain.

4 CONCLUSION

This work described an easy-to-deploy tool for collecting and storing behavioural data from the popular social VR tool Mozilla Hubs. Our solution can be integrated into a running instance of Hubs to gather metrics from participants within a browser environment. Collected data is stored off-site using an optimised purpose-built web server in a compressed format, making it possible to store substantial amounts of data without placing too much load on the host system.

ACKNOWLEDGMENTS

This work was in part supported through the European Commission Horizon Europe program under grant 101070109 TRANSMIXR (https:// transmixr.eu/).

REFERENCES

[1] Sara Arlati, Vera Colombo, Daniele Spoladore, Luca Greci, Elisa Pedroli, Silvia Serino, Pietro Cipresso, Karine Goulene, Marco Stramba-Badiale, Giuseppe Riva, et al. 2019. A social virtual reality-based application for the physical and cognitive training of the elderly at home. *Sensors* 19, 2 (2019), 261.

[2] A-Frame Authors. 2015. *A-Frame*. https://aframe.io/.

[3] ThreeJS Authors. 2010. *ThreeJS*. https://threejs.org.

[4] Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. 1995. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 242–249.

[5] Steve Benford, Chris Greenhalgh, Tom Rodden, and James Pycock. 2001. Collaborative virtual environments. *Commun. ACM* 44, 7 (2001), 79–85.

[6] Frank Biocca and Mark R. Levy. 1995. Virtual reality as a communication system. *Communication in the age of virtual reality* (1995), 15–31.

[7] Elizabeth F Churchill and Dave Snowdon. 1998. Collaborative virtual environments: an introductory review of issues and systems. *virtual reality* 3 (1998), 3–15.

[8] Mozilla Corporation. 2017. *Mozilla Hubs*. https://hubs.mozilla.com.

[9] Sebastian J Friston, Ben J Congdon, David Swapp, Lisa Izzouzi, Klara Brandstätter, Daniel Archer, Otto Olkkonen, Felix Johannes Thiel, and Anthony Steed. 2021. Ubiq: A system to build flexible social virtual reality experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[10] Simon Gunkel, Hans Stokking, Martin Prins, Omar Niamut, Ernestasia Siahaan, and Pablo Cesar. 2018. Experiencing virtual reality together: Social VR use case study. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*. 233–238.

[11] Rec Room Inc. 2016. *Rec Room*. https://recroom.com.

[12] Spatial Systems Inc. 2017. *Spatial*. https://www.spatial.io.

[13] VRChat Inc. 2014. *VRchat*. https://hello.vrchat.com.

[14] Duc Anh Le, Blair Maclntyre, and Jessica Outlaw. 2020. Enhancing the experience of virtual conferences in social virtual environments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, New York, NY, USA, 485–494.

[15] Quang Tuan Le, Akeem Pedro, and Chan Sik Park. 2015. A social virtual reality based construction safety education system for experiential learning. *Journal of Intelligent & Robotic Systems* 79, 3 (2015), 487–506.

[16] Jie Li, Guo Chen, Huib De Ridder, and Pablo Cesar. 2020. Designing a social vr clinic for medical consultations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[17] Jie Li, Vinoba Vinayagamoorthy, Julie Williamson, David A Shamma, and Pablo Cesar. 2021. Social VR: A new medium for remote communication and collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[18] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping pro-social interaction in VR: an emerging design framework. In *Proceedings of the ACM CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[19] Yanni Mei, Jie Li, Huib de Ridder, and Pablo Cesar. 2021. Cakevr: A social virtual reality (vr) tool for co-designing cakes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[20] Ignacio Reimat, Yanni Mei, Evangelos Alexiou, Jack Jansen, Jie Li, Shishir Subramanyam, Irene Viola, Johan Oomen, and Pablo Cesar. 2022. Mediascape XR: A Cultural Heritage Experience in Social VR. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6955–6957.

[21] Silvia Rossi, Irene Viola, Jack Jansen, Shishir Subramanyam, Laura Toni, and Pablo Cesar. 2021. 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)*. 1–7.

[22] Asreen Rostami, Kasper Karlgren, and Donald McMillan. 2022. Kintsugi VR: Designing with Fractured Objects. In *ACM International Conference on Interactive Media Experiences* (Aveiro, JB, Portugal) *(IMX '22)*. Association for Computing Machinery, New York, NY, USA, 95âĂŞ108. https://doi.org/10.1145/3505284.3529966

[23] Ralph Schroeder. 2010. *Being There Together: Social interaction in shared virtual environments*. Oxford University Press.

[24] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. Vr2gather: A collaborative social vr system for adaptive multi-party real-time communication. *IEEE MultiMedia* (2023).

[25] Julie Williamson, Jie Li, Vinoba Vinayagamoorthy, David A Shamma, and Pablo Cesar. 2021. Proxemics and social interactions in an instrumented virtual reality workshop. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[26] Julie R Williamson, Joseph O'Hagan, John Alexis Guerra-Gomez, John H Williamson, Pablo Cesar, and David A Shamma. 2022. Digital proxemics: Designing social and collaborative interaction in virtual environments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[27] Chiara Zizza, Adam Starr, Devin Hudson, Sai Shreya Nuguri, Prasad Calyam, and Zhihai He. 2018. Towards a social virtual reality learning environment in high fidelity. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 1–4.

Untethered Real-Time Immersive Free Viewpoint Video

Javier Usón, Carlos Cortés, Victoria Muñoz, Teresa Hernando, Daniel Berjón, Francisco Morán,
Julián Cabrera and Narciso García
Grupo de Tratamiento de Imágenes,
Information Processing and Telecommunications Center,
ETSI Telecomunicación, Universidad Politécnica de Madrid
Madrid, Spain
j.usonp@upm.es

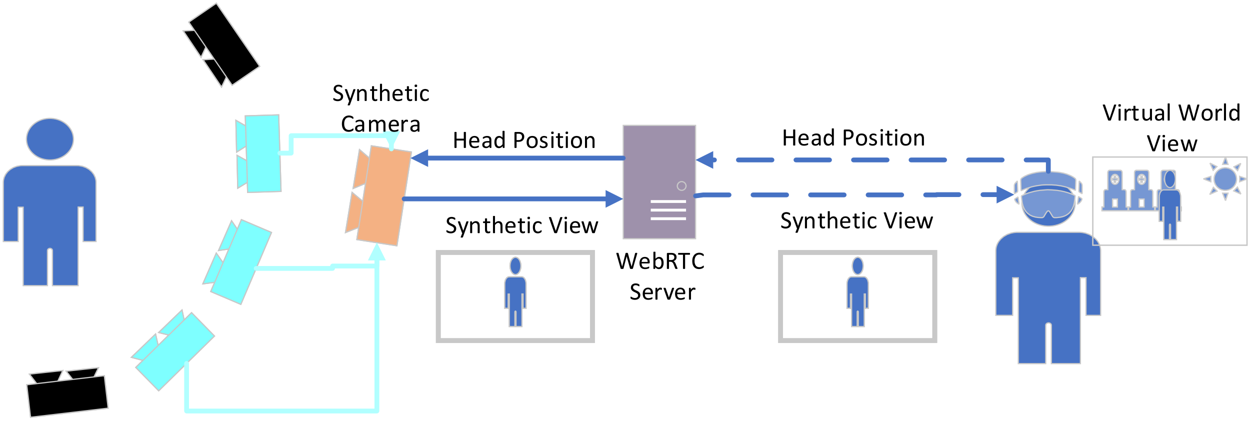


Figure 1: *FVV Live* Immersive System Diagram.

ABSTRACT

The recent development of new video capture systems has led to the adoption of volumetric video technologies to replace 2D video in use cases such as videoconference, where this enhancement promises to solve videoconference fatigue. In particular, volumetric capture allows the content to be viewed from different points of view, enabling more natural interaction during the videoconference. One of the solutions proposed for this scenario is Free Viewpoint Video (FVV). It makes use of a set of calibrated cameras that allows the use of real life information to generate a synthetic view from any arbitrary point in space. Although there are real-time capture developments of FVV systems, they make use of 2D displays and joysticks to control the point of view. In our opinion, this undermines the possibilities of volumetric video for the videoconferencing use case. Building on a previously developed FVV system, we present a novel untethered HMD-based immersive visualization system that enables point of view control with the user's natural position and visualization of live volumetric content in a 3D environment. Synthetic views are generated in real-time by the FVV system, and streamed with low latency protocols to a Meta quest 3 HMD using a WebRTC-based server. This work discusses the architecture of the end-to-end system and describes the bitrate, framerate and latency values at which the system works.

CCS CONCEPTS

• Human-centered computing → Visualization systems and tools; **• Information systems → Multimedia streaming**.

ACM Reference Format:
Javier Usón, Carlos Cortés, Victoria Muñoz, Teresa Hernando, Daniel Berjón, Francisco Morán, Julián Cabrera and Narciso García. 2024. Untethered Real-Time Immersive Free Viewpoint Video. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3652212.3652214

1 INTRODUCTION

The latest technological developments in the area of video transmission are directing the focus towards the volumetric video. Recent studies link 2D streaming media with negative effects on user experience [14]. This is known as videoconferencing fatigue [1]. Furthermore, these studies relate this effect to the technological limitations of 2D video. In particular, the impossibility of free movement and two-dimensional representation. In this context, volumetric video is postulated as a solution because it allows the natural visualization of content from different points of view, allowing freedom of movement [11, 13, 14]. In the literature there are different methods

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04…$15.00
https://doi.org/10.1145/3652212.3652214

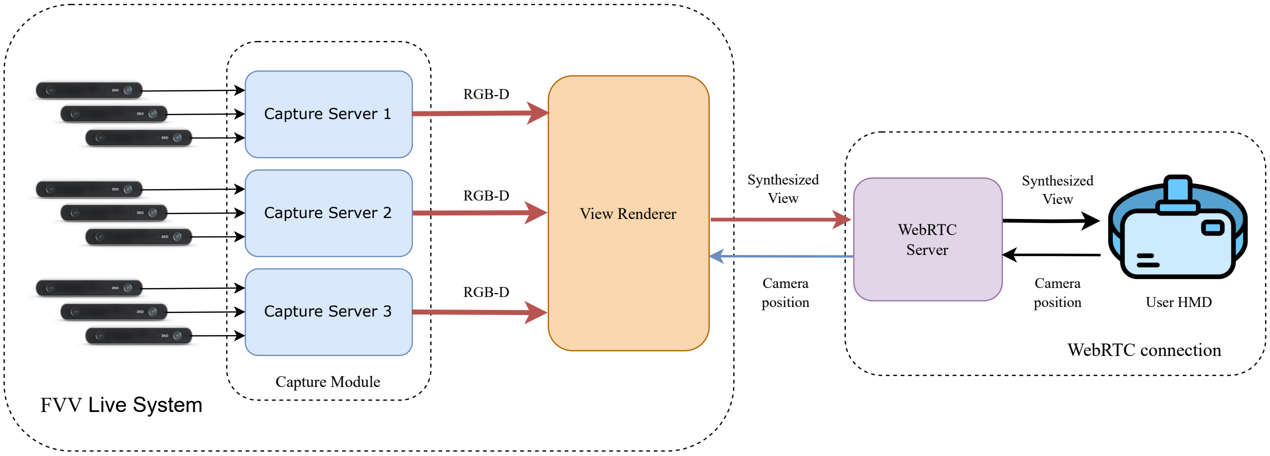


Figure 2: Diagram of the complete architecture of the immersive FVV system. The left side presents the components from *FVV Live*: RGB-D information from the scene is obtained thanks to a set of stereo cameras and transmitted to the view renderer, which is capable of synthesizing a new view from any arbitrary point of view. The right side presents the wireless connection of the user HMD to the view renderer through a WebRTC server: the parameters of the HMD virtual camera are transmitted using a WebRTC data channel, and the requested point of view is rendered and forwarded to the HMD using RTP.

to generate volumetric video. For example, [16] presents a volumetric videoconference system based on active depth cameras to generate a 3D volume. Another way to generate volumetric video is free-viewpoint, this technique consists of generating an on-demand 3D view using the information from a set of cameras. Although there are different implementations of this type of video [2, 15, 18], these systems use a display based on 2D. Moreover, these works implement techniques to modify the point of view using devices such as mouse and joysticks, lacking natural interaction. An alternative solution to solve these issues is the use of head mounted displays (HMDs). HMDs allow the visualization of volumetric content through the use of displays in front of your eyes, Furthermore, the current HMDs incorporate a head tracking system with 6DoF that allows the user to move around the virtual world. Thus, taking advantage of all the benefits of volumetric video. This type of solution has already been addressed in [18]. However, due to technical limitations at the time, their solution requires a wired connection, limiting the user's freedom of movement. Another problem is the impossibility of that FVV system to display video in real time, making interactive use cases impossible. In this paper, we present what to our knowledge is the first development of real-time FVV with immersive visualization and interaction. In summary, this work contributes to:

- Present the first implementation of a videoconference FVV system with immersive control and display.
- Present an analysis of the technical characteristics including bitrate, latency and resolution.

The paper is structured as follows: section 2 presents literature related to 3D video systems; section 3 presents the operation scheme of our FVV system with a detailed description of its processes; section 4 presents an analysis of the performance of our system with other 3D video systems; section 5 presents the conclusions of the work along with directions for future work.

2 RELATED WORK

2.1 Volumetric Video

Volumetric video consists of capturing three-dimensional space through the use of video cameras [15]. Thanks to this technology it is possible to offer multimedia content with greater interactivity due to the freedom of movement [11]. This is because the content can be viewed from different points of view. Among the techniques that can be found to generate volumetric video are: LIDAR based, structured light based, lightfields or calibrated stereo cameras [8].

Volumetric video is being adapted for integration into a multitude of use cases: surgery [12], creative storytelling [19] and immersive videoconference or Social XR [16]. Specifically, volumetric video is very promising in the area of videoconferencing. It is postulated as a technology that can help overcome current issues, like video conferencing fatigue, mainly caused by the lack of free movement or flat representation of users [14]. In addition, there are studies that claim that this type of video, coupled with immersive visualization, improves the user experience substantially, taking videoconferencing to its maximum exponent of realism [11, 14].

One way to generate volumetric video is through free viewpoint video. In this case, a set of calibrated cameras is used to generate a synthetic view at a given point of view. By modifying this point, the system generates a new viewport, allowing to see the content from different perspectives. Finally, FVV systems usually generate the final viewport, not requiring a point cloud or mesh format. Thanks to this, it is possible to take advantage of all the existing multimedia transmission infrastructure to enable volumetric video. Therefore, this solution does not need to adapt the transcoding stages to specific volumetric data type.

2.2 FVV Systems

FVV systems, unlike those based on generating complete volumes, try to generate the view that the user demands at that moment. In the literature, there are different implementations that make use of this technique to generate volumetric video [2, 15, 18]. The advantages of this form of volumetric capture over others is that it allows higher image quality, as it does not depend on lower resolution sensors such as active depth cameras. Thus being able to take full advantage of the color resolution of the camera. Another advantage is the ease with which this type of solution can be integrated into established video transcoding and transmission workflows. Since the generated view is a video that follows pre-existing encoding standards. This allows to reuse all the software/hardware with which current video players and decoders are already compatible. However, free viewpoint video is slower when it comes to changing the viewpoint, since the display device does not receive the full volume, being at the mercy of the network when changing the viewpoint.

Viewport-based FVV systems are ideal for immersive systems. Current HMDs are battery-dependent and their processing power is limited. In addition, their mobile GPUs make volume-based volumetric video rendering more complex, whereas current HMDs are very well prepared to receive and decode standard video.

This work presents an extension to the *FVV Live* system proposed in [2, 10] to enable untethered immersive HMDs. Previous work covered the end-to-end FVV pipeline: volumetric capture of the scene, media encoding and transmission, and view synthesis, all while working in real-time with low latency.

3 SYSTEM ARCHITECTURE

Figure 1 presents an overview of the architecture proposed. The system can be divided into three blocks: The *FVV Live* system with all its internal components, a WebRTC server, and the user HMD working as a wireless WebRTC client. Figure 2 provides a detailed diagram of this architecture at the process level.

FVV Live is formed by a set of Stereolabs ZED cameras that capture RGB-D information from the scene. These media streams are processed, encoded and transmitted from a group of capture servers to a view render server. This server chooses the three closest cameras to the desired viewpoint and uses their RGB-D information to render the new view point. The *FVV Live* pipeline is a black box for the other blocks, the view renderer takes a camera position and synthesizes the view from that view point. The WebRTC server is in charge of communicating the *FVV Live* view renderer with the HMD using low latency protocols.

Live volumetric video transmission requires a huge amount of resources, namely processing power, network bandwidth, a set of volumetric cameras and a stage to be recorded. With the approach proposed, all of these requirements are managed by the *FVV Live* system, leaving the end user with minimal processing and bandwidth restrictions.

3.1 *FVV Live*

The *FVV Live* system is in charge of performing the volumetric capture of the scene and rendering said scene from the viewpoint requested by the HMD.

The volumetric capture stage is performed by a group of synchronized stereo cameras which allows the computation of the geometrical information from the scene as depth images. Depth information is heavily affected by traditional lossy video compression, so it is encoded with a lossless scheme. To reduce the bitrate for the transmission, foreground segmentation is performed to send only foreground depth information.

Since the system discards depth data from the background, this information has to come from an alternative source. Assuming the background is static, it can be reconstructed offline, free of real-time constraints, using more compute-intensive techniques such as structure from motion (SfM) [4].

As explained by [5], the system performs a layered synthesis separating the live (online) foreground from the pre-computed (offline) background. Firstly, to synthesize the foreground, online depth maps from the closest three reference cameras are combined to build a virtual depth map that can then be used to trace back the colour to the reference cameras using backwards Depth Image Based Rendering (DIBR). Secondly, to synthesize the background, the high-quality offline depth maps are used in a similar way. Finally, the holes between both layers are filled.

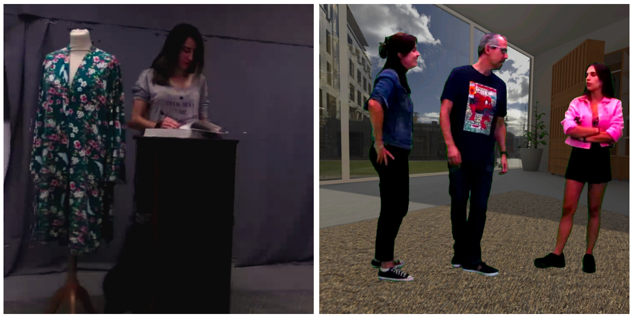
3.2 WebRTC server

The wireless transmission of synthetic view to the HMD is provided through WebRTC (Web Real-Time Communication) [17], a widely-used and open source real-time communication protocol for web applications, allowing peer-to-peer communication. The *FVV Live* system is connected to a WebRTC server that enables bidirectional data transmission. The choice of WebRTC is motivated by its ability to handle real-time media streaming with low latency and adaptability to various network conditions. This WebRTC server acts as an intermediary between the FVV system and the HMD. The server is implemented using aiortc [9], a library for WebRTC and Object Real-Time Communication (ORTC) in Python.

In WebRTC, media and data streams are transmitted via a peer-to-peer connection. Session Description Protocol (SDP) [6] and Interactive Connectivity Establishment (ICE) [7] are two important protocols used in WebRTC connection establishment. While SDP is responsible for negotiating the parameters of a multimedia session between the devices, ICE is responsible for providing the connection between devices over the network. To establish the WebRTC connection, the peers need to complete a signaling process first.

This process starts with an “Offer” SDP from the HMD client so the media streams details such as the type of codec, transport protocol and other related information can be negotiated. After that, once the WebRTC server receives the “Offer” SDP previously sent by the HMD client, It returns an “Answer” SDP to the HMD client containing its media details. In order to establish the connection between the server and the HMD client, the “ICE gathering” takes place. This process starts with the HMD client sending its network address (known as ICE Candidate) to the WebRTC server so when the latter checks for unprocessed Ice Candidates and receives the one sent by the HMD client, it returns its own ICE Candidate. This information is used to determine the best available network path for the session, ensuring a successful connection.

Once the signaling process has finished, the connection is established and transmission can begin. In the proposed approach, the streams involved are a reliable data channel, and both video and audio streams being transmitted to the HMD.



(a) View fully rendered by *FVV Live* (b) Avatars rendered by *FVV Live*

Figure 3: Examples from the two proposed scenarios. In (a), the full scene is rendered by *FVV Live*. In (b), only the avatars are rendered by *FVV Live*, while the rest of the scene is a virtual scenario rendered by the HMD.

3.3 Integration with HMDs

The client application which runs on the HMD was developed using Unity. It handles WebRTC connection establishment, presentation of the received video and sending the camera position information.

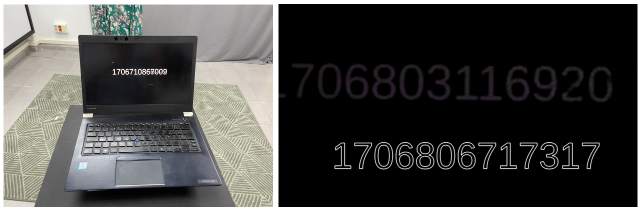
For the presentation of the received video, the virtual environment consists of a plane displaying the video which is placed in front of the user's point of view, covering their entire field of vision. The plane is attached to the virtual camera, so it follows the user when he/she moves its head. Finally, the Unity engine is the one in charge of generating the left and right views to be presented by the HMD.

The camera pose from the Unity scene is transmitted to the *FVV Live* system so the virtual cameras from both virtual scenarios match accurately. This transmission is performed by a WebRTC data channel every frame. The pose is encoded as a JSON containing the position (X, Y and Z coordinates) in meters, rotation (Euler angles) in degrees, and the horizontal field of view in degrees of the main camera ("center eye"). This camera is controlled by the HMD using the user's head movement.

With this configuration, the user is able to visualize and control a FVV scene in an immersive way: the new synthetic view is rendered by the *FVV Live* and visualized on an HMD. In this configuration, the synthetic view occupies the user's entire viewport in the virtual world. Additionally, a different scenario is proposed, where only some elements from the FVV scene are presented on top of a virtual world. Figure 3 shows an example of such scenario, where the *FVV Live* only renders people, and those "avatars" are integrated with a virtual scene.

To achieve this integration, the live scene has to be segmented, and as explained in subsection 3.1, the *FVV Live* system already performs segmentation. To take advantage of this, all the discarded pixels are given an specific color (green), so the HMD can remove them and show what is behind in the virtual scene.

If virtual elements are added to the Unity scene, the user will be able to visualize them behind the FVV avatars. This approach has the particularity that background objects are rendered by the HMD, free from the delay and bandwidth restrictions of transmission.



(a) Time reference (b) HMD display

Figure 4: End-to-end latency measurements method, involving one computer with a clock reference synchronized to the HMD. The HMD presents a timestamp shown by the screen of the computer captured by the cameras and rendered by the system next to its current timestamp. The end-to-end latency is the difference between them.

4 PERFORMANCE ANALYSIS

This section addresses the performance of the proposed system in terms of transmission resolution, framerate, bitrate and latency. A Meta Quest 3 headset was used for these tests.

Resolution and framerate restrictions are imposed by the *FVV Live* system. Given the capture hardware used, the system can work either with 720p or 1080p resolution and at 5, 10, 15, 20 and 30 fps. The resolution can be reduced in the WebRTC transmission to assure low latency in challenging network conditions.

Regarding the bitrate, the information transmitted to and from the client involves one 2D video stream, one monophonic audio stream, and a secure data stream to send camera information. The video stream bitrate can be adapted depending on the content and network conditions. As an example, for a videoconference scenario with people moving around the scene, with 1080p resolution at 30 fps, the resulting bitrate was 3 Mbps. The other streams require negligible bandwidth, with audio generating approximately 125 kbps and data 120 kbps.

Volumetric video transmission from the *FVV Live* capture module to the view renderer requires high bandwidth given the lossless encoding of depth information, up to 100 Mbps per reference camera. Nevertheless, the system acts as a black box in the scenario proposed, so the communication can be performed under a controlled wired network. This way, bandwidth restrictions do not affect the end user.

To measure end-to-end latency, a USB cable is used to synchronize the clocks of a PC and the HMD. Then, using the capture system, images are taken of the clock. In the HMD, the capture of the PC clock can be observed together with the internal clock of the HMD. Figure 4 shows a snapshot of the view in the HMD. Here the time at which the snapshot was taken (upper clock) and the current time (lower clock) can be seen. This technique is an Android adapted version of [3]. The results show an average end-to-end delay of 380±16 ms.

Table 1 presents a summary of the results from the performance analysis carried out on the proposed system. This analysis shows how the system is capable of delivering high resolution requiring low network bandwidth and with latency similar to state-of-the-art immersive systems such as [16].

Table 1: Summary of the system performance analysis

Resolution	Framerate	Bitrate	End-to-end latency
720p or 1080p	from 5 to 30 fps	Adjustable ~3Mbps	380±16 ms

5 CONCLUSIONS

In this work we propose what to our knowledge is the first implementation of real-time FVV with immersive visualization and interaction. It allows the user to navigate a live scene freely using an HMD, and covers the end-to-end pipeline: volumetric capture, synthetic view rendering, transmission to the HMD and visualization.

The implementation involves the *FVV Live* system, which is in charge of capturing the live scene and rendering the synthetic views, and a WebRTC server that enables communication between the client HMD and the rendering process. The HMD requests an specific point of view, the rendering server synthesises said view and the result is transmitted, encoded as a video, to the HMD.

Two different scenarios are studied, the first one being the simplest approach, where the *FVV Live* renders the full scene and the HMD only presents the received video. The second one involves segmentation of the FVV scene (e.g. *FVV Live* only renders people), so it can be integrated with a virtual scenario rendered by the HMD. Thus enabling its integration into Social XR. Figure 3 shows both scenarios.

Performance tests were conducted on the system, with results showing that this approach is able to provide HD resolution (views rendered at 1080p and 30 fps) with low end-to-end delay (~380 ms) on a Meta Quest 3 headset, all with a low bandwidth requirement for the end user (~3 Mbps).

This solution for volumetric video transmission is an important contribution to the area of immersive communications. Specifically, this solution concentrates the efforts of viewpoint synthesis and encoding under the same infrastructure, leaving the client as a simple receiver of 2D video. In this way, all the infrastructure present in the world of video transmission is fully valid. However, we have yet to study the delay and bitrate implications of our proposal when it is taken out of the lab. As future work in this regard, we propose the evaluation of different technical parameters in the QoE to elucidate the limits of the system when transmitting over the network and to compare it to other state-of-the-art solutions. Moreover, further development of the WebRTC server is proposed to allow for more manual control over the video transmission to reduce motion-to-photon latency (M2P). Furthermore, adding transmission of depth information (RGB-D) will also be explored, aiming to use it for the stereo view generation in the HMD, and to be able to correctly solve occlusions in the virtual scene.

ACKNOWLEDGMENTS

This work was partially supported by project HORIZON-IA-101070250 (XRECO) funded by the European Union, by project PID2020-115132RB (SARAOS) funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government, and by project UNICO-5G I+D TSI-063000-2021-80 (DISRADIO – Pilotos) funded by the Ministry of Digital Transformation of the Spanish Government and the NextGenerationEU (Recovery, Transformation and Resilience Plan - PRTR).

REFERENCES

[1] Andrew A Bennett, Emily D Campion, Kathleen R Keeler, and Sheila K Keener. 2021. Videoconference fatigue? Exploring changes in fatigue after videoconference meetings during COVID-19. *Journal of Applied Psychology* 106, 3 (2021), 330.

[2] Pablo Carballeira, Carlos Carmona, César Díaz, Daniel Berjón, Daniel Corregidor, Julián Cabrera, Francisco Morán, Carmen Doblado, Sergio Arnaldo, María del Mar Martín, and Narciso García. 2022. FVV Live: A Real-Time Free-Viewpoint Video System With Consumer Electronics Hardware. *IEEE Transactions on Multimedia* 24 (2022), 2378–2391. https://doi.org/10.1109/TMM.2021.3079711

[3] Robert Gruen, Eyal Ofek, Anthony Steed, Ran Gal, Mike Sinclair, and Mar Gonzalez-Franco. 2020. Measuring System Visual Latency through Cognitive Latency on Video See-Through AR devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 791–799. https://doi.org/10.1109/VR46266.2020.00103

[4] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision.* Cambridge university press.

[5] Teresa Hernando, Daniel Berjón, Francisco Morán, Javier Usón, Cesar Díaz, Julián Cabrera, and Narciso García. 2023. Real-Time Layered View Synthesis for Free-Viewpoint Video from Unreliable Depth Information. In *Proceedings of the 15th International Workshop on Immersive Mixed and Virtual Environment Systems* (Vancouver, BC, Canada) *(MMVE ’23)*. Association for Computing Machinery, New York, NY, USA, 7–11. https://doi.org/10.1145/3592834.3592881

[6] IETF. 2006. Session Description Protocol (SDP). https://tools.ietf.org/html/rfc4566.

[7] IETF. 2018. Interactive Connectivity Establishment (ICE). https://tools.ietf.org/html/rfc8445.

[8] Yili Jin, Kaiyuan Hu, Junhua Liu, Fangxin Wang, and Xue Liu. 2023. From Capture to Display: A Survey on Volumetric Video. arXiv:2309.05658 [cs.MM]

[9] Jeremy Lainé. 2024. aiortc. https://github.com/aiortc/aiortc.

[10] Pablo Pérez, Daniel Corregidor, Emilio Garrido, Ignacio Benito, Ester González-Sosa, Julián Cabrera, Daniel Berjón, César Díaz, Francisco Morán, Narciso García, Josué Igual, and Jaime Ruiz. 2022. Live Free-Viewpoint Video in Immersive Media Production Over 5G Networks. *IEEE Transactions on Broadcasting* 68, 2 (2022), 439–450. https://doi.org/10.1109/TBC.2022.3154612

[11] Pablo Pérez, Ester Gonzalez-Sosa, Jesús Gutiérrez, and Narciso García. 2022. Emerging Immersive Communication Systems: Overview, Taxonomy, and Good Practices for QoE Assessment. *Frontiers in Signal Processing* 2 (2022). https://doi.org/10.3389/frsip.2022.917684

[12] Moritz Queisner, Michael Pogorzhelskiy, Christopher Remde, Johann Pratschke, and Igor M Sauer. 2022. VolumetricOR: A New Approach to Simulate Surgical Interventions in Virtual Reality for Training and Education. *Surg Innov* 29, 3 (Feb. 2022), 406–415.

[13] Oliver Schreer, Ingo Feldmann, Sylvain Renault, Marcus Zepp, Markus Worchel, Peter Eisert, and Peter Kauff. 2019. Capture and 3D Video Processing of Volumetric Video. In *2019 IEEE International Conference on Image Processing (ICIP)*. 4310–4314. https://doi.org/10.1109/ICIP.2019.8803576

[14] Janto Skowronek, Alexander Raake, Gunilla H. Berndtsson, Olli S. Rummukainen, Paolino Usai, Simon N. B. Gunkel, Mathias Johanson, Emanuël A. P. Habets, Ludovic Malfait, David Lindero, and Alexander Toet. 2022. Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. *IEEE Access* 10 (2022), 63885–63931. https://doi.org/10.1109/ACCESS.2022.3176369

[15] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand. 2006. 3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards. In *2006 IEEE International Conference on Multimedia and Expo*. 2161–2164. https://doi.org/10.1109/ICME.2006.262683

[16] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia* 30, 2 (2023), 48–59. https://doi.org/10.1109/MMUL.2023.3263943

[17] WebRTC Working Group. 2021. Web Real-Time Communication (WebRTC). https://www.w3.org/TR/webrtc/.

[18] Yuko Yoshida and Tetsuya Kawamoto. 2014. [DEMO] Displaying free-viewpoint video with user controlable head mounted display DEMO. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 389–390. https://doi.org/10.1109/ISMAR.2014.6948503

[19] Gareth W. Young, Néill O’Dwyer, and Aljosa Smolic. 2023. Chapter 21 - Volumetric video as a novel medium for creative storytelling. In *Immersive Video Technologies*, Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar (Eds.). Academic Press, 591–607. https://doi.org/10.1016/B978-0-32-391755-1.00027-4

V2RA: a Grid-Based Rate-Adaptation Logic for Volumetric Video

Zafer Gurel[*], Alperen F. Zengin[*], Ali C. Begen[*], Saba Ahsan[★], Lukasz Kondrad[κ], Kashyap Kammachi-Sreedhar[★], Serhan Gül[κ], Gazi Illahi[★] and Igor D.D. Curcio[★]
[*]Ozyegin University, [★]Nokia,[κ]Nokia
[*]Turkiye,[★]Finland,[κ]Germany

ABSTRACT

The quality-bitrate relationship is not necessarily as straightforward in volumetric video content as in 2D video. This is caused by the different volumetric video components affecting the overall quality of the content disproportional to their bitrates. Therefore, switching to a combination of higher bitrate components to improve the quality may not always produce the best outcome when making rate-adaptation decisions for streaming. To address this problem, this study proposes a new rate-adaptation logic named Volumetric Video Rate Adaptation (V2RA). The experiments performed using the MPEG Immersive Video (MIV) standard show that V2RA can significantly reduce bandwidth consumption at the expense of an acceptable loss in quality. In some cases, V2RA even achieves quality gains together with bandwidth savings.

CCS CONCEPTS

• **Networks** → **Application layer protocols**; • **Information systems** → *Multimedia streaming*.

KEYWORDS

Viewport-dependent streaming, immersive video, MIV, 3D, virtual reality, adaptive streaming.

ACM Reference Format:
Zafer Gurel[*], Alperen F. Zengin[*], Ali C. Begen[*], Saba Ahsan[★], Lukasz Kondrad[κ], Kashyap Kammachi-Sreedhar[★], Serhan Gül[κ], Gazi Illahi[★] and Igor D.D. Curcio[★]. 2024. V2RA: a Grid-Based Rate-Adaptation Logic for Volumetric Video. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652226

1 INTRODUCTION

The most recent evolution of the media has been the development of volumetric video, in particular, MPEG Immersive Video (MIV). The MIV standard defines a framework for coding and representation of immersive visual content, also known as volumetric video. MIV is a part of the family of Visual Volumetric Video-based Coding (V3C) standards. More details about the V3C standards are available in [2, 5, 6, 8, 9]. As with any other change, the MIV standard introduces unique challenges to overcome. One challenge is to adapt to the irregular bitrate vs. quality relationship among different components of MIV content, such as texture and geometry.

MIV content is created using multiple camera views and depth maps from a 3D environment. Multiple view-depth map pairs captured from different perspectives in the scene are then pruned to remove the inter-view redundancies. Each resulting pruned video is called an *atlas*. The MIV main profile has two types of atlases: geometry and texture.

The geometry atlases are created from the depth maps, as illustrated in Figure 1. This atlas type contains the depth information of the objects in the scene. It is usually significantly smaller than the attribute atlas in size since it requires only one channel. Effects of the quality of this atlas are most evident when there is parallax in the scene. The texture atlases are created using the camera views. This atlas type contains information about the color of the pixels in the scene. The texture component typically has the largest bitrate among all components.

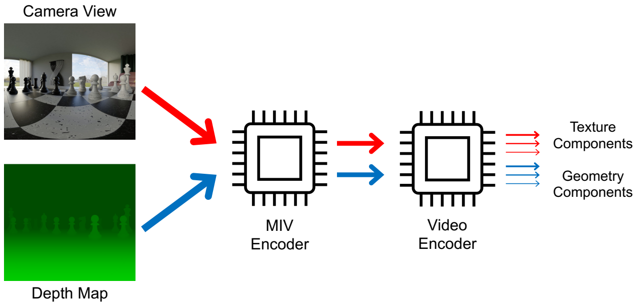


Figure 1: An MIV encoding pipeline.

The atlases are essentially 2D video bitstreams. Since V3C is codec agnostic, any existing 2D video codec, *e.g.*, AVC, HEVC, VVC, can be used to encode the atlases. Encoded atlases are then sent to the client and decoded to render a 3D environment using the texture and depth information about the scene. For adaptive streaming, the atlases can be encoded in multiple qualities to stream the content in a rate-adaptive manner. In such a scenario, the client can independently request the geometry and texture components, as illustrated in Figure 2. Different quality geometry and texture components can be combined to achieve different overall qualities.

We argue that the combined bitrate of a set of MIV components and the overall quality resulting from this combination are not always positively correlated. In our testing, we observed that some combinations could achieve very close or even better picture quality compared to another combination despite totaling a significantly lower bitrate. In such cases, the geometry component was the deciding factor for the overall quality. Some parts of an MIV content

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652226

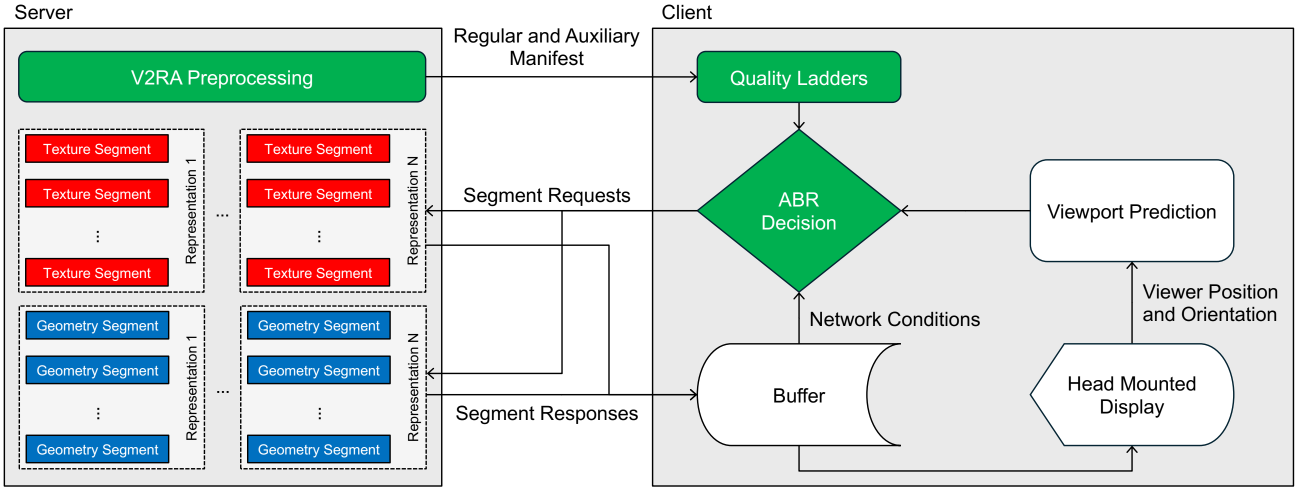


Figure 2: Workflow for adaptive streaming of MIV content. V2RA's main components are shown in green color.

composed of lower-quality texture and higher-quality geometry components could outperform those composed of higher-quality texture and lower-quality geometry components in terms of overall quality despite being smaller in size. This implies that when making adaptive bitrate decisions, we cannot assume that higher bitrate components will have a higher quality. Since there are cases where MIV content can be delivered in higher quality using less bandwidth, adaptation decisions need to be made smartly based on the content and the user viewport.

Current adaptive bitrate (ABR) algorithms work on the assumption that higher bitrates lead to better quality. This causes the common ABR algorithms to make sub-optimal decisions when used for MIV content. To prevent such decisions, the overall quality of different component combinations must be assessed beforehand to create a *quality ladder*. Since the quality of the content can differ significantly for different viewports even when using the same components, the quality ladders must be created in a viewport-dependent fashion.

In this paper, we propose a preprocessing step to create quality ladders for a subset of possible viewports. Since measuring the quality of every possible viewport is not viable, we sample the whole viewing space using some representative viewports on a grid. By rendering the content using different quality component combinations for these viewports on the grid, we can approximate a quality ladder for any viewport. After generating the quality ladders, we demonstrate how these ladders can be used to stream MIV content more efficiently. The proposed rate-adaptation logic is called Volumetric Video Rate Adaptation (V2RA) and its main components are shown in Figure 2.

2 RELATED WORK

Streaming immersive video typically requires higher resolution (4K or more) and frame rates (at least 60 fps) than 2D video to provide a satisfactory user experience and avoid motion sickness that may be induced if the viewing experience is not in sync with the user's movement. However, in most cases, only a part of the scene is visible to the user at a given time. Therefore, streaming only the part of the scene that lies inside the user's viewport or streaming the background with lower quality/resolution (viewport-dependent streaming) is a feasible solution to achieve high bandwidth savings. For example, in [10], the visible parts of the atlases inside the user's viewport were extracted to reduce the video bitrate. [18] demonstrated how VVC [7] subpictures could be used to implement viewport-dependent streaming for volumetric video. [11] developed a visual saliency-based tiling and QoE-based transmission scheme for optimal transport of volumetric video.

One of the most critical aspects of viewport-dependent streaming is viewport prediction. As the segments are requested for a future viewport, correctly predicting the future viewports is the key to achieving the full potential of viewport-dependent streaming. Thus, viewport prediction has been an active research field for 360-degree videos. Some approaches used the past data for prediction (*e.g.*, [14, 19, 20]), while others used models for head motion (*e.g.*, [15, 21]) or learning algorithms (*e.g.*, [13, 22]). Some techniques have also been developed for six degrees of freedom (6DoF) [4].

Content saliency in a 2D video is not equally distributed over the whole scene, which is also true for 3D videos. By extracting the salient regions in a 3D scene and allocating more bits to such regions, the available bandwidth can be used more effectively to increase the quality of the regions the user's attention is directed to. The authors of [17] explored the advantages of such delivery methods for 3D video. [16] evaluated visual attention models for omnidirectional videos using publicly available testbed and subjective user data.

3 THE V2RA LOGIC

In this section, we introduce V2RA, a new rate-adaptation logic for streaming MIV content. It uses a grid to approximate the viewport and makes ABR decisions based on the approximated viewport. To achieve this, we draw a grid in the viewing space. Then, we render the video for every point on the grid for multiple orientations, using every possible quality combination of the available components.

After the quality assessment of all the renders, a quality ladder is created for every point and a set of pre-determined orientations on the grid. The combination with the lowest bitrate with a quality score close enough to the highest quality available is chosen as the preferred combination for each segment. The desired closeness to the highest available quality can be adjusted for different content and applications. When viewing the content, the client finds the closest grid point to itself and the closest pre-determined orientation and uses the quality ladder for that point and orientation.

3.1 Grid-Based Viewport Approximation

To use a viewport-dependent approach for rate adaptation, we first need to be able to compute the user's viewport efficiently. However, matching the exact pixels within the MIV atlases to viewports is difficult because of the complex nature of the MIV encoding and complications, such as occlusions. Hence, we reduce the possible viewing space to a smaller representative subset of viewports. The viewports included in this set are evenly spaced on the intersections of an n-by-m grid. For each grid point, *i.e.*, the intersections of vertical and horizontal lines, there can be multiple viewports because of different orientations on the same grid point. In Figure 3, a top-down view of a 4-by-5 grid is depicted. In this example, there are four orientations for each point on the grid, with the yaw values of 0, 90, 180 and 270 degrees.

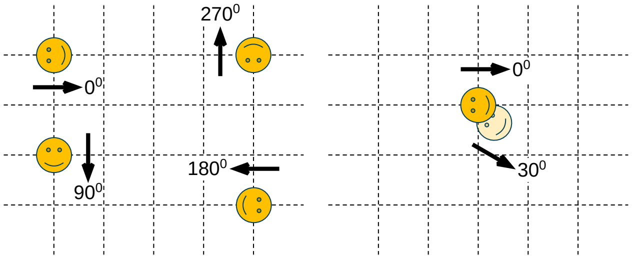


Figure 3: An example 4-by-5 grid with four orientations for each intersection (left) and viewport approximation (right).

After setting the viewport positions and creating the quality ladder, as described in Section 3.2, the client needs to approximate its actual viewport to one of the viewports on the grid. The Euclidean distance between the viewer position and the grid intersections is calculated to find the closest viewport. After that, the viewport on that point with the closest orientation to the actual viewport is chosen as the user's approximated viewport, and the calculated quality ladder for that viewport is used when making ABR decisions. In Figure 3, the actual viewport is illustrated as the semi-transparent smiling face, facing 30 degrees clockwise from the x-axis. The opaque smiling face represents the approximated viewport for the actual pose. It is on the closest grid intersection and the orientation it faces is the closest to 30 degrees.

3.2 Quality Ladder

In 2D video, a bitrate ladder is used to switch the quality up and down as the available bandwidth changes. However, for MIV content, the relationship between bitrate and quality may not be as straightforward as it is for 2D video. Since an MIV sequence consists of different video components such as texture and geometry, the joint effect of adapting the bitrate of each of these components on the quality of the reconstructed video needs to be considered. Especially in cases where a higher-quality texture component and a lower-quality geometry component are paired, they usually consume more bandwidth than a lower-quality texture, higher-quality geometry pair. However, in some cases, the latter has superior quality despite consuming less bandwidth. In some other cases, the small quality gain by changing between the two does not justify the bandwidth waste. Therefore, we propose a quality ladder consisting of the objective quality metrics calculated for the viewports on the grid. The video is rendered for each viewport on the grid and an objective quality score is calculated for every possible combination.

The quality ladder is then sent to the client for the decision process. There are multiple ways to relay this information to the client. The most likely scenario would be to send the quality ladders for every segment as an auxiliary manifest upfront. The SARA algorithm proposed in [1] uses a similar auxiliary manifest to communicate the specific segment sizes. A similar approach for the manifest delivery can be used for the quality ladders.

After receiving the manifest, the client can use the quality ladder for the viewport belonging to the closest grid point to make the decisions. In order to achieve high quality using a reasonable bandwidth, the client requests the segments for each component of the MIV bitstream (geometry and texture components in our experiments) with the lowest total bitrate that also has an overall quality close enough (within a specified margin) to the highest possible one for the particular available bandwidth. The pseudocode for pair selection logic is given in Algorithm 1.

Algorithm 1 Select a Pair of Segments

```
procedure SelectPairOfSegments(availablePairs, margin)
    $bestPair \leftarrow$ maxQuality($availablePairs$)
    $candidates \leftarrow \{\}$
    $chosenPair \leftarrow bestPair$
    for each $pair$ in $availablePairs$ do
        $difference \leftarrow \left(\frac{bestPair.\text{quality} - pair.\text{quality}}{bestPair.\text{quality}}\right)$
        if $difference \leq$ margin then
            $candidates$.append($pair$)
        end if
    end for
    for each $pair$ in $candidates$ do
        if $pair$.bitrate $<$ $chosenPair$.bitrate then
            $chosenPair \leftarrow pair$
        else if $pair$.bitrate == $chosenPair$.bitrate and $pair$.quality > $chosenPair$.quality then
            $chosenPair \leftarrow pair$
        end if
    end for
    return $chosenPair$
end procedure
```

4 EXPERIMENTAL RESULTS

4.1 Setup

To evaluate the performance of our approach, we simulated the adaptation process. We used the Chess video, containing multiple moving and stationary objects with varying distances to the viewer, and the Classroom video, containing a smaller number of objects and a more dynamic texture but no movement in geometry. The dynamic elements of the texture components of the Classroom video are lighting changes and some film grain that increases the texture component’s size. We used different quantization parameters (QP) for each component for different quality levels. QP values of 4 and 11 were used for the geometry component, and for the texture component, QP values of 22 and 27 were used. These values were selected among the empirical values used in [2]. All components were encoded using an HEVC encoder[1] with the following parameters:

- Segment duration: 16 frames
- Group of pictures (GoP) duration: 16 frames
- Frame rate: 30 fps
- Video duration: 10 seconds (Chess) and four seconds (Classroom)

The grid was drawn around the 3-by-3 square around the origin. Every grid point was 0.25 units away from its neighbors. For each position, there were four rotations 0, 90, 180 and 270 degrees around the y-axis. Each grid pose was rendered with a slight uniform motion to account for the effects of parallax. We first rendered the necessary grid poses with all quality combinations and computed their qualities. Then, we created a quality ladder for the ABR decisions. Finally, we rendered videos using natural pose traces and compared the total bandwidth usage and the resulting overall quality by (*i*) using the quality ladder, and (*ii*) blindly picking the highest possible bitrate.

Since we did not have the means to stream and render MIV content in real time, the decision process was simulated after rendering the video using a previously known pose trace. In practice, the viewport belonging to the closest grid point cannot be known before sending a request. Therefore, this closest grid point and the user orientation must be predicted before selecting the most suitable pair of segments[2].

Two quality metrics were used for the experiments in this paper: Immersive Video Peak Signal-to-Noise Ratio (IV-PSNR) and Video Multimethod Assessment Fusion (VMAF). To assess the quality of the MIV content, we first took renders using pose traces and evaluated the resulting viewport.

- IV-PSNR is a specialized metric designed for assessing the quality of 3D video content adjusted for the common artifacts in 3D video [3]. It is a weighted PSNR calculation with weights tuned specifically for 3D video content. IV-PSNR is particularly relevant in 3D video rendering, where the dynamic nature of the content influences the viewer’s experience.
- VMAF is a comprehensive quality metric developed by Netflix that combines multiple assessment methods to provide a unified score for video quality evaluation [12]. It integrates various perceptual quality metrics, including spatial and temporal aspects, to emulate human perception accurately.

4.2 Results

Table 1 shows the calculated IV-PSNR scores for the segments of the Chess video. The values outside the $[-1.5\%, 0]$ range are highlighted with a colored background. A similar table for the VMAF scores of the Chess video’s segments is given in the Appendix (see Table 2). The overall sizes for both the texture and geometry components are shown in Figure 4 for various QP combinations.

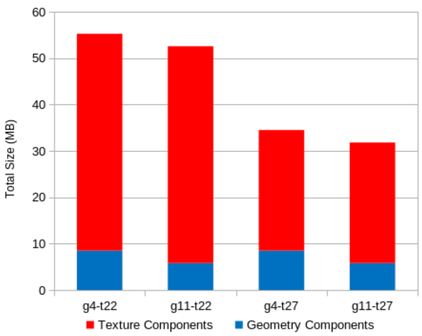


Figure 4: The total size of the component combinations for the Chess video (10 seconds).

The V2RA logic selects the pair of segments with the lowest total bitrate within the desired margin of the highest possible overall quality for the available bandwidth. Since only the qualities of the grid positions are known beforehand, decisions are made based on the quality of the closest grid position and not the viewer’s actual pose trace. The grid positions are named after the x and y coordinates and the yaw angles in that order. The closest grid positions for Pose 01 are 25_0_0 for the first four segments, 0_25_90 for the subsequent four segments, -25_0_180 for the following seven segments and 0_-25_270 for the last four segments.

The initial hypothesis was that the quality levels of the combinations were not directly correlated with the overall bitrate of the videos. We see that this hypothesis holds for several segments. For example, in Table 1 under the last (Pose 01) column, the fourth through the ninth segments have higher quality for the g4-t27 combination than the g11-t22 combination despite having a lower bitrate. Another example in Table 1 is under the 0_25_90 column. Most of the IV-PSNR scores are higher for the g4-t27 combination compared to the g11-t22 combination despite having a lower bitrate. In such cases, switching to a higher bitrate can reduce the overall quality in addition to wasting bandwidth.

When the available bandwidth is insufficient for the highest bitrate, we observe V2RA’s superiority over the blind approach. By choosing the combination with the highest approximated quality among the available combinations instead of the highest bitrate one,

[1] FFmpeg: Available at https://ffmpeg.org/

[2] In this study, we assume perfect viewport prediction to better demonstrate the effects of the proposed rate-adaptation logic.

	-25_0_180				0_-25_270				0_25_90				25_0_0				Pose 01			
GoP	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**
1	51.7	51.7	51.0	51.1	49.0	48.6	48.5	48.3	49.6	49.2	49.2	48.9	52.4	52.1	51.8	51.6	54.6	54.2	53.4	53.4
2	46.8	46.8	46.4	46.5	48.3	48.0	47.9	47.7	46.9	46.7	46.7	46.5	50.3	50.0	50.1	49.8	52.6	52.2	52.0	51.8
3	48.0	48.0	47.6	47.6	48.7	48.5	48.4	48.3	47.3	47.1	47.1	46.9	51.0	50.6	50.6	50.3	49.6	49.4	49.3	49.1
4	49.0	48.9	48.3	48.3	49.4	49.2	48.8	48.7	49.2	48.8	48.9	48.5	52.4	52.3	51.9	51.8	49.7	49.3	49.4	49.2
5	53.0	52.7	51.7	51.6	49.4	49.3	49.0	48.9	50.7	50.3	50.3	49.9	53.3	53.0	52.6	52.4	49.9	49.5	49.7	49.3
6	49.7	49.4	49.2	49.0	48.6	48.4	48.1	48.0	49.4	49.1	49.2	48.9	52.1	51.9	51.5	51.4	51.7	51.2	51.3	50.8
7	52.3	51.9	51.6	51.4	48.8	48.7	48.4	48.3	49.4	48.7	49.1	48.5	52.9	52.9	52.2	52.3	51.7	50.7	51.3	50.3
8	50.1	50.1	49.4	49.4	49.8	49.7	49.3	49.2	49.4	48.8	49.2	48.5	53.2	53.0	52.6	52.4	48.7	48.2	48.4	48.0
9	51.7	51.6	50.9	51.0	49.5	49.3	49.1	49.0	49.3	49.0	48.9	48.7	52.7	52.0	52.1	51.5	48.6	48.1	48.3	47.9
10	46.9	46.9	46.6	46.6	47.9	47.8	47.6	47.5	46.7	46.5	46.5	46.3	49.5	48.6	49.2	48.4	50.8	50.5	50.5	50.2
11	48.7	48.7	48.3	48.3	48.5	48.4	48.2	48.1	46.3	46.1	46.1	45.9	51.2	50.8	50.6	50.4	51.5	51.0	51.2	50.7
12	50.1	50.0	49.4	49.4	49.9	49.7	49.4	49.3	47.6	47.3	47.4	47.0	51.4	51.1	50.7	50.6	53.3	52.9	52.7	52.4
13	52.6	52.0	52.0	51.6	49.1	49.0	48.7	48.6	49.6	49.2	49.3	48.9	48.4	47.7	47.8	47.2	53.6	53.3	52.3	52.2
14	49.7	49.6	49.2	49.2	48.3	48.2	47.9	47.8	48.8	48.4	48.6	48.2	49.6	49.3	49.2	48.9	48.9	48.6	48.4	48.3
15	51.8	51.6	51.3	51.2	48.4	48.3	48.1	48.0	48.3	47.7	48.0	47.4	45.3	45.3	45.2	45.1	47.9	47.9	47.5	47.5
16	50.2	50.1	49.5	49.5	49.5	49.3	49.0	48.8	49.0	48.3	48.7	48.0	49.8	49.4	49.5	49.1	48.9	48.8	48.6	48.5
17	51.7	51.7	51.1	51.1	49.4	49.2	49.0	48.9	49.2	49.0	48.9	48.7	49.7	49.0	49.4	48.8	50.2	49.8	49.9	49.5
18	46.7	46.7	46.4	46.4	47.9	47.8	47.6	47.5	46.5	46.3	46.4	46.1	50.0	49.7	49.7	49.4	52.5	52.2	52.1	51.7
19	47.8	47.8	47.5	47.5	48.0	47.9	47.8	47.6	46.8	46.7	46.6	46.5	48.8	48.8	48.6	48.5	47.9	47.6	47.4	47.1

Table 1: IV-PSNR scores for the Chess video segments. Red indicates values outside $[-1.5\%, 0]$ range.

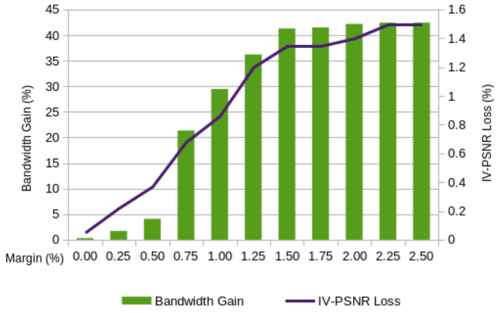


Figure 5: Bandwidth gain and IV-PSNR score loss for different IV-PSNR margins for the Chess video.

we can save 11.21% (and increase quality by 0.02%) or 5.88% (and decrease quality by 0.03%) bandwidth, if we make a decision based on IV-PSNR or VMAF, respectively. The reason for the increase in the IV-PSNR score (despite using less bandwidth) is that the blind approach preferred the g11-t22 combination (higher bitrate) over the g4-t27 combination (lower bitrate) even when the latter had a higher quality.

Another implication of this irregular bitrate-quality relationship is a less drastic one. Sometimes, switching to a higher bitrate combination may increase the quality, but the increase in quality may not be significant enough to justify the extra bandwidth to make the switch. The proposed approach is to pick the lower-quality combinations if their qualities are within a certain range of the highest available quality for the available bandwidth. By changing the acceptable quality loss margin, we can analyze the bandwidth gain vs. quality loss tradeoff.

Figure 5 shows the bandwidth gain and IV-PSNR score loss by changing the acceptable quality margin when no bandwidth restrictions are assumed. The suitable margin for any desirable application and content may vary; therefore this analysis must be conducted for different use cases. The plot for the same analysis on VMAF can be found in the Appendix (see Figure 6).

By choosing the lowest bitrate combination with a VMAF score within the $[-3.5\%, 0]$ range from the highest score for the segment, we can achieve 38.97% bandwidth savings while only losing 2.08% of the overall quality. By using the IV-PSNR score as the deciding factor, and accepting qualities within the $[-1.5\%, 0]$ range from the highest quality for the segment, we can save 41.30% of the bandwidth while only losing 1.35% in the quality.

5 CONCLUSION AND FUTURE WORK

This paper introduces a new method to make better ABR decisions based on some pre-rendered scenes for MIV content. We created a testbed and evaluated the quality loss and bandwidth savings that can be achieved with this method. By making the switch decisions based on the predicted qualities of the viewport, we saw bandwidth savings up to 25% with only a 0.75% decrease in the IV-PSNR score.

The main purpose of the study was to find a better approach to creating a bitrate ladder, considering the peculiarities of the bitrate quality relationship in MIV content. Nevertheless, it is essential to validate these initial findings through subjective studies.

APPENDIX

Here we present the VMAF scores for the segments of the Chess video in Table 2 with the values outside a margin of 3.0% highlighted in red. For the same video, the corresponding bandwidth gain and VMAF score loss vs. the acceptable quality margin graph is presented in Figure 6. In addition, we present the same results for the Classroom video in Tables 3 and 4 and Figure 7.

REFERENCES

[1] A. C. Begen, M. N. Akcay, A. Bentaleb, and A. Giladi. Adaptive streaming of content-aware-encoded videos in dash. js. *SMPTE Motion Imaging Journal*, 131(4):30–38, 2022 (DOI: 10.5594/JMI.2022.3160560).

[2] J. M. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V. K. M. Vadakital, and L. Yu. MPEG immersive video coding standard. *Proc. IEEE*, 109(9):1521–1536, 2021.

[3] A. Dziembowski, D. Mieloch, J. Stankowski, and A. Grzelka. Iv-psnr—the objective quality metric for immersive video applications. *IEEE Trans. Circuits and Systems for Video Technology*, 32(11):7575–7591, 2022.

[4] B. Han, Y. Liu, and F. Qian. ViVo: visibility-aware mobile volumetric video streaming. In *ACM MobiCom*, 2020.

	-25_0_180				0_-25_270				0_25_90				25_0_0				Pose 01			
GoP	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**
1	91.2	90.6	88.0	87.4	86.7	84.7	84.1	82.0	94.6	93.1	93.0	91.6	94.4	93.4	92.2	91.2	96.8	96.1	94.1	93.3
2	86.0	85.1	83.2	82.4	82.7	80.6	80.4	78.4	90.1	88.4	88.7	87.1	92.7	91.8	90.7	89.8	100.0	99.9	98.7	97.9
3	88.0	87.2	84.9	84.2	84.3	82.2	82.2	80.4	92.7	91.2	91.1	89.4	93.0	92.0	90.8	90.0	100.0	99.9	99.4	98.6
4	90.8	90.0	87.6	86.8	86.2	84.6	83.5	81.9	94.5	93.0	93.0	91.6	94.3	93.3	91.9	90.9	100.0	100.0	100.0	99.7
5	94.4	93.6	90.1	89.3	85.2	83.5	83.1	81.5	96.5	95.4	94.8	93.7	92.2	90.4	89.9	88.1	100.0	100.0	100.0	99.9
6	89.3	86.8	85.4	83.0	82.2	80.7	80.1	78.7	94.0	92.3	92.5	90.9	90.5	89.0	88.1	86.6	100.0	99.6	99.1	97.7
7	92.5	91.0	89.2	87.7	84.2	82.5	82.2	80.5	94.6	92.1	92.6	90.3	91.7	90.6	89.5	88.5	99.4	97.5	97.4	95.2
8	92.1	91.0	88.8	87.7	85.6	84.2	83.3	81.9	95.2	92.9	93.5	91.2	93.2	92.1	90.9	89.8	95.8	93.8	93.8	91.9
9	91.4	90.5	88.4	87.5	85.5	83.5	83.5	81.6	95.8	94.6	94.2	92.9	89.3	87.5	86.7	84.9	94.6	93.4	92.3	91.0
10	86.6	85.6	83.8	82.8	81.5	80.1	79.4	78.0	89.7	88.1	88.3	86.6	89.5	86.5	86.7	83.8	98.5	97.5	95.9	95.1
11	88.1	87.0	85.3	84.2	83.0	81.7	81.1	79.7	92.7	91.0	91.0	89.3	91.5	89.6	89.0	87.2	99.8	99.4	97.9	97.1
12	91.0	89.7	88.0	86.8	84.2	83.1	82.0	80.8	94.7	93.0	93.0	91.3	94.1	92.1	91.4	89.5	100.0	100.0	99.3	98.5
13	92.2	90.7	89.1	87.6	83.2	81.6	81.0	79.6	96.2	94.4	94.5	92.8	87.3	84.0	83.5	80.3	100.0	100.0	98.6	97.7
14	90.2	89.3	87.5	86.6	81.2	79.6	78.9	77.4	94.0	91.4	92.5	90.0	88.9	87.1	85.8	84.0	95.7	94.4	92.8	91.5
15	92.3	91.5	89.4	88.7	82.6	81.3	80.6	79.3	93.4	90.0	91.5	88.1	89.5	88.7	87.5	86.6	93.5	92.3	91.6	90.4
16	92.0	91.3	88.9	88.2	83.9	82.7	81.4	80.2	95.4	92.8	93.7	91.0	95.5	94.2	93.0	91.7	92.7	91.3	90.8	89.3
17	91.3	90.7	88.2	87.6	84.1	82.9	82.1	80.7	95.1	93.8	93.5	92.3	93.5	90.5	91.1	88.0	93.3	91.8	90.8	89.5
18	85.6	85.2	83.0	82.6	80.6	79.7	78.4	77.7	89.0	86.9	87.6	85.5	93.7	92.3	91.4	90.0	90.3	89.0	87.7	86.5
19	86.7	85.9	84.0	83.2	81.9	80.8	79.9	79.2	91.3	89.7	89.6	87.9	92.3	91.1	90.2	89.1	82.9	81.7	80.4	79.3

Table 2: VMAF scores for the Chess video segments. Red indicates values outside [-3.0%, 0] range.

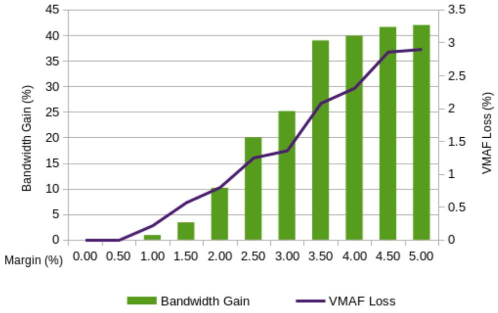


Figure 6: Bandwidth gain and VMAF score loss for different VMAF margins for the Chess video.

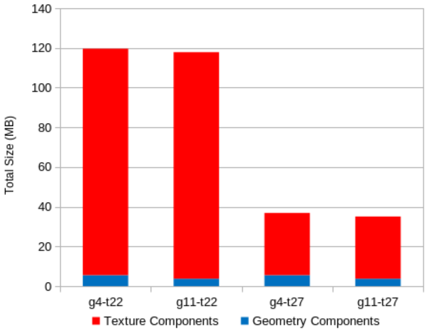


Figure 7: The total size of the component combinations for the Classroom video (four seconds).

	0_0_0				Pose 01			
GoP	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**
1	48.0	47.9	48.0	47.9	50.1	50.1	50.4	50.4
2	46.5	46.5	46.2	46.2	48.5	48.5	48.8	48.8
3	47.3	47.3	47.5	47.4	48.1	48.1	48.4	48.4
4	47.5	47.5	47.8	47.8	48.0	48.0	48.4	48.4
5	48.2	48.2	48.1	48.2	49.7	49.8	49.9	49.9
6	46.8	46.7	46.7	46.7	49.6	49.7	49.9	49.9
7	47.4	47.3	47.2	47.2	50.3	50.3	50.2	50.2
8	40.6	40.4	40.5	40.3	43.7	43.6	43.6	43.5

Table 3: IV-PSNR scores for the segments of the Classroom video.

	0_0_0				Pose 01			
GoP	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**	**g4-t22**	**g11-t22**	**g4-t27**	**g11-t27**
1	93.6	91.7	93.6	91.6	88.3	86.0	88.2	86.0
2	91.8	89.9	91.8	89.9	86.8	84.9	86.8	84.9
3	89.5	87.8	89.5	87.8	88.1	86.4	88.1	86.4
4	88.4	86.5	88.4	86.5	96.5	94.6	96.6	94.6
5	90.0	87.1	90.1	87.2	97.1	94.1	97.1	94.1
6	88.0	85.6	88.2	85.7	95.6	92.9	95.6	92.9
7	89.1	86.7	89.4	87.0	99.9	98.6	99.9	98.6
8	88.3	85.5	88.5	85.7	100.0	99.7	100.0	99.7

Table 4: VMAF scores for the segments of the Classroom video.

[5] S. S. Ilola L., Kondrad L. and H. A. An overview of the MPEG standard for storage and transport of visual volumetric video-based coding. *Front. Signal Proc.*, 2, 2022.

[6] ISO/IEC. ISO/IEC 23090-10:2022 Information technology – Coded representation of immersive media Part 10: Carriage of visual volumetric video-based coding data. [Online] Available: https://www.iso.org/standard/78991.html, 2022. Accessed on Jan. 26, 2024.

[7] ISO/IEC. ISO/IEC 23090-3:2022 Information technology – Coded representation of immersive media Part 3: Versatile video coding. [Online] Available: https://www.iso.org/standard/83531.html, 2022. Accessed on Jan. 26, 2024.

[8] ISO/IEC. ISO/IEC 23090-12:2023 Information technology – Coded representation of immersive media Part 12: MPEG immersive video. [Online] Available: https://www.iso.org/standard/79113.html, 2023. Accessed on Jan. 26, 2024.

[9] ISO/IEC. ISO/IEC 23090-5:2023 Information technology – Coded representation of immersive media Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC). [Online] Available: https://www.iso.org/standard/83535.html, 2023. Accessed on Jan. 26, 2024.

[10] S. Lee, J.-B. Jeong, and E.-S. Ryu. Implementing partial atlas selector for viewport-dependent MPEG immersive video streaming. In *ACM NOSSDAV*, 2023.

[11] J. Li, C. Zhang, Z. Liu, R. Hong, and H. Hu. Optimal volumetric video streaming with hybrid saliency based tiling. *IEEE Trans. Multimedia*, 25:2939–2953, 2023.

[12] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, et al. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[13] P. Maniotis and N. Thomos. Viewport-aware deep reinforcement learning approach for 360 video caching. *IEEE Trans. Multimedia*, 24:386–399, 2021.

[14] A. T. Nasrabadi, A. Samiei, and R. Prakash. Viewport prediction for 360 videos: a clustering approach. In *ACM NOSSDAV*, 2020.

[15] A. Nguyen, Z. Yan, and K. Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *ACM Multimedia*, 2018.

[16] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *IEEE QoMEX*, 2018.

[17] B. Salahieh, W. Cochran, and J. Boyce. Delivering object-based immersive video experiences. *Electronic Imaging*, 2021(18):103–1, 2021.

[18] M. Santamaria, V. K. Malamal Vadakital, L. Kondrad, A. Hallapuro, and M. M. Hannuksela. Coding of volumetric content with MIV using VVC subpictures. In *IEEE MMSP*, 2021.

[19] L. Sun, Y. Mao, T. Zong, Y. Liu, and Y. Wang. Flocking-based live streaming of 360-degree video. In *ACM MMSys*, 2020.

[20] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360 immersive videos. In *IEEE CVPR*, 2018.

[21] H. Yuan, S. Zhao, J. Hou, X. Wei, and S. Kwong. Spatial and temporal consistency-aware dynamic adaptive streaming for 360-degree videos. *IEEE Jour. Selected Topics in Signal Processing*, 14(1):177–193, 2019.

[22] X. Zhang, G. Cheung, Y. Zhao, P. Le Callet, C. Lin, and J. Z. Tan. Graph learning based head movement prediction for interactive 360 video streaming. *IEEE Trans. Image Processing*, 30:4622–4636, 2021.

Human Trajectory Forecasting in 3D Environments: Navigating Complexity under Low Vision

Franz Franco Gallo*
franz.franco-gallo@inria.fr
Université Côte d'Azur, Inria
Sophia-Antipolis, France

Hui-Yin Wu
Université Côte d'Azur, Inria
Sophia-Antipolis, France
hui-yin.wu@inria.fr

Lucile Sassatelli
Université Côte d'Azur, CNRS, I3S.
Institut Universitaire de France
Sophia-Antipolis, France
lucile.sassatelli@univ-cotedazur.fr

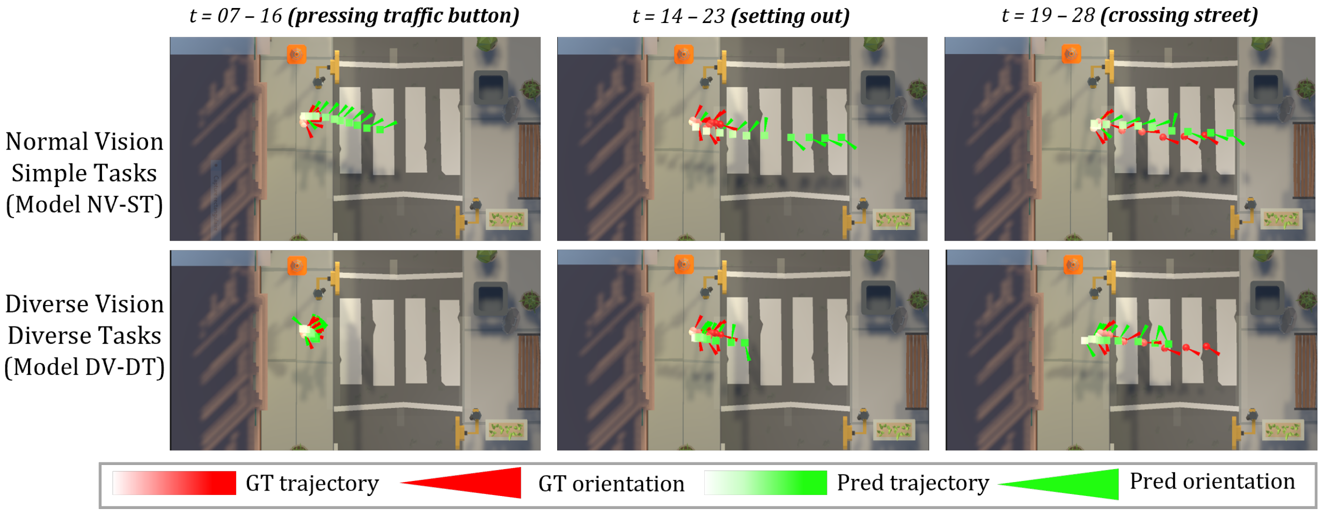

t = 07 – 16 (pressing traffic button)
t = 14 – 23 (setting out)
t = 19 – 28 (crossing street)
Normal Vision Simple Tasks (Model NV-ST)
Diverse Vision Diverse Tasks (Model DV-DT)
GT trajectory
GT orientation
Pred trajectory
Pred orientation

Figure 1: Predicted trajectory and orientation of two GIMO model training variations evaluated in a low-vision, complex task scene. The diverse model takes into account complex interactions (pressing traffic button), and adapts to user walking speed.

ABSTRACT

This work tackles the challenge of predicting human trajectories while carrying out complex tasks in contextually-rich virtual environments. We evaluate the CREATIVE3D multimodal dataset on human interaction and navigation in 3D virtual reality (VR). In the dataset, navigating traffic crossings with simulated visual impairments are used as an example of complex or unpredictable situations. We establish evaluations for a base multi-layer perceptron (MLP) and two state-of-the-art models: TRACK (RNN) and GIMO (transformer), on tasks with varying levels of complexity and visual impairment conditions. Our findings indicate that a model trained on normal visual conditions and simple tasks does not generalize on test data with complex interactions and simulated visual impairments, despite including 3D scene context and user gaze. In comparison, a model trained on diverse visual and task conditions is more robust, with up to 84% decrease in positional error and 9% in orientation error, but with the trade-off of lower accuracy for simpler tasks. We believe this work can benefit real-world applications such as autonomous driving, and enable context-aware computing for diverse scenarios and populations.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Motion capture**; **Activity recognition and understanding**.

KEYWORDS

human motion prediction, virtual reality, context, RNN, transformer

ACM Reference Format:
Franz Franco Gallo, Hui-Yin Wu, and Lucile Sassatelli. 2024. Human Trajectory Forecasting in 3D Environments: Navigating Complexity under Low Vision. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652223

1 INTRODUCTION

Human trajectory forecasting aims to predict future human movements and is of strong interest especially for high-stake scenarios such as pedestrian behavior prediction and understanding in self-driving applications [5]. The complexity of pedestrian behavior,

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04
https://doi.org/10.1145/3652212.3652223

characterized by their individual intentions, interactions with other pedestrians, vehicles and the environment, present a significant challenge for autonomous systems [16]. Acknowledging these complexities, our work focuses on ensuring that models for autonomous vehicles consider the diverse behaviors of pedestrians, including those with vision inequalities and their interactions with traffic lights, to enhance prediction accuracy and fairness.

Multiple approaches to human motion prediction have been proposed using machine learning and deep learning techniques, with an unsolved challenge of efficiently taking into account scene and social context [1]. In this work, we are the first to approach model performance for pedestrian trajectory and attention prediction under the light of fairness for the visually impaired. We do so thanks to virtual reality (VR) technologies where realistic scenarios can be simulated to investigate human behaviour in-context. One such dataset is the CREATTIVE3D multimodal dataset of user behavior in VR [15][1] which collects user behaviours in complex tasks such as seeking, taking, and transporting objects, and real walking in simulated road intersections [12], including conditions with simulated visual impairments – a virtual scotoma (area in the central visual field with little or no acuity).

We present evaluations of motion prediction models on this newly introduced dataset to identify key weaknesses of existing reference prediction models and research challenges ahead. Specifically, our contributions are:

- We identify how brittle the models are in case of distribution shifts, that is (1) when training on normal-vision data and predicting on low-vision data (up to 90% and 8% error increase in position and attention prediction, respectively), (2) when training on simple tasks and predicting on complex tasks (up to 900% and 44% error increase), and
- We show how a diverse training set with different types of vision conditions and tasks can alleviate the performance unfairness. We reveal that the models exhibit performance trade-offs between the different populations and scenarios when trained on a diverse and balanced dataset (e.g., error increase of up to 25% and 10% for position and attention of normal vision data, 72.3% and 1% for simple tasks), hence exhibiting their inability to properly condition the output based on context. We propose future important research avenues based on the findings.

We first present the related work in Sec. 2. We then introduce our models and testing conditions in Sec. 3 and present the results in Section 4. Finally, we provide a discussion in Sec. 5 and conclusions with the future research avenues Sec. 6.

2 RELATED WORK

Predicting human motion or attention trajectory from multiple modalities has been a long-standing endeavor in various application scenarios (pedestrian trajectory forecast for self-driving, optimization of VR rendering, etc.). We briefly discuss and position our approach within the existing prediction models, and the available datasets with varied contextual conditions (type of vision, type of tasks, physical environment) and representation (unstructured such as point cloud, or structured such as scene graphs).

2.1 Models for human motion prediction

The prediction of human motion is adequately approached as a sequence-to-sequence problem, with prior movements providing the basis for forecasting subsequent sequences, and possibly informed by the context. Current models employ a variety of architectures, notably Recurrent Neural Networks (RNN), Graph Convolutional Networks, Generative Adversarial Networks , and Transformers. An example of a multimodal RNN-based prediction model is TRACK, which predicts attention in 3 Degrees of Freedom (DoF) VR [13]. Leveraging correlations within a single modality and across several modalities has known substantial progress with Transformers [14], fueling so-called cross-modal Foundation Models that are pre-trained on large-scale datasets [9, 11]. An example of human motion prediction using attention mechanisms to exploit spatial and temporal correlation between joints is STTran[2].

However, transformers are plagued with quadratic complexity in the size of the input, often high-dimensional for images, videos and text, incurring heavy computational costs both in train and test. Several approaches aim to counteract the high complexity, amongst which the family of Perceiver models [8], avoiding computationally-heavy self-attention on a high-dimensional input. Recently, Zheng et al. introduced a Perceiver-based architecture for motion prediction in 6 DoF VR, named GIMO. The GIMO model [17] exploits gaze data from an eponym dataset to improve human motion prediction (center of gravity displacements and positions of joints).

In the present work, we consider TRACK and GIMO as reference representatives of RNN-based and Transformer-based models for motion prediction in VR. To our knowledge, no existing prediction model has neither considered the impact of low vision on prediction accuracy, nor that of different tasks.

2.2 Datasets for human motion prediction

The context in which actions are carried out by people, including the vision condition, environment and tasks, can be represented in three forms: images (2D), point clouds (3D), and scene graphs. The endeavor to accurately model human motion is extensively pursued through the utilization of high-caliber motion capture datasets. These range from the more compact CMU Graphics Lab motion capture database[4] to large collections like AMASS [10] and Human3.6M dataset [7]. The latter is distinguished by its high-quality motion capture with a multi-view camera system, establishing itself as a benchmark for motion prediction and 3D pose estimation. For rich contexts, datasets such as GIMO [17] and CIRCLE [3] have emerged taking advantage of virtual and augmented reality technologies, concentrating on simple tasks like reaching for an object or navigating to a location.

Nevertheless, these datasets do not portray realistic interactions with the environment that are often chained and overlapping. The recent CREATTIVE3D dataset [15] addresses this gap having key interesting features to address our objective. Indeed, it is the largest dataset of human motion in context (over 2.6 million poses), it is captured in fully annotated and dynamic 3D scenes with multivariate – gaze, physiology, and motion – data, and it investigates the impact of simulated low-vision conditions using dynamic eye tracking under real walking and simulated walking conditions. It therefore allows the analysis of predicted pedestrian behavior disaggregated

[1] https://zenodo.org/doi/10.5281/zenodo.8269108

over simple and complex tasks, such as interacting with the traffic light before crossing, as depicted in Fig. 1, and over normal and simulated low-vision conditions. It also provides point clouds of the environments, which can be processed as input and incorporated into existing models such as TRACK and GIMO.

3 METHODS

We investigate how predictive models trained on normal-vision and simple navigation tasks perform on simulated low vision and higher task complexity at inference time. We introduce the dataset for this analysis and the models chosen for benchmarking.

3.1 Problem Definition

We consider predicting the future trajectory of a human, modeled by the head position and orientation in 3D space from past position and possibly context (depending on the models).

The human model comprises of, at any given time t (in frames), the head **position** $\mathbf{p}_t \in \mathbb{R}^3$, each component in meters, and head **orientation** $\mathbf{r}_t \in \mathbb{R}^3$ in Euler angles (roll, pitch, yaw). Head position represents the user's absolute position in the scene where they walk physically with a 1:1 ratio between the real and virtual distance in the 10 by 4 meters tracked space. Head orientation corresponds to the direction of the center of the headset field of view.

The problem consists in predicting a full motion sequence over a future horizon H, represented as $\mathbf{M}_{t+1:t+H} = \{(\hat{\mathbf{p}}_{t+1}, \hat{\mathbf{r}}_{t+1}), \ldots, (\hat{\mathbf{p}}_{t+H}, \hat{\mathbf{r}}_{t+H})\}$ from a given time t. We employ a sampling rate of 2 fps, utilize 3 seconds of past motion and gaze data for input, and aim to forecast motion for the subsequent 5 seconds. Specifically, for any time-stamp t, our prediction spans $\{\mathbf{M}_{t+s}\}_{s=1}^{H}$ for each time-step s across a horizon of $H = 10$. The model accounts for a past motion history $\mathbf{M}_{t-L+1:t}$ with $L = 6$.

3.2 The CREATTIVE3D multimodal dataset

We take advantage of our newly released CREATTIVE3D dataset [15] of human interactions and navigation in VR, specifically scenes of road crossings. The CREATTIVE3D dataset includes an extensive collection of simulated pedestrian behaviors, designed to capture a wide range of human activities, from basic motion to complex interactions with objects and urban infrastructure. Its richness lies in the multimodal data collected allowing for an in-depth analysis of pedestrian dynamics under varying conditions.

This dataset stands out due to its comprehensive multivariate data including gaze, physiology, and motion in fully-annotated dynamic 3D scenes. It explores the impact of simulated low-vision conditions, incorporating real-time eye tracking to simulate visual impairments. The dataset supports a broad spectrum of research, from cognitive studies to computational modeling for understanding human behavior in VR. The dataset includes 6 scenarios of two task complexities: simple tasks (ST) with only navigation, and complex tasks (CT) with simultaneous navigation and interaction. An example of a complex tasks consisting of interacting with the traffic light then crossing is shown in Fig. 1. Each scenario is further observed under two visual conditions: Normal Vision (NV) and simulated Low Vision (LV).

Training. We consider 4 types of models: trained on normal vision and simple tasks, as well as a combinations of low vision or complex task. Specifically, we designed training and validation datasets that (1) for the scenario, comprise of either simple tasks only (ST) or diverse simple and complex tasks (DT), and (2) either normal vision only (NV) or diverse normal and low vision (DV). The resulting four training and validation sets, along with the number of samples per scenario/visual condition are summarized in Table 1. Note that each sample across all models is unique to ensure a diverse and comprehensive dataset for model training and validation.

Table 1: Summary of models: training and validation sets

Model	Training Samples	Validation Samples
NV-ST	NV-ST: 251	NV-ST: 63
NV-DT	NV-ST: 139 NV-CT: 139	NV-ST: 35 NV-CT: 35
DV-ST	NV-ST: 139 LV-ST: 138	NV-ST: 35 LV-ST: 35
DV-DT	NV-ST: 139 NV-CT: 139 LV-CT: 138	NV-ST: 35 NV-CT: 35 LV-CT: 35

Test. To investigate the robustness of reference models to distribution shifts over vision conditions and task complexities, we consider 4 test sets (with number of samples) to assess their performance across the spectrum of tasks and visual conditions: Test NV-ST (78), Test NV-CT (44), Test LV-ST (42), Test LV-CT (43)

3.3 Prediction models

We evaluate three reference models for human motion prediction: MLP, TRACK, and GIMO, as depicted in Fig. 2 on their accuracy and robustness across various vision conditions and task complexities. The models take as input different feature vectors processed from scene point cloud, gaze point cloud, and pose data. As shown in top of Figure 2 using PointNet++ for feature extraction, we obtain a per-point feature map (F_P) and a global scene descriptor (F_O).

MLP baseline model [6]. includes fully connected layers, transpose operations, and layer normalization to merge information across frames effectively. Each MLP block has a fully connected layer and layer normalization, iteratively applied to capture the temporal dynamics in the motion sequence, as shown in Figure 2-(a). Our adaptation uses 4 MLP blocks, leveraging its ability to effectively model temporal dependencies for improved accuracy.

TRACK [13]. based on RNN, a sequence-to-sequence model where the past ego motion and scene features are each processed by individual LSTMs, before being fused by a third LSTM. As shown in Figure 2-(b), the scene context is the gaze-interpolated feature f_g. Given the per-point feature F_P, the gaze point feature f_g is computed through inverse distance-weighted interpolation [17], this interpolated gaze feature thus encapsulates relevant scene information, offering clues to deduce the subject's intention.

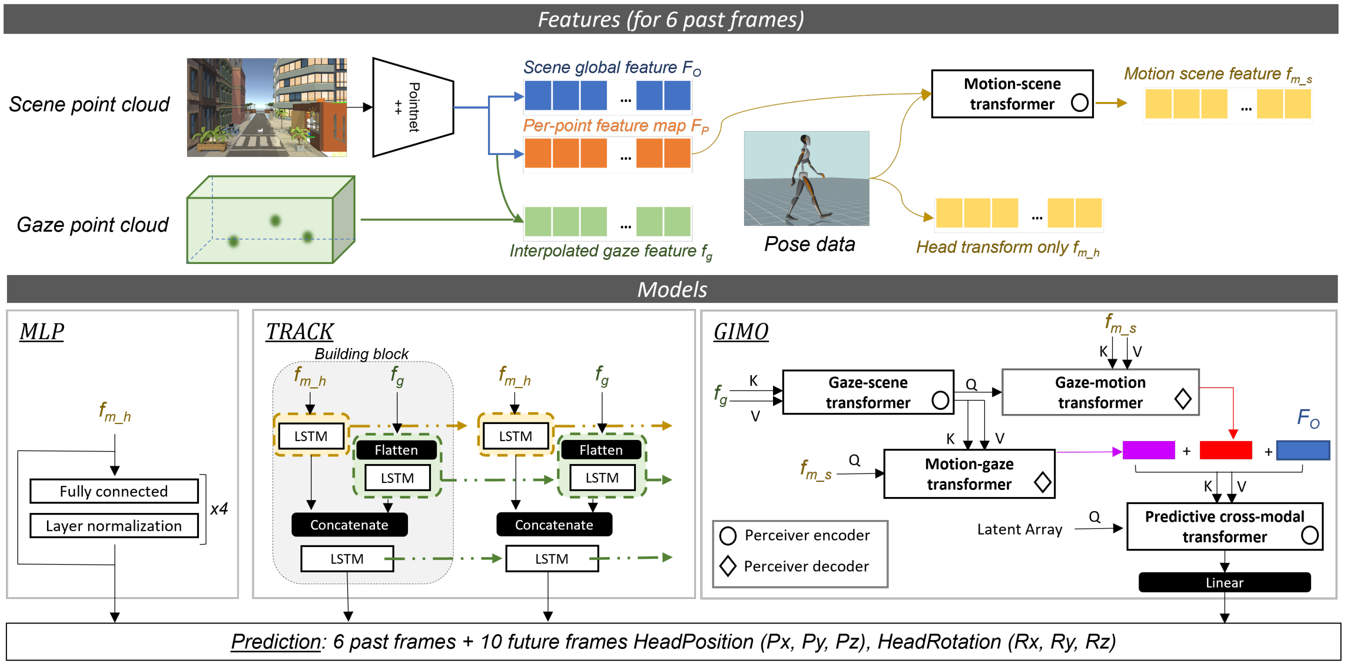


Figure 2: Our workflow takes into account scene, gaze, and human motion data, building different feature vectors. We evaluate the dataset on three models: a baseline MLP, TRACK (LSTM) and GIMO (transformer).

GIMO model [17]. a tranformer model composed of three cross-attention modules, where self-attention is first applied to the key-value modality, followed by cross-attention with the query modality (Figure 2-(c)). The modalities attending to each other are the position, the scene context around the body, and the scene context around the gaze target. All three latent vectors are then combined in a last cross-modal transformer, producing estimates of the future positions and orientations over a prediction horizon of 5 seconds with a history length of 3 seconds. The hyper-parameters are kept as in the original GIMO article.

Each of the three architectures is trained on the four training and validation sets detailed in the previous section. Each of the three architectures undergoes training on the four training and validation sets outlined in the preceding section. Each of the three architectures is subjected to learning processes on the four training and validation sets outlined in the preceding section.

3.4 Evaluation Metrics

For assessing prediction accuracy on position and head orientation, we use the following metrics:

Position error. We measure prediction error on position with the Mean Squared Error (MSE). MSE calculates the average distance, in square meters, between ground truth and predicted trajectory position across all time steps in the future horizon H (5 seconds).

Orientation error. The error prediction on head orientation is measured with the Orthodromic Distance (OD). OD quantifies the average angular distance, in radians, between ground truth and predicted orientations across all time steps in the future horizon H. The OD is defined as:

$$OD = \frac{1}{H}\sum_{i=1}^{H} 2\arccos(|\mathbf{r}_{t+i} \cdot \hat{\mathbf{r}}_{t+i}|) \qquad (1)$$

where $H = 10$ is the total number of predictions, $\mathbf{r}_{t+i}$ and $\hat{\mathbf{r}}_{t+i}$ are the unit quaternions representing the ground truth and predicted orientations for the t+ith prediction.

4 EXPERIMENTAL EVALUATION

In this section, we address the following research questions:

RQ1 How do the models compare to each other in different train-test configurations, and can we identify a superior model?

RQ2 To what extent can models trained on normal vision tasks maintain accuracy in low vision scenarios? Does refining the training dataset to reflect low vision test conditions optimize predictions, and what inherent model limitations does this approach reveal?

RQ3 How well do models designed for tasks under normal vision adapt to more complex challenges? Is the accuracy of predictions enhanced by aligning the training data with the complexities of the test environment, or does this strategy expose fundamental flaws in the models?

4.1 Global analysis

Table 2 shows the median values of MSE and OD for the position and orientation predictions respectively. GIMO has the lowest MSE values in all tests except LV-ST.

GIMO's architecture under simple tasks (NV-ST, LV-ST) has the lowest OD, with relatively consistent performance across different conditions. TRACK and MLP architectures on the other hand, seem sensitive to test conditions, as evidenced by fluctuating OD values.

Table 2: MSE and OD values for position and orientation on (1) three architectures (MLP, TRACK and GIMO), (2) four model variations (Table 1), and (3) on the four test sets.

	Tests	NV-ST		NV-CT		LV-ST		LV-CT	
Arch	Model	MSE	OD	MSE	OD	MSE	OD	MSE	OD
MLP	NV-ST	0.141	0.838	0.854	0.889	0.160	0.725	0.761	0.936
	NV-DT	0.718	0.805	0.267	0.877	0.285	0.704	0.214	0.819
	DV-ST	0.152	0.862	0.668	0.901	**0.142**	0.774	0.607	0.939
	DV-DT	0.744	0.797	0.261	0.854	0.520	0.691	0.252	0.833
TRACK	NV-ST	0.088	0.864	0.845	0.913	0.174	0.790	0.767	1.066
	NV-DT	0.284	0.669	0.250	0.697	0.260	0.709	0.201	0.777
	DV-ST	0.136	0.929	0.527	0.861	0.163	0.832	0.598	0.917
	DV-DT	0.260	0.631	0.193	0.685	0.184	0.616	0.188	0.732
GIMO	NV-ST	**0.083**	**0.557**	0.829	0.804	0.157	0.597	0.832	0.714
	NV-DT	0.143	0.562	**0.167**	0.688	0.225	**0.590**	0.186	0.679
	DV-ST	0.104	0.618	0.855	0.806	0.144	0.646	0.879	0.762
	DV-DT	0.223	0.649	**0.167**	**0.640**	0.180	0.646	**0.137**	**0.647**

Answer to RQ1: Under the NV-ST test for simple tasks, the MLP model displays moderate to high MSE values, achieving its best performance at $0.141m^2$ for the NV-ST model. TRACK demonstrates an improvement over MLP, while GIMO surpasses both MLP and TRACK in the NV-ST configuration by recording the lowest MSE of $0.083m^2$. Upon including low vision and complex tasks into the test, the performances of MLP, TRACK, and GIMO vary, with each showing their best results in DV-DT model. GIMO's architecture outperforms with both NV-ST and DV-DT models, offering the most accurate predictions. GIMO consistently exhibits lower OD values, which confirms it as the superior model across all tests.

4.2 GIMO analysis

We conduct a detailed examination of the GIMO architecture's performance. The whisker plots in Figure 3 show the MSE and OD distribution along the prediction horizon and the median estimation using the sliding window along the task sequence length, computed between the ground truth and predicted future motion across the four different training and validation sets for GIMO.

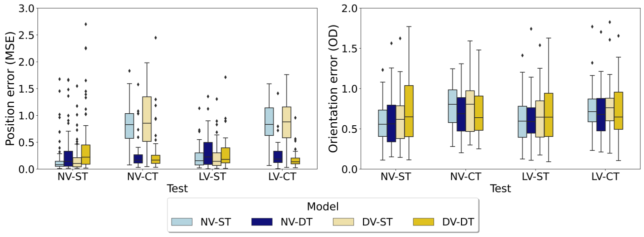


Figure 3: Comparative Analysis of Position and Orientation Errors in the GIMO Architecture: (Left) MSE Box plots for position estimation and (Right) OD for orientation estimation.

Model NV-ST stands out for its consistency in NV-ST test as depicted with Figure 3 (left), demonstrated by tight interquartile range (IQRs) and low median MSE value. However, wider IQRs in the NV-CT and LV-CT tests indicate significant prediction errors under complex conditions. Model NV-DT shows wider IQRs in NV-ST and LV-ST, reflecting greater MSE variability for simple tasks than NV-ST, however reducing the IQRs in NV-CT and LV-CT. Model DV-ST maintains narrow IQRs in NV-ST and LV-ST tests, indicating stable performance, but struggles with increased variability in NV-CT and LV-CT. Model DV-DT exhibits similar trends, with variable median MSE values and considerable outliers in NV-ST, NV-CT, and LV-CT, underscoring challenges in complex and low vision conditions. Overall, while NV-DT and DV-DT models offer accuracy and consistency, NV-ST and DV-ST highlight increased variability and occasional large errors, especially in complex task scenarios.

Across all models, the transition from simple to complex tasks tends to result in a slight increase in orientation error under both normal and low vision conditions. The variability of OD, as shown by the IQR and outliers in Figure 3 (right), does not change drastically, which could mean that the models maintain a similar level of consistency in orientation prediction despite task complexity. In the following sections we extend our evaluation to focus on the impact of vision conditions and task complexity using Table 2.

4.2.1 *Impact of vision condition.*

Comparing the NV-ST model's performance across tests sets reveals a significant shift when going from normal to low vision: position MSE increases by 89.16% ($+0.074\,m^2$) and orientation OD by 7.18%. Training on diverse vision (DV-ST) introduces a 8.2% ($-0.013\,m^2$) decrease in position MSE over the base model (NV-ST), but an increase (also 8.2%) in orientation OD. Meanwhile, The MSE and OD for the NV-ST test condition also increase by 25.30% ($+0.021\,m^2$) and 10.95% respectively, further reinforcing the notion of a trade-off in model performance.

Answer to RQ2: The Model NV-ST trained on normal vision and simple tasks exhibit poor robustness when predicting on low vision with increase in position (MSE) and orientation (OD) error. Modifying the training set to include diverse vision conditions (model DV-ST) improves the position error but worsens the orientation error, and even more increases the position and orientation error for the simple task test NV-ST. While the model becomes more adaptable to low-vision trajectories, its performance slightly degrades in normal-vision conditions.

4.2.2 *Impact of task complexity.*

Comparing the NV-ST model performance on various test sets, we notice a 898.80% ($+0.746\,m^2$) increase in position MSE and 44.34% increase in orientation OD when going from simple to complex tasks. The substantial increase in both MSE and OD under the complex task condition reflects that task complexity has a more profound impact on the model's performance than changes in vision conditions.

In contrast, the model trained on diverse tasks (NV-DT) outperforms the NV-ST model on complex tasks, with 79.98% ($-0.662\,m^2$) decrease in position MSE, and 14.43% decrease in OD. However, this is at the expense of accuracy for NV-ST test, with an 72.29% ($+0.06\,m^2$) increase for position MSE. The orientation OD is less impacted, only resulting in a 0.9% increase. This reflects that training on diverse tasks (NV-DT) improves the model's ability to tackle complex challenges, enhancing both positional accuracy and orientation precision. However, this focused improvement on complex tasks can potentially lead to a reduction in performance on simpler, baseline tasks (NV-ST).

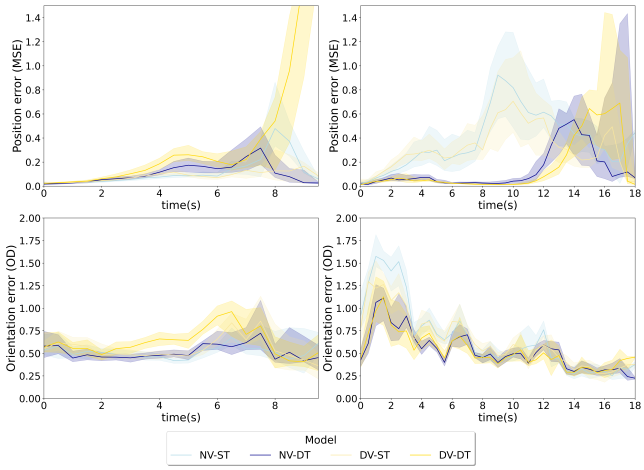


Figure 4: Disaggregated results for MSE and OD along the task duration, under NV-ST (left) and LV-CT (right) test conditions.

Finally, if we evaluate the performance of the DV-DT model against the baseline model NV-ST across the four different test conditions, we observe that the DV-DT model shows a remarkable improvement in handling position prediction in complex tasks and scenarios involving low vision (-83.53% on position MSE $-0.695\ m^2$), with also a moderate improvement to orientation OD (-9.38%). However, when evaluated under low vision conditions with simple tasks (LV-ST test), the model's performance slightly deteriorates. Significant concern arises from the model's performance in standard test conditions (NV-ST test), where the MSE position error increases by 168.67% ($+0.14\ m^2$), with also a notable increase in OD orientation error (+16.52%).

Answer to RQ3: The Model NV-ST trained on normal vision and simple tasks exhibit poor robustness when predicting on complex tasks. Modifying the training set to include diverse tasks, generally improves model performance in those specific conditions. However, this focused improvement comes with compromises, on the baseline tasks (NV-ST). The quantified data reveal that training on a broad spectrum of conditions significantly improves performance, evidenced by up to an 89% reduction in positional error and 20% in orientation error, but introduces a trade-off, resulting in reduced accuracy for simpler baseline tasks.

The disaggregated plots in Figure 4 shows MSE and OD across models under NV-ST (right) and LV-CT (left) test. The model NV-ST demonstrates impressive accuracy, contrasting with the model DV-DT, where a pronounced increase in MSE is observed towards the task's end. In the LV-CT test, models NV-ST and DV-ST exhibit heightened MSE at the task's onset due to their training void of complex tasks, whereas NV-DT and DV-DT models show initial MSE reductions, only to rise again as tasks progress. This trend is paralleled by escalating OD errors from the outset, particularly when individuals engage with traffic lights, highlighting increased positional uncertainty. The influence of low vision introduces amplified uncertainty in tracking ground truth positions and orientations, notably exacerbating as tasks conclude. Moreover, the onset of complex tasks elevates orientation errors, especially during initial traffic light interactions, leading to escalated positional errors by the task's end.

5 DISCUSSIONS

The experimental evaluation described in Section 4 details the impact of low-vision and task complexity conditions on human motion prediction.

Our findings reveal GIMO as the superior model, taking advantage of the extra contextual information in this model, consistently outperforming MLP and TRACK in prediction for both position and orientation, especially in the NV-ST and DV-DT tests. However, performing a deeper analysis on the models trained with GIMO, reveals issues in the NV-ST model's ability to predict tasks with low vision and complex task conditions. And even if we refine the training data set to include diversity of conditions (NV-DT, DV-ST and DV-DT) marginally improved position and orientation prediction errors, but at the cost of increased prediction errors in tasks with normal conditions.

Our study specifically addresses the challenges faced by individuals with scotoma, a condition resulting in partial vision loss, in the context of human motion forecasting but also highlights the urgent need for a more inclusive approach in subsequent research efforts.

6 CONCLUSIONS

Our study provides a foundational understanding of model performance in predicting human motion in immersive environments under low visual conditions and complex tasks. We found that models trained with a diverse range of task conditions stands out for its robustness in reducing position and orientation prediction errors by 84% and 9%, respectively. Nonetheless, the subtle trade-offs observed across normal vision and simple task conditions highlight the complexity of designing predictive models. Future work could explore more training strategies or architectural improvements to enhance performance under these challenging conditions, a promising direction involves using context annotations from the CREATIVE3D dataset to deepen our understanding of human behavior during task execution, particularly in complex scenarios.

7 ACKNOWLEDGEMENTS

This work has been partially supported by the French National Research Agency through the ANR CREATTIVE3D project ANR-21-CE33-0001 and UCA^{JEDI} Investissements d'Avenir ANR-15-IDEX-01 (IDEX reference center for extended reality XR^{2}C^2).

REFERENCES

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. 2020. Socially and Contextually Aware Human Motion and Pose Forecasting. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6033–6040. https://doi.org/10.1109/LRA.2020.3010742 Number: 4.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 565–574.

[3] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. 2023. CIRCLE: Capture In Rich Contextual Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21211–21221.

[4] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. 2009. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. (2009).

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *Int. J. of Robotics Research* 32, 11 (2013), 1231–37.

[6] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2023. Back to mlp: A simple baseline for human motion prediction. In *Proc. IEEE Winter Conf. on Appl. of Computer Vision*. 4809–19.

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

[8] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*.

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[10] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proc. IEEE/CVF international conference on computer vision*. 5442–5451.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[12] Florent Robert, Hui-Yin Wu, Lucile Sassatelli, Stephen Ramanoël, Auriane Gros, and Marco Winckler. 2023. An Integrated Framework for Understanding Multimodal Embodied Experiences in Interactive Virtual Reality. In *Proc. 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France). Association for Computing Machinery, New York, NY, USA, 14–26.

[13] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. 2021. TRACK: A New Method From a Re-Examination of Deep Architectures for Head Motion Prediction in 360^o Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 5681–5699.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Hui-Yin Wu, Florent Alain Sauveur Robert, Franz Franco Gallo, Kateryna Pirkovets, Clément Quere, Johanna Delachambre, Stephen Ramanoël, Auriane Gros, Marco Winckler, Lucile Sassatelli, Meggy Hayotte, Aline Menin, and Pierre Kornprobst. 2023. Exploring, walking, and interacting in virtual reality with simulated low vision: a living contextual dataset. (2023). https://inria.hal.science/hal-04429351 preprint.

[16] Chi Zhang and Christian Berger. 2023. Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Transactions on Intelligent Transportation Systems* 24, 10 (2023), 10279–10301. https://doi.org/10.1109/TITS.2023.3281393

[17] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J Guibas. 2022. GIMO: Gaze-Informed Human Motion Prediction in Context. *arXiv preprint arXiv:2204.09443* (2022). arXiv:2204.09443

Received 07 February 2024; revised 08 March 2024; accepted xx March 2024

Quality Assessment and Modeling for MPEG V-PCC Volumetric Video

Yuang Shi
National University of Singapore
Singapore
yuangshi@u.nus.edu

Samuel Rhys Cox
National University of Singapore
Singapore
samcox@comp.nus.edu.sg

Wei Tsang Ooi
National University of Singapore
Singapore
ooiwt@comp.nus.edu.sg

ABSTRACT

Volumetric video, which is typically represented by 3D point clouds, requires efficient point cloud compression (PCC) technologies for practical storage and transmission. Particularly, developed by the Moving Picture Experts Group (MPEG), video-based PCC (V-PCC) converts 3D point clouds into 2D image maps and compresses them with 2D video codecs, showing excellent compression performance. However, understanding the impact of compression on the perceptual quality of volumetric videos, which consist of both geometry and texture components, remains challenging. In this study, we propose a quality of experience (QoE) model to predict the subjective quality with respect to the compression level of geometry and texture, quantifying the impact of geometry and texture compression on perceptual quality. To the best of our knowledge, this study is the first to accurately model the perceptual quality of V-PCC-encoded volumetric videos. The QoE model is built based on a volumetric video quality assessment dataset, VOLVQAD, collected by us. We further evaluate our QoE model on the vsenseVVDB2 dataset, which was collected from diverse study settings, to validate its robustness and generalization ability. Both evaluations demonstrate the effectiveness of our model in various compression scenarios. This study makes a valuable contribution to our understanding of the factors that influence the QoE in V-PCC-encoded volumetric videos. The proposed model also holds potential for various other applications, such as adaptive bitrate allocation.

CCS CONCEPTS

• **Computing methodologies** → **Volumetric models**; • **Information systems** → **Multimedia streaming**.

KEYWORDS

Volumetric video; MPEG V-PCC; Subjective quality evaluation; Subjective quality modeling

ACM Reference Format:
Yuang Shi, Samuel Rhys Cox, and Wei Tsang Ooi. 2024. Quality Assessment and Modeling for MPEG V-PCC Volumetric Video. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652218

1 INTRODUCTION

Volumetric video is a promising media format used in virtual/augmented reality systems. The 3D point cloud is a popular way to represent volumetric video. To enable efficient point cloud compression (PCC), the Moving Picture Experts Group (MPEG) standardized two compression technologies: video-based PCC (V-PCC) and geometry-based PCC (G-PCC) [11]. In particular, V-PCC uses a projection-based approach to convert 3D point clouds into multiple 2D image maps that represent the texture, geometry, and occupancy of the points. These image sequences are then compressed as a 2D video using state-of-the-art video codecs. V-PCC shows significant potential for the applications of volumetric videos.

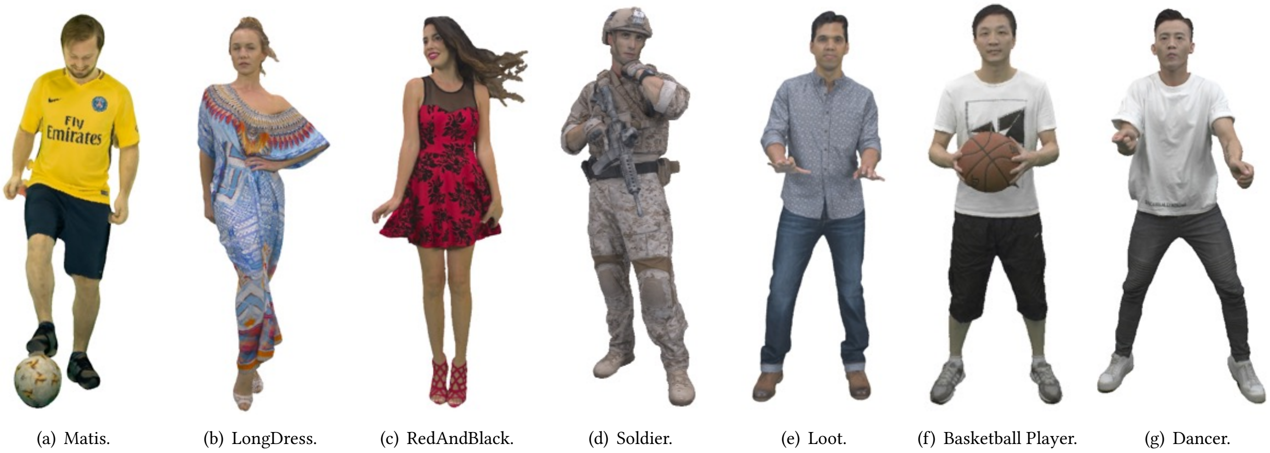
Unlike 2D videos, volumetric videos consist of both geometry (the shape and structure) and texture (the color and visual appearance). When these components are compressed, they may exhibit different types of visual imperfections or artifacts that can impact the overall perceptual quality of the video. Understanding the specific effects of compression on geometry and texture is crucial for optimizing the perceptual quality of volumetric videos. By analyzing these effects, researchers can develop techniques to minimize artifacts and enhance the viewing experience for users.

There are limited works available in the literature for evaluating the perceptual quality of volumetric video [5, 10, 25, 29–31], with only one study considering the individual impact of geometry and texture compression. Among these studies, Cox et al. [3] took the initial step by investigating the effect of compression levels for texture and geometry maps on the perceptual quality of V-PCC-encoded volumetric videos. They conducted a subjective quality assessment study, creating a volumetric video quality assessment (VVQA) dataset called VOLVQAD. The volumetric videos from 8i dataset [6] were encoded using MPEG V-PCC with sixteen different compression levels for texture and geometry. Participants were asked to provide mean opinion scores (MOS) based on the rendered test videos. By analyzing the subjective assessment data, they made a qualitative observation that compressing the texture map resulted in a more significant reduction in perceptual quality compared to compressing the geometry map. Nevertheless, they failed to model the relationship between compression artifacts and perceptual quality.

Building upon this previous work, we take a further step by developing a statistical model to quantify the impact of geometry and texture compression on perceptual quality. We trained a machine-learning model using the VOLVQAD dataset, which was generated from 8i dataset [6]. To evaluate our model, we conducted additional subjective quality assessment experiments with 36 users

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652218



(a) Matis. (b) LongDress. (c) RedAndBlack. (d) Soldier. (e) Loot. (f) Basketball Player. (g) Dancer.

Figure 1: Volumetric videos for user study selected from (a) vsenseVVDB1 [29], (b-e) 8i dataset [6], and (f-g) Owlii dataset [28].

and created a new VVQA dataset [1] that consists of 864 ratings of V-PCC-encoded videos with different compression levels of geometry and texture. The raw volumetric videos of this evaluation dataset were taken from not only the 8i dataset but also Owlii dataset [28]. Our model yields great performance with an overall Pearson correlation coefficient (PCC) of 0.98, Spearman rank correlation coefficient (SRCC) of 0.93, root mean square error (RMSE) of 0.50, and mean absolute error (MAE) of 0.45. Additionally, we included the vsenseVVDB2 VVQA dataset [30], which was collected from different user study settings. This additional VVQA dataset allows for a more comprehensive assessment of our model across diverse scenarios. Our model shows outstanding performance with PCC of 0.99, SRCC of 1.00, RMSE of 0.09, and MAE of 0.07. These results of our evaluation demonstrate the effectiveness and generalization ability of our model, highlighting its capability to accurately predict perceptual quality in varying compression scenarios.

The paper is structured as follows: Section 2 provides an overview of related datasets and highlights the distinctions between our datasets and existing ones. In Section 3, we explain the process of generating volumetric video sequences and conducting the subjective quality assessment study. The training and evaluation of our proposed QoE model are discussed in Section 4, followed by the conclusion in Section 5.

2 RELATED WORK

The establishment of perceptual quality prediction models for 2D videos has been extensively investigated in the literature. Numerous studies on video quality assessment [8, 9, 15, 16] have revealed that video quality is influenced by encoder-related parameters, such as quantization factor. For instance, Eden proposed a video quality prediction model that highlights the quantization parameter of the encoder as a primary factor impacting QoE [9]. Khan et al. developed a QoE prediction model by considering the distortions caused by the encoder [16].

As video capturing and processing techniques advance, the exploration of perceptual quality assessment and modeling has extended to the realm of 3D space. This research can be broadly categorized into two groups based on the data format used to construct the 3D models: perceptual quality modeling of point cloud [1, 2, 4, 12, 17, 23, 24] and mesh [19, 20]. Among the works focusing on point cloud QoE modeling, the majority predominantly investigate impairments introduced to single-frame 3D models. For instance, Alexiou et al. conducted subjective studies to explore the influence of V-PCC and G-PCC-induced distortions on QoE for static point cloud models [2]. However, volumetric videos possess a dynamic nature, which introduces additional factors affecting QoE, including motion smoothness, temporal consistency, and the perception of movement within the scene.

Although a few recent studies have addressed volumetric video quality assessment and modeling [3, 10, 21, 25, 27, 29, 30], the individual impact of geometry and texture compression of V-PCC has not been adequately considered in most of them. Among these works, Cox et al. [3] explored the individual roles of geometry and texture compression and developed the VVQA dataset called VOLVQAD. Our work serves as a follow-up study to VOLVQAD, focusing on the development of a statistical model to quantify the impact of geometry and texture compression on perceptual quality. In addition, we conducted further subjective quality assessment experiments involving 36 users, resulting in the creation of a new VVQA dataset comprising 864 ratings of V-PCC-encoded videos with varying levels of geometry and texture compression.

3 SUBJECTIVE QUALITY ASSESSMENT

3.1 Stimuli Generation

Volumetric Video Dataset. We utilize a set of seven dynamic point clouds as our raw data, which are depicted in Figure 1. These point clouds are sourced from various datasets, including the vsenseVVDB1 [29], the 8i dataset [6], and the Owlii dataset [28]. The first volumetric video, named Matis, originates from the vsenseVVDB1

[1] This dataset is publicly available at https://github.com/nus-vv-streams/qoe-model for sharing with the research community.

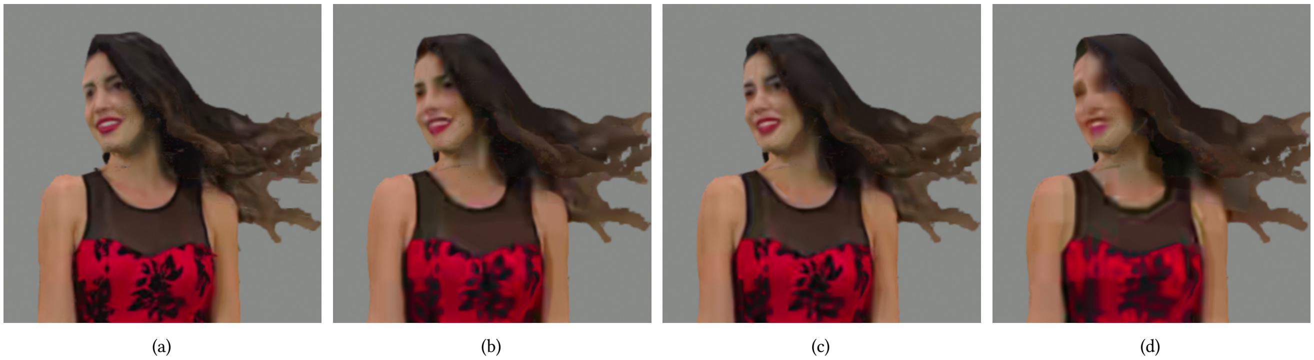


Figure 2: Sample frames of the RedAndBlack model showing quality levels: (a) (GR0, TR3), (b) (GR3.5, TR1), (c) (GR4, TR2.5), (d) (GR5, TR0).

and is employed for the training task in our user study. The remaining four point cloud sequences (LongDress, RedAndBlack, Soldier, and Loot) are extracted from the 8i dataset. We additionally include two dynamic point clouds, Basketball Player and Dancer, which have been obtained from the Owlii dataset. It is worth noting that the Basketball Player and Dancer sequences are twice as large as the avatars in the 8i dataset. The size of the point cloud can have an impact on compression and rendering processes. To ensure consistency in point density, we address this by down-scaling and down-sampling the Owlii dataset. This adjustment allows us to maintain the same point density as the 8i dataset, ensuring comparable results in compression and rendering.

Compression. V-PCC projects the volumetric video into 2D geometry and texture maps and compresses them separately. The overall compression rate of V-PCC is thus determined jointly by the geometry compression rate (GR) and texture compression rate (TR) with each having its quantization parameter (QP): geometry QP (QP^g) and texture QP (QP^t). Based on the V-PCC Common Test Condition (CTC) [7], the compression rates are defined into five levels, labeled as R1 to R5. Additionally, to explore the rate-distortion performance comprehensively, we further increase the QP values for R1 to obtain an additional compression rate called R0, as demonstrated in previous works [3, 22]. Following this notation, we represent these compression rates by GR and TR so that Ri can be denoted as (GRi, TRi), where $i \in \{0, 1, 2, 3, 4, 5\}$.

Cox et al. [3] made the first work to explore the impact of geometry and texture compression on QoE. They built a VVQA dataset called VOLVQAD which was collected from a user study. In the user study, two sets of volumetric videos that encompassed a total of sixteen quality levels were generated for rating. The first set of videos was generated with compression rates (GRi, TRi), $i \in \{0, 1, 2, 3, 4, 5\}$. The second set of videos were generated by varying the GR across the quality levels while maintaining a constant TR of TR5, that is, (GRi, TR5), $i \in \{0, 1, 2, 3, 4\}$. They also generate the videos with compression rates (GR2, TRi), $i \in \{0, 1, 3, 4, 5\}$, which means varying the TR across the quality levels (TR0 to TR5) while keeping the GR constant at GR2.

As our QoE prediction model is built on the VOLVQAD dataset, we made a careful selection of the compression rates for building our testing dataset, so that we can assess the performance of our model on unseen data with different compression settings. We specifically choose four settings of GR and TR: (GR0, TR3), (GR3.5, TR1), (GR4, TR2.5), and (GR5, TR0). These settings allow us to generate four distinct sets of videos with different compression configurations than those in VOLVQAD. Besides, the settings of (GR3.5, TR1) and (GR4, TR2.5) change the GR and TR in finer granularity respectively, which can help in understanding the impact of small variations in GR and TR on the perceptual quality of volumetric videos. The MPEG V-PCC reference software (v15.0) [2] is used to encode the raw point cloud sequences. Table 1 summarizes our encoder settings.

Table 1: Settings of V-PCC encoder for the testing set

	(GR0, TR3)	(GR3.5, TR1)	(GR4, TR2.5)	(GR5, TR0)
QP^g	36.0	22.0	20.0	16.0
QP^t	32.0	42.0	34.5	47.0

Rendering and Video Generation. We follow the guidelines presented by Cox et al. [3] for rendering and video generation. Specifically, we decode the compressed V-PCC streams and use the Open3D Python library (v0.14.1) to generate images for each frame in the point cloud sequences. The image size is fixed at 600×1080, and we set the camera viewport to a frontal view with the object positioned at the center. The background color is gray (#898B88) with a point size of 1, and these settings remain consistent for all frames of the same model. For video generation, we utilize FFmpeg [3] to create videos at 30fps with a duration of 10 seconds. We apply visually lossless H.264 parameters (-c:v libx264 -crf 15) to control distortion while maintaining high quality. Figure 2 shows the rendered frames of RedAndBlack with four quality levels.

3.2 Participants

We conducted participant recruitment through a university advertisement web page, where potential participants were invited to

[2] https://github.com/MPEGGroup/mpeg-pcc-tmc2/releases/tag/release-v15.0
[3] https://ffmpeg.org

take part in our study. The recruitment criteria specified that participants needed to be at least 18 years old, have no (uncorrected) visual impairments or color blindness and possess no prior experience in picture quality evaluation. We offered a reimbursement of S$6 upon completion of the user study, which typically took around 10 to 15 minutes to finish. Before proceeding with the evaluation of the videos, participants underwent in-person vision tests, which were conducted by our research team. Participants who did not pass the vision tests were unable to continue with the user study. In total, we recruited 36 participants for the study, with an average age of 21.5 (ranging from 19 to 24 years old). Out of the participants, 20 identified as female, while 16 identified as male.

3.3 Procedure

The user studies were conducted following the recommendations of the International Telecommunication Union (ITU) [13, 14]. Participants were placed in a dimly lit room and seated at a fixed viewing distance of 120 cm (four times the display height) away from the display (Dell P2319H). At the start of the user study, the study's workflow and goals were explained to the participants. The participants were also asked to provide their consent to participate. Following the ITU guidelines, participants completed a visual acuity test (using a Snellen eye chart) and a test for normal color vision (using Beck color plates).

Participants who successfully passed the vision tests received detailed instructions regarding the tasks they would perform. They then went through training to familiarize themselves with the interface and experimental procedures. During the training phase, participants completed five video rating tasks, following the same procedure as the main study. The model used in training is not used in the main study. The videos shown during training covered a range of low- and high-quality to ensure participants were familiar with a full spectrum of quality impairments before the start of the study. A member of the research team was present in the room throughout the user studies to address any questions that participants may have had.

Upon completing the training, participants proceeded to the main rating tasks. In this phase, participants viewed two videos, both from the same model, displayed side-by-side on the screen for ten seconds. The reference video, which represented unimpaired quality, appeared on the left; while the trial video, potentially containing impairments, appeared on the right. Participants were then instructed to rate the quality impairment of the trial video using the degradation category rating (DCR) method. After the videos finished playing, participants were asked to rate the quality impairment of the trial video relative to the reference video using a scale consisting of the following options: "1 - Very annoying," "2 - Annoying," "3 - Slightly annoying," "4 - Perceptible but not annoying," and "5 - Imperceptible."

During the rating process, participants had the option to replay videos as many times as they deemed necessary, and no time limit was imposed on providing video ratings. Participants were presented with 24 pairs of videos, and the order of video presentation was randomized for each participant.

4 SUBJECTIVE QUALITY MODELING

4.1 Experimental Settings

QoE Training Set. VOLVQAD is used as the training set to train our QoE prediction model. VOLVQAD comprises 376 video sequences and 7,680 ratings collected from 120 users. These video sequences are encoded with MPEG V-PCC using four avatar models from the 8i dataset, covering 16 different quality levels controlled by QP^g and QP^t. The subjective quality assessment methodology employed in VOLVQAD is consistent with our study.

QoE Testing Set. In addition to the testing dataset collected by ourselves, as mentioned in Section 3, we also incorporate vsenseVVDB2 [30] as an additional testing set. VsenseVVDB2 consists of ratings for eight volumetric videos sourced from vsenseVVDB1 [29] and 8i dataset. These videos were encoded using V-PCC with compression rates ranging from R0 to R5. Because the ratings range from 0 to 100, we perform a discretization process to map these ratings into the range [1, 2, 3, 4, 5]. Specifically, we divided the rating range into five equal intervals of 20 units each. Each interval was then assigned a discrete value, starting from 1. Finally, we mapped each rating to its corresponding discrete value based on the interval it fell into.

It is worth noting that the user study procedure for vsenseVVDB2 differs from ours. In their study, the videos were placed within a scene, and the camera was set to orbit the avatar's initial origin twice in a clockwise direction within a 10-second interval. The inclusion of this additional testing data collected from different study settings allows us to evaluate the generalization ability of our model, assessing its performance beyond the specific conditions of our dataset.

Evaluation Metrics. Four metrics are used to evaluate the QoE model: PCC, SRCC, RMSE, and MAE. The PCC and SRCC provide insights into the correlation and ranking accuracy of the predictions of the model. Meanwhile, RMSE considers the average magnitude of these differences, and MAE focuses on their average absolute value.

4.2 Model Training and Selection

To achieve accurate quality prediction, we utilize supervised machine learning (ML) algorithms. We consider five ML models: (i) Polynomial Regression (PR), (ii) Support Vector Regression (SVR), (iii) Random Forest (RF), (iv) Multi-Layer Perceptron (MLP), and (v) K-Nearest Neighbors (K-NN). During the training stage, we conduct 5-fold cross-validation to tune the hyperparameters using the grid search algorithm. Our primary objective criterion for hyperparameter tuning is RMSE.

Following the hyperparameter tuning process, we carefully select specific hyperparameter values for each ML model. The PR model is configured with two degrees, while the SVR model utilizes a linear kernel. The RF model incorporates 100 estimators and a maximum depth of 5. The MLP model employs a learning rate of 0.001 and a hidden layer consisting of 100 neurons. Lastly, the K-NN model is set to consider two neighbors with uniform weights. All other hyperparameters are maintained at their default values.

The prediction performance of the five models is reported in Table 2. We find that PR achieves the best performance with 0.99 PCC, 1.00 SRCC, 0.06 RMSE, and 0.20 MAE, compared to other methods.

We thus adopt PR to predict the subjective quality of volumetric videos with respect to QP^g and QP^t, given its effectiveness and efficiency.

Formally, for a volumetric video encoded by V-PCC with QP^g and QP^t, we can predict its subjective quality $\hat{q}$ with:

$$\hat{q} = \vec{X}^T \cdot \vec{\beta} + \varepsilon, \tag{1}$$

where

$$\begin{aligned} \vec{X}^T &= \left[QP^g, QP^t, \left(QP^g\right)^2, QP^g \cdot QP^t, \left(QP^t\right)^2\right], \\ \vec{\beta} &= [-0.002, 0.208, -0.005, 0.006, -0.007]^T, \quad \varepsilon = 2.29. \end{aligned} \tag{2}$$

Table 2: Comparison of the QoE models

Model	PCC ↑	SRCC ↑	RMSE ↓	MAE ↓
SVR	0.16	0.10	0.53	0.47
MLP	0.45	0.30	0.92	0.77
KNN	0.81	0.75	0.40	0.34
RF	0.96	0.90	0.49	0.41
PR	**0.99**	**1.00**	**0.06**	**0.20**

4.3 Model Evaluation

We first evaluate the performance of our QoE prediction model on the testing dataset collected by ourselves, which consists of 864 ratings of V-PCC-encoded videos from 8i dataset and the Owlii dataset. Figure 3 plots the predicted MOS of our model. Table 3 presents the performance of the QoE model on the encoded videos from the 8i and Owlii datasets.

Table 3: Performance of QoE model on 8i and Owlii datasets

Dataset	PCC ↑	SRCC ↑	RMSE ↓	MAE ↓
8i	0.99	1.00	0.43	0.42
Owlii	0.98	0.80	0.55	0.48
Overall	0.98	0.93	0.50	0.45

Recall that the training set is collected based on the encoded videos from the 8i dataset, which serves as the foundation for training our model. Therefore, when predicting the QoE of encoded video from 8i dataset, our model shows exceptional performance results, with a high PCC of 0.99, a perfect SRCC of 1.00, a low RMSE of 0.43, and a small MAE of 0.42. On the other hand, the Owlii dataset is not part of the training process, thus containing new and unseen features that the model may not learn from the 8i dataset. Nevertheless, the QoE model exhibited a high PCC of 0.98 and a good SRCC of 0.80, indicating a strong positive linear relationship and a reasonable monotonic relationship between the predicted and actual ratings. The RMSE of 0.55 and MAE of 0.48 indicate slightly larger average differences between the predicted and actual ratings compared to the 8i dataset. Despite these small differences, the performance of our model on the Owlii dataset highlights its ability to generalize well to new and unseen data, demonstrating its robustness and generalization ability.

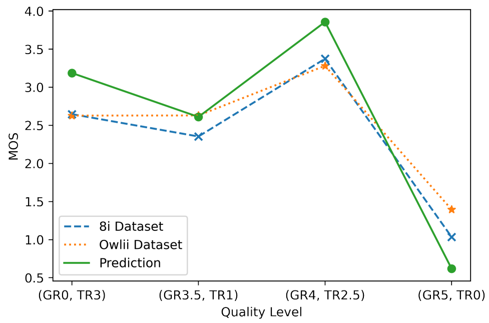


Figure 3: Prediction of QoE model on 8i and Owlii datasets.

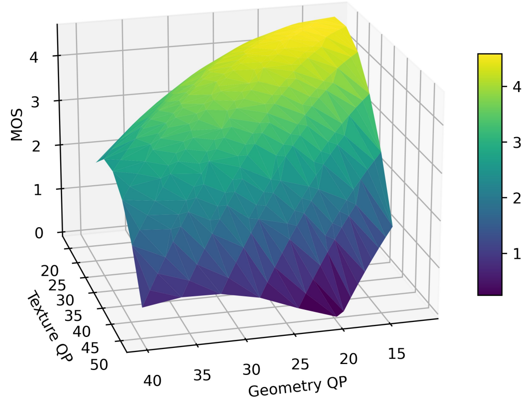


Figure 4: The QoE surface fitted by the predictions.

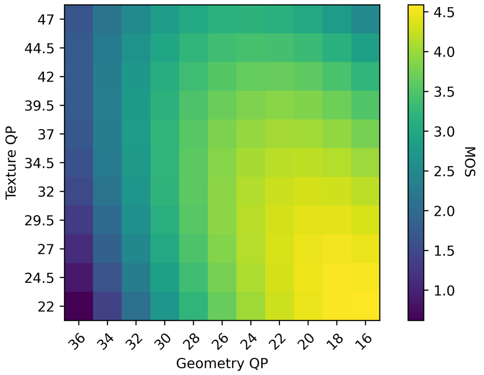


Figure 5: The QoE matrix of the predictions.

Moreover, using the predictions generated by our model, we create a visualization of the results by fitting a surface of QoE in a 3D space, as shown in Figure 4. We also plot the QoE matrix constructed from the prediction of our model in Figure 5. Both visualizations allow us to gain insights into how the QoE varies based on different combinations of geometry QP and texture QP

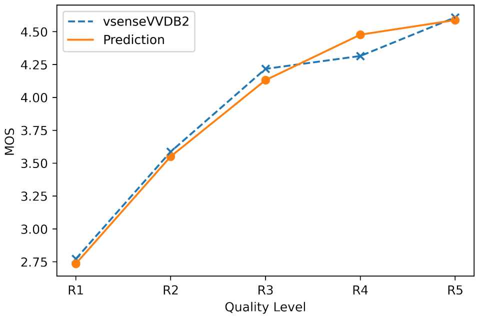


Figure 6: Prediction of QoE model on vsenseVVDB2.

values. It is generally observed that the MOS tends to increase as the geometry QP and texture QP values decrease. However, an interesting finding is that when the texture QP is high, the MOS remains consistently low regardless of the change in geometry QP. On the contrary, when the texture QP is low, the MOS increases as the geometry QP decreases. This observation suggests that users are more sensitive to changes in texture quality compared to geometry quality, which aligns with previous research findings [3, 18]. It implies that variations in texture QP have a more pronounced impact on the perceived quality of the content.

Table 4: Performance of QoE model on vsenseVVDB2

Dataset	PCC ↑	SRCC ↑	RMSE ↓	MAE ↓
vsenseVVDB2	0.99	1.00	0.09	0.07

Besides, we evaluate our model on vsenseVVDB2. We only selected the ratings of volumetric videos from the 8i dataset. This is because the other avatar models from vsenseVVDB1 have sparser point clouds, resulting in fewer data points to represent the details of texture and geometry. Therefore, the generated videos from these models tend to have lower quality compared to the volumetric videos from the 8i dataset. We plot the predictions and actual ratings from vsenseVVDB2 in Figure 6 and present the performance of our model in Table 4. As can be seen, our QoE model demonstrates excellent performance on the vsenseVVDB2 dataset, with PCC of 0.99, SRCC of 1.00, RMSE of 0.09, and MAE of 0.07. The results indicate that our model effectively generalizes its predictions beyond specific conditions of the training dataset, and accurately predicts the QoE for videos in different user study settings.

5 CONCLUSION

In this paper, we conduct a subjective quality assessment study and develop a QoE model that predicts the quality of experience by considering the features of V-PCC. The evaluation results demonstrate the effectiveness and generalization ability of our model. This work contributes to the understanding of factors influencing QoE in V-PCC-encoded volumetric videos and provides a valuable tool for video encoding and delivery optimizations to enhance user satisfaction. However, the proposed QoE model focuses primarily on geometry QP and texture QP of V-PCC. While these factors have been shown to influence QoE, there may be other important aspects, such as lighting, color accuracy, or motion smoothness, that were not explicitly considered in the model. Future research could explore incorporating additional relevant features to further improve the accuracy and completeness of the QoE model. Moreover, different visualization modalities, such as using Head-Mounted Displays (HMDs), can introduce unique factors that may influence users' perception, immersion, and overall satisfaction [26]. Examining the performance of the proposed prediction model with data from experiments involving different visualization modalities is crucial for understanding the impact of these modalities on the quality of experience.

ACKNOWLEDGMENTS

This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (T1 251RES2038).

REFERENCES

[1] Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, Giuseppe Valenzise, and Patrick Le Callet. BASICS: Broad quality assessment of static point clouds in a compression scenario. *IEEE Transactions on Multimedia*, pages 1–13, Early Access, 2024.

[2] Evangelos Alexiou, Irene Viola, Tomás M Borges, Tiago A Fonseca, Ricardo L De Queiroz, and Touradj Ebrahimi. A comprehensive study of the rate-distortion performance in mpeg point cloud compression. *APSIPA Transactions on Signal and Information Processing*, 8:e27, Nov 2019.

[3] Samuel Rhys Cox, May Lim, and Wei Tsang Ooi. VOLVQAD: an MPEG V-PCC volumetric video quality assessment dataset. In *Proceedings of the 14th Conference on ACM Multimedia Systems, MMSys 2023, Vancouver, BC, Canada, June 7-10, 2023*, pages 357–362. ACM, 2023.

[4] Luís Alberto da Silva Cruz, Emil Dumic, Evangelos Alexiou, João Prazeres, Carlos Rafael Duarte, Manuela Pereira, António M. G. Pinheiro, and Touradj Ebrahimi. Point cloud quality evaluation: Towards a definition for test conditions. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience, QoMEX 2019, Berlin, Germany, June 5-7, 2019*, pages 1–6. IEEE, 2019.

[5] Sam Van Damme, Maria Torres Vega, and Filip De Turck. A full- and no-reference metrics accuracy analysis for volumetric media streaming. In *Proceedings of the 13th International Conference on Quality of Multimedia Experience, QoMEX 2021, Montreal, Canada, June 14-17, 2021*, pages 225–230. IEEE, 2021.

[6] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 8i voxelized full bodies-a voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006*, 7(8):11, 2017.

[7] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. Common test conditions for point cloud compression. *ISO/IEC JTC1/SC29/WG11 w17766*, 2018.

[8] Arnd Eden. No-reference image quality analysis for compressed video sequences. *IEEE Transactions on Broadcasting*, 54(3):691–697, Sept 2008.

[9] Rosario Feghali, Filippo Speranza, Demin Wang, and Andr Vincent. Video quality metric for bit rate control via joint adjustment of quantization and frame rate. *IEEE Transactions on Broadcasting*, 53(1):441–446, Mar 2007.

[10] Mateus M. Gonçalves, Luciano Agostini, Daniel Palomino, Marcelo Schiavon Porto, and Guilherme Corrêa. Encoding efficiency and computational cost assessment of state-of-the-art point cloud codecs. In *Proceedings of the 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 3726–3730. IEEE, 2019.

[11] Danillo Graziosi, Ohji Nakagami, Satoru Kuma, Alexandre Zaghetto, Teruhiko Suzuki, and Ali Tabatabai. An overview of ongoing point cloud compression standardization activities: Video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing*, 9:e13, Apr 2020.

[12] Zhouyan He, Gangyi Jiang, Mei Yu, Zhidi Jiang, Zongju Peng, and Fen Chen. TGP-PCQA: Texture and geometry projection based quality assessment for colored point clouds. *Journal of Visual Communication and Image Representation*, 83:103449, Jan 2022.

[13] ITU-R. Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation BT.500-13*, 2012.

[14] ITU-T. Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation*, page 910, 2008.

[15] Shraboni Jana, An (Jack) Chan, Amit Pande, and Prasant Mohapatra. QoE prediction model for mobile video telephony. *Multimedia Tools and Applications*, 75(13):7957–7980, Jun 2015.

[16] Asiya Khan, Lingfen Sun, and Emmanuel C. Ifeachor. QoE prediction model and its application in video quality adaptation over umts networks. *IEEE Transactions on Multimedia*, 14(2):431–442, Apr 2012.

[17] Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang. Point cloud quality assessment: Dataset construction and learning-based no-reference metric. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):80:1–80:26, Feb 2023.

[18] A Mike Burton. Why has research in face recognition progressed so slowly? the importance of variability. *Quarterly Journal of Experimental Psychology*, 66(8):1467–1485, Aug 2013.

[19] Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric. *ACM Transactions on Graphics*, 42(3), Jun 2023.

[20] Yana Nehmé, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Visual quality of 3D meshes with diffuse colors in virtual reality: Subjective and objective evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2202–2219, Nov 2021.

[21] Minh Nguyen, Shivi Vats, Sam Van Damme, Jeroen van der Hooft, Maria Torres Vega, Tim Wauters, Christian Timmerer, and Hermann Hellwagner. Impact of quality and distance on the perception of point clouds in mixed reality. In *Proceedings of the 15th International Conference on Quality of Multimedia Experience, QoMEX 2023, Ghent, Belgium, June 20-22, 2023*, pages 87–90. IEEE, 2023.

[22] Yuang Shi, Pranav Venkatram, Yifan Ding, and Wei Tsang Ooi. Enabling low bitrate MPEG V-PCC-encoded volumetric video streaming with 3D sub-sampling. In *Proceedings of the 14th Conference on ACM Multimedia Systems, MMSys 2023, Vancouver, BC, Canada, June 7-10, 2023*, pages 108–118. ACM, 2023.

[23] Honglei Su, Zhengfang Duanmu, Wentao Liu, Qi Liu, and Zhou Wang. Perceptual quality assessment of 3d point clouds. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 3182–3186. IEEE, 2019.

[24] Eric M. Torlig, Evangelos Alexiou, Tiago A. Fonseca, Ricardo L. de Queiroz, and Touradj Ebrahimi. A novel methodology for quality assessment of voxelized point clouds. In *Applications of Digital Image Processing XLI*, volume 10752, pages 174–190. International Society for Optics and Photonics, SPIE, 2018.

[25] Jeroen van der Hooft, Maria Torres Vega, Christian Timmerer, Ali C. Begen, Filip De Turck, and Raimund Schatz. Objective and subjective qoe evaluation for adaptive point cloud streaming. In *Proceedings of the 12th International Conference on Quality of Multimedia Experience, QoMEX 2020, Athlone, Ireland, May 26-28, 2020*, pages 1–6. IEEE, 2020.

[26] Irene Viola, Shishir Subramanyam, Jie Li, and Pablo Cesar. On the impact of VR assessment on the quality of experience of highly realistic digital humans: A volumetric video case study. *Quality and User Experience*, 7(1):3, May 2022.

[27] Jannis Weil, Yassin Alkhalili, Anam Tahir, Thomas Gruczyk, Tobias Meuser, Mu Mu, Heinz Koeppl, and Andreas Mauthe. Modeling quality of experience for compressed point cloud sequences based on a subjective study. In *Proceedings of the 15th International Conference on Quality of Multimedia Experience, QoMEX 2023, Ghent, Belgium, June 20-22, 2023*, pages 135–140. IEEE, 2023.

[28] Yi Xu, Yao Lu, and Ziyu Wen. Owlii dynamic human mesh sequence dataset. *ISO/IEC JTC1/SC29/WG11 m41658*, 7(8):11, 2017.

[29] Emin Zerman, Pan Gao, Cagri Ozcinar, and Aljosa Smolic. Subjective and objective quality assessment for volumetric video compression. In *Image Quality and System Performance XVI, Electronic Imaging 2019, IQSP, Burlingame, CA, USA, 13-17 January 2019*. Society for Imaging Science and Technology, 2019.

[30] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. Textured mesh vs coloured point cloud: A subjective study for volumetric video compression. In *Proceedings of the 12th International Conference on Quality of Multimedia Experience, QoMEX 2020, Athlone, Ireland, May 26-28, 2020*, pages 1–6. IEEE, 2020.

[31] Xuemei Zhou, Irene Viola, Evangelos Alexiou, Jack Jansen, and Pablo César. QAVA-DPC: eye-tracking based quality assessment and visual attention dataset for dynamic point cloud in 6 DoF. In *Proceedings of the 2023 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2023, Sydney, Australia, October 16-20, 2023*, pages 69–78. IEEE, 2023.

Perceptual Impact of Facial Quality in MPEG V-PCC-encoded Volumetric Videos

Yuang Shi
National University of Singapore
Singapore
yuangshi@u.nus.edu

Wei Tsang Ooi
National University of Singapore
Singapore
ooiwt@comp.nus.edu.sg

ABSTRACT

Volumetric video, a technique used in augmented reality (AR) and virtual reality (VR) applications, presents unique challenges in rendering and compression. To enable efficient compression, video-based point cloud compression (V-PCC) techniques have been introduced by the Moving Picture Experts Group (MPEG). Given the interaction nature of volumetric videos, it is important to understand the impact of user behavior for the optimizations of volumetric video transmission and compression. In this study, we investigate the influence of rendering face quality of the avatars on users' viewing experience in MPEG V-PCC-encoded volumetric videos. We conducted a subjective quality assessment study using the Degradation Category Rating (DCR) method, manipulating facial quality by controlling the compression level of V-PCC. Our analysis reveals the significant role of facial quality in influencing users' overall perceptual quality in volumetric videos. The generated videos and subjective assessment data is made public at https://github.com/nus-vv-streams/facial-quality to support further research.

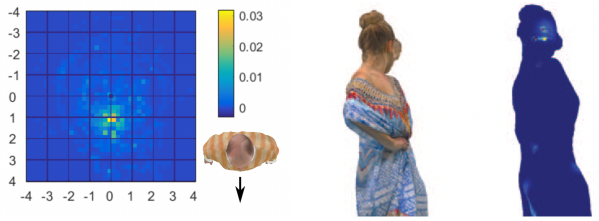
CCS CONCEPTS

• **Computing methodologies → Volumetric models**; • **Information systems → Multimedia streaming**.

KEYWORDS

Volumetric video; MPEG V-PCC; Subjective quality evaluation; Visual saliency

ACM Reference Format:
Yuang Shi and Wei Tsang Ooi. 2024. Perceptual Impact of Facial Quality in MPEG V-PCC-encoded Volumetric Videos. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3652212.3652221



(a)

(b)

Figure 1: The face of the avatar in volumetric videos usually attracts more attention: (a) Heat map of users' location [14], (b) Visual attention map of *Longdress* [15].

1 INTRODUCTION

Volumetric video is an emerging technique utilized for creating content in augmented reality (AR) and virtual reality (VR) applications. Typically, these videos are represented as either point clouds or 3D textured mesh sequences, offering viewers the freedom to observe the content from various angles. To facilitate efficient compression of point cloud data, the Moving Picture Experts Group (MPEG) has introduced video-based point cloud compression (V-PCC) techniques, which have been standardized and widely adopted [5]. V-PCC employs a projection-based approach to transform complex 3D point clouds into multiple 2D image maps. These image sequences are subsequently compressed using cutting-edge video codecs, treating them as conventional 2D videos. The implementation of V-PCC holds immense promise for enhancing the storage and transmission efficiency of volumetric videos.

Volumetric videos present distinct challenges compared to traditional 2D videos, necessitating a deeper understanding of the effects of rendering and compression on perceptual quality [1]. One notable difference is the freedom for users to move and view volumetric videos from various angles, introducing user behavior as a new factor in assessing the impact of rendering and compression on perceptual quality. A particularly complex aspect is visual attention, as different regions of the volumetric video naturally draw varying levels of viewer focus. Consequently, compression techniques may introduce distortions that affect the overall perceptual quality differently across different regions.

Numerous prior studies [7, 10–12, 14, 15] have consistently shown that users tend to direct their attention towards the frontal body of avatars in volumetric videos, with particular focus on the avatar's face. For instance, Zerman et al. [14] conducted an AR user study to collect behavior data in volumetric videos, revealing that participants allocated a substantial amount of their viewing time towards the face and frontal body of the volumetric avatars. Figure 1(a) shows an example of a heat map of users' location collected by them. Similarly, Zhou et al. [15] developed a comprehensive eye-tracking-based visual attention dataset. They observed that

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652221

subjects predominantly fixated on the faces and front view of dynamic point clouds, despite the random rotation of the avatar, as shown in Figure 1(b).

Building upon these findings, a subjective quality assessment study was conducted to investigate the influence of facial quality on users' viewing experience. In the user study, the Degradation Category Rating (DCR) method is used to evaluate MPEG V-PCC-encoded volumetric videos. The quality of the avatar's face was manipulated by controlling the compression level of V-PCC, and enlisted participants to rate the videos based on their perception and satisfaction. Upon analyzing the assessment data collected from the participants, we found that the volumetric videos with higher facial quality get up to 39.7% higher mean opinion score (MOS) compared with the control group, demonstrating the crucial role of facial quality in influencing users' overall perceptual quality with volumetric videos. The findings highlight the importance of optimizing compression techniques to preserve the quality and realism of facial features, as they significantly contribute to users' immersive viewing experience. The generated videos and collected data is made publicly available to support further studies [1].

2 SUBJECTIVE QUALITY ASSESSMENT

2.1 Volumetric Video Generation

Volumetric Video Dataset. Four dynamic point clouds obtained from the 8i dataset [3] are employed for our study. These point clouds, namely *LongDress*, *RedAndBlack*, *Soldier*, and *Loot*, consist of 300 frames captured at a frame rate of 30 frames per second over 10 seconds.

Compression. In the compression step, we utilize the MPEG V-PCC reference software TMC2 (v15.0) [2] to encode the raw point cloud sequences. The compression rates in V-PCC are controlled by the geometry and texture quantization parameter (QP). The V-PCC common test condition (CTC) [4] defines five compression rates denoted as R5 to R1, where R5 corresponds to the highest quality (lowest compression), and R1 represents the lowest quality (highest compression). Additionally, an extra compression level is introduced, denoted as R0, which exhibits higher distortion levels compared to R1. The geometry QP and texture QP are set to 36 and 47, respectively. Detailed information regarding the CTC encoder settings, along with our additional setting, is summarized in Table 1.

Rendering. For rendering, we decode the compressed V-PCC streams using the MPEG V-PCC reference software TMC2 (v15.0) and employ the Open3D Python library (v0.14.1) to generate images for each frame in the point cloud sequences. The image dimensions are fixed at 600×1080, and the camera viewport is set to a frontal view with the object positioned at the center. The background color is gray (#898B88), and a point size of 1 is maintained consistently across all frames of the same model.

To introduce variations in the quality of the face region, we utilize the OpenCV Python library (v4.6.0) to detect the face of the avatar in each frame. This detection process provides us with the precise boundaries of the detected face region. To ensure a gradual transition in quality, a foveated-rendering-like strategy [6] is adopted. In addition to the detected face region, we include a peripheral area that encompasses the head and neck of the avatar. The size of this peripheral area is determined empirically, considering the anatomical proportions.

Table 1: Settings for MPEG V-PCC Reference Encoder

Rate Level	Occupancy Resolution	Q Factor (Geometry Map)	Q Factor (Attribute Map)
5	2	16	22
4	4	20	27
3	4	24	32
2	4	28	37
1	4	32	42
0	4	36	47

Therefore, each generated video consists of three distinct regions with decreasing qualities. The first region is the detected face region, which represents the highest-quality portion of the video. The second region is the peripheral area of the face, covering the head and neck, which exhibits a slightly lower quality level. The third region comprises the remaining body parts, exhibiting the lowest quality among the three regions. Figure 2(a) provides a visual representation of these regions for better understanding.

For comparative analysis, we incorporate a control group where higher qualities are assigned to the center of the human model. In these videos, three regions with decreasing qualities are present: (i) the region centered around the human body, with a size equivalent to that of the detected face region, (ii) the peripheral region, which matches the size of the peripheral region in the experimental (face) group, and (iii) the remaining body regions.

Consequently, for each avatar model, a total of eight videos are generated, each exhibiting a distinct quality switch pattern.

- Face-Li, $i = 1, 2, 3, 4$: Video quality starts with high-quality representation from the detected face region and gradually transitions to lower quality levels. This gives us four patterns: R5-R4-R3, R4-R3-R2, R3-R2-R1, and R2-R1-R0, which are labeled as Face-L4 to Face-L1.
- Center-Li, $i = 1, 2, 3, 4$: Video quality starts high from the body center region, and progresses to low quality, following the same four patterns: R5-R4-R3, R4-R3-R2, R3-R2-R1, R2-R1-R0. These patterns are denoted as Center-L4 to Center-L1, respectively.

Video Generation. Finally, we utilize FFmpeg [3] to create videos at 30fps with a duration of 10 seconds. Visually lossless H.264 parameters (-c:v libx264 -crf 15) are applied to control distortion while maintaining high quality.

2.2 Participants and Procedure

Participants were recruited through a university advertisement web page, meeting the criteria of being at least 18 years old, having normal vision, and no prior experience in picture quality evaluation. A total of 36 participants (21.5 years old on average) completed the study, with 20 identifying as female and 16 as male. They underwent

[1] https://github.com/nus-vv-streams/facial-quality

[2] https://github.com/MPEGGroup/mpeg-pcc-tmc2/releases/tag/release-v15.0

[3] https://ffmpeg.org

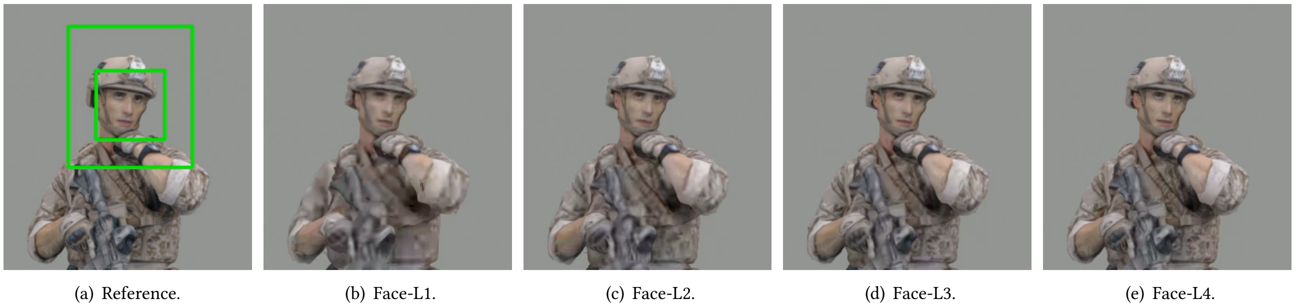


Figure 2: Sample frames of the *Soldier* model: (a) Reference, (b) Face-L1 (R2-R1-R0), (c) Face-L2 (R3-R2-R1), (d) Face-L3 (R4-R3-R2), (e) Face-L4 (R5-R4-R3). The quality switch regions are plotted in (a) for better illustration.

in-person vision tests before video evaluation. Participants failing the tests were excluded. Reimbursement of S$6 was offered for study completion, requiring 10-15 minutes on average.

The study procedure was conducted following the methodology proposed by Cox et al. [2], while adhering to the guidelines recommended by the International Telecommunication Union (ITU) [8, 9]. The user studies took place in a dimly lit room, with participants positioned at a fixed viewing distance equivalent to four times the height of the displayed model. Before the commencement of the study, participants were provided with a detailed explanation of the workflow and objectives, and their informed consent was obtained. Following ITU guidelines, participants underwent visual acuity tests using a Snellen eye chart to assess visual acuity, as well as Beck color plates to evaluate normal color vision.

After successfully passing the vision tests, participants received detailed instructions for the tasks and completed training to familiarize themselves with the interface and experimental procedures. During training, the participants were shown quality variations of the Matis (football player) model from the VSenseVVDB1 [13]. Following the training, participants proceeded to the main rating tasks. They viewed side-by-side videos for ten seconds, both originating from the same model. The left video represented unimpaired quality (reference), while the right video potentially contained impairments (trial). During the rating process, participants had the freedom to replay videos as needed, and no time limit was imposed for providing video ratings. Using the DCR method, participants were asked to rate the trial video's quality impairment using the scale:

- "1 - Very annoying,"
- "2 - Annoying,"
- "3 - Slightly annoying,"
- "4 - Perceptible but not annoying," and
- "5 - Imperceptible."

A total of 32 pairs of videos were presented, with the order of video presentation randomized for each participant.

3 RESULT ANALYSIS

We compare differences in mean ratings between Face-Li and Center-Li, $i = 1, 2, 3, 4$. Independent samples t-test is used to compare the mean rating. Table 2 shows mean differences and p-values for each comparison. Figure 3 plots the mean ratings with the change in quality levels. We can observe consistent and statistically significant differences ($p < 0.001$) in mean ratings between the Face-Li and Center-Li conditions across all quality switch patterns. The Face conditions consistently yield superior viewing quality in comparison to the Center conditions, with a notable increase of up to 39.7% in Mean Opinion Scores (MOS).

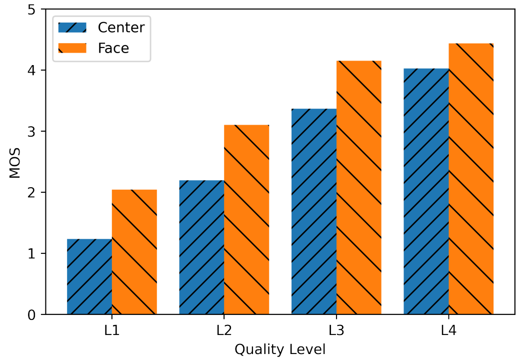


Figure 3: Mean ratings with the change of quality levels.

Table 2: Difference between mean ratings for Face-Li and Center-Li, $i = 1, 2, 3, 4$. The SD (standard deviations), SE (standard errors), the mean differences, and the p-values are also shown.

Quality Switch	N	Mean	SD	SE	Mean Difference	t-test p-value
Face-L1	36	2.04	0.59	0.10	0.81	<0.001
Center-L1	36	1.23	0.30	0.05	-0.81	<0.001
Face-L2	36	3.10	0.64	0.10	0.91	<0.001
Center-L2	36	2.19	0.57	0.09	-0.91	<0.001
Face-L3	36	4.15	0.39	0.06	0.78	<0.001
Center-L3	36	3.37	0.53	0.09	-0.78	<0.001
Face-L4	36	4.44	0.40	0.07	0.41	<0.001
Center-L4	36	4.03	0.46	0.07	-0.41	<0.001

Moreover, we can find that the mean differences at lower quality levels (L1 and L2) are larger than at higher quality levels (L3 and L4). Meanwhile, the mean difference increases with the increase of the quality level until L2, then decreases in subsequent levels, as shown in Figure 4. These observations indicate that when the quality level is lower, participants may be more sensitive to any improvement in the face stimuli since even small enhancements can have a noticeable impact on perceived quality. As the quality level increases, participants may reach a point where the quality is already satisfactory or meets their expectations. At this stage, further increases in quality may become less discernible or have diminishing returns in terms of perceived improvement. This could lead to a ceiling effect, where participants find it challenging to differentiate or appreciate additional quality increases.

4 CONCLUSION

In this paper, we conducted a user study to investigate the influence of facial quality on users' overall perceptual quality in MPEG V-PCC-encoded volumetric videos. The findings from the subjective quality assessment study emphasize the significant influence of facial quality on the perceptual experience when viewing volumetric videos. Our research highlights the importance of optimizing compression techniques to preserve the quality and realism of facial features, as they significantly contribute to users' immersive viewing experience. Further studies can build upon these findings to improve the design and implementation of compression algorithms for volumetric videos, considering the perceptual impact of facial quality and user behavior.

It is important to recognize the constraints of our work. In our study, we worked with volumetric videos which contain 3D data. However, when it comes to how our participants interacted with these videos, they did so using 2D displays. This is because the prevailing way people consume volumetric videos at present is through standard screens that provide a 2D viewing experience. We thus conducted our quality assessments based on how the participants perceived the videos in this 2D format. However, we are fully aware of the significance of assessing perceptual quality in a more immersive and realistic 3D context. In the future, we plan to expand our research to explore how viewers perceive volumetric videos when experienced in a true 3D environment.

ACKNOWLEDGMENTS

This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (T1 251RES2038).

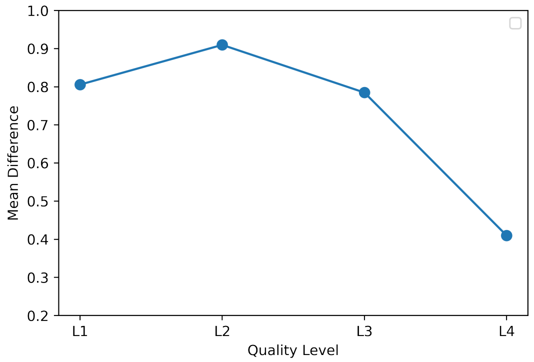


Figure 4: Mean Difference with the change of quality levels.

REFERENCES

[1] Evangelos Alexiou, Yana Nehmé, Emin Zerman, Irene Viola, Guillaume Lavoué, Ali Ak, Aljosa Smolic, Patrick Le Callet, and Pablo Cesar. Chapter 18 - subjective and objective quality assessment for volumetric video. In *Immersive Video Technologies*, pages 501–552. Academic Press, 2023.

[2] Samuel Rhys Cox, May Lim, and Wei Tsang Ooi. VOLVQAD: an MPEG V-PCC volumetric video quality assessment dataset. In *Proceedings of the 14th Conference on ACM Multimedia Systems, MMSys 2023, Vancouver, BC, Canada, June 7-10, 2023*, pages 357–362. ACM, 2023.

[3] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 8i voxelized full bodies-a voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006*, 7(8):11, 2017.

[4] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. Common test conditions for point cloud compression. *ISO/IEC JTC1/SC29/WG11 w17766*, 2018.

[5] Danillo Graziosi, Ohji Nakagami, Satoru Kuma, Alexandre Zaghetto, Teruhiko Suzuki, and Ali Tabatabai. An overview of ongoing point cloud compression standardization activities: Video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing*, 9:e13, Apr 2020.

[6] Yongjie Guan, Xueyu Hou, Nan Wu, Bo Han, and Tao Han. MetaStream: Live volumetric content capture, creation, delivery, and rendering in real time. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2023, Madrid, Spain, October 2-6, 2023*, pages 29:1–29:15. ACM, 2023.

[7] Kaiyuan Hu, Haowen Yang, Yili Jin, Junhua Liu, Yongting Chen, Miao Zhang, and Fangxin Wang. Understanding user behavior in volumetric video watching: Dataset, analysis and prediction. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1108–1116. ACM, 2023.

[8] ITU-R. Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation BT.500-13*, 2012.

[9] ITU-T. Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation*, page 910, 2008.

[10] Silvia Rossi, Irene Viola, and Pablo César. Behavioural analysis in a 6-DoF VR system: Influence of content, quality and user disposition. In *Proceedings of the 1st Workshop on Interactive eXtended Reality, IXR@MM 2022, Lisbon, Portugal, 14 October 2022*, pages 3–10. ACM, 2022.

[11] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo César. User centered adaptive streaming of dynamic point clouds with low complexity tiling. In *Proceedings of the 28th ACM International Conference on Multimedia, MM 2020, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3669–3677. ACM, 2020.

[12] Irene Viola, Shishir Subramanyam, Jie Li, and Pablo Cesar. On the impact of VR assessment on the quality of experience of highly realistic digital humans: A volumetric video case study. *Quality and User Experience*, 7(1):3, May 2022.

[13] Emin Zerman, Pan Gao, Cagri Ozcinar, and Aljosa Smolic. Subjective and objective quality assessment for volumetric video compression. In *Image Quality and System Performance XVI, Electronic Imaging 2019, IQSP, Burlingame, CA, USA, 13-17 January 2019*. Society for Imaging Science and Technology, 2019.

[14] Emin Zerman, Radhika Kulkarni, and Aljosa Smolic. User behaviour analysis of volumetric video in augmented reality. In *Proceedings of the 13th International Conference on Quality of Multimedia Experience, QoMEX 2021, Montreal, Canada, June 14-17, 2021*, pages 129–132. IEEE, 2021.

[15] Xuemei Zhou, Irene Viola, Evangelos Alexiou, Jack Jansen, and Pablo César. QAVA-DPC: eye-tracking based quality assessment and visual attention dataset for dynamic point cloud in 6 DoF. In *Proceedings of the 2023 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2023, Sydney, Australia, October 16-20, 2023*, pages 69–78. IEEE, 2023.

Progressive Coding for Deep Learning based Point Cloud Attribute Compression

Michael Rudolph
michael.rudolph@uni-due.de
University of Duisburg-Essen

Aron Riemenschneider
aron.riemenschneider@stud.uni-due.de
University of Duisburg-Essen

Amr Rizk
amr.rizk@uni-due.de
University of Duisburg-Essen

ABSTRACT

Progressive coding is a valuable technique for networked immersive media. As users approach objects in an immersive environment, progressive coding enables a gradual improvement of content quality. This effectively reduces bandwidth consumption compared to non-progressive methods that require to fully exchange a content representation by an independent, new representation.

In this work, we introduce an approach to progressively code point cloud attributes in a learned manner by compressing quantization residuals of each preceding representation through a learned, lightweight transformation in the entropy bottleneck. This allows to progressively reduce quantization errors using a single model in an end-to-end learning manner given the quantization residuals. In contrast to the state of the art that conditions the compression on a fixed rate-distortion, i.e. it requires an *ensemble of models* to build an adaptive streaming system, our approach requires only a *single model* during compression and decompression. We present preliminary results of our method, showing bandwidth savings for the scenario of a user approaching an object and gradually transitioning from low to high quality representations.

CCS CONCEPTS

• **Information systems** → Multimedia streaming; • **Computing methodologies** → Point-based models.

KEYWORDS

Virtual Reality, 6DOF, Point Cloud, Adaptive Streaming

ACM Reference Format:
Michael Rudolph, Aron Riemenschneider, and Amr Rizk. 2024. Progressive Coding for Deep Learning based Point Cloud Attribute Compression. In *16th International workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652217

1 INTRODUCTION

Point clouds are a popular representation format for volumetric data. They are easy to acquire through sampling points on an object's surface allowing to model arbitrary shapes. In terms of multimedia applications, that require textured representations of the object, an attribute vector is assigned to each point. One main challenge in the processing and distribution of point clouds stems from their immense data demand. Coupled with the user free movement to explore an immersive environment this effectively results in only a small subset of content actually being visible in the users' view-port at the cost of requiring very high data rates.

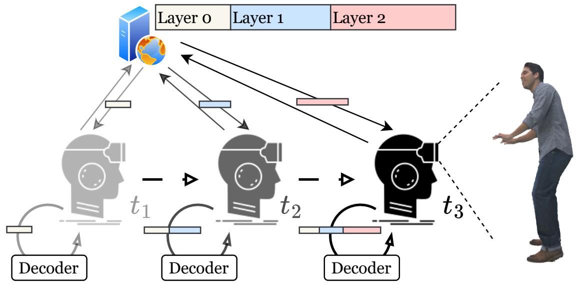


Figure 1: As a user approaches content of interest, progressive coding allows to gradually increase the quality through a number of enhancement layers.

Adaptive streaming techniques, being the de-facto standard in video streaming [2, 31] and allowing considerable bandwidth savings in free view-port video [18], have been transferred to point cloud content [19, 38] to reduce the bandwidth demand in scenes with multiple objects. Here, the proximity of the content to the users' view-port introduces an additional dimension [30] when adaptively selecting a set of qualities to maximize the Quality-of-Experience (QoE) given the clients' bandwidth constraints. With this in mind, we assume that content will be repeatedly re-transmitted, gradually exchanging the representation to increase the quality as the object gains of importance for a user. This stands in contrast to the currently available point cloud compression methods. V-PCC [1] and G-PCC [3] operate in static manner, i.e., they compress an independent bitstream per quality representation. Similarly, promising rate-distortion performance has been shown in a number of learning-based point cloud compression algorithms, handling the geometry [16, 32, 34, 43, 44], attributes [37, 42] or both modalities together [46]. However, with the exception of the latter, all learned compression approaches are *conditioned on a fixed rate-distortion trade-off*, leading to the *need for an ensemble of models* to build an adaptive streaming system.

In this work, we argue that progressive coding, i.e. a coding technique that compresses the point cloud into a layered bitstream, is a perfect match for the described requirements, allowing gradual increase in quality without the re-transmission of independent representations. While progressive coding has been prominently

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04…$15.00
https://doi.org/10.1145/3652212.3652217

employed in JPEG [40] and used in video compression [35], it regained traction recently, as it allows to convert learned, fixed-rate image compression methods into progressive codecs [24, 25, 27].

Summarizing our contributions given the need for progressive point cloud compression:

- We propose an approach for progressive point cloud attribute compression, requiring a single encoder and decoder model. The approach is optimized on all quality layers in an end-to-end manner.
- By stacking a sequence of entropy bottlenecks, we allow to extract individual features of the latent presentation to iteratively refine the reconstruction.
- Our evaluation shows significant potential for bandwidth savings when gradually transitioning from low quality to high quality in comparison to fixed-rate models.

2 PROBLEM STATEMENT

To introduce our problem we first briefly review the transform coding framework from [4]. There, the source data $\mathbf{x}$ is transformed in an autoencoder, consisting of an encoder used for the analysis transform g_a and a synthesis transform g_s for the decoder as

$$\mathbf{y} = g_a(\mathbf{x}; \phi_a) \tag{1}$$

$$\hat{\mathbf{y}} = Q(\mathbf{y}) \tag{2}$$

$$\hat{\mathbf{x}} = g_s(\hat{\mathbf{y}}; \phi_s) \tag{3}$$

where g_s and g_a are realized as neural networks parameterized by ϕ_a and ϕ_s, respectively. After quantization, the latent representation $\hat{\mathbf{y}}$ is entropy coded for compression. To further exploit spatial correlation in the down-sampled latent representation $\mathbf{y}$, it is common practice to estimate the local means μ and variances σ of the elements in $\mathbf{y}$, which can be compressed as side-information through a hyperprior model [5, 29]. This hyperprior model consists of a hyperanalysis h_a and a hypersynthesis h_s, allowing to locally model the distribution of latents $\mathbf{y}$ through a Gaussian distribution.

During training, quantization is then substituted through additive uniform noise, serving as a proxy for the actual quantization effect while allowing to back-propagate gradients through the bottleneck. As a result, this approach is *restricted to a single rate configuration*, imposed by introducing a trade-off parameter λ during training to balance the loss terms for rate and distortion as in

$$\mathcal{L} = \mathcal{R} + \lambda \mathcal{D} \tag{4}$$

Now the main problem is that allowing encoding to *multiple rate configurations* requires training an ensemble of models conditioned by λ, resulting in a set of parameters for the encoder and decoder, namely $\mathbf{\Theta}_a = \{\theta_a^{(i)} | i = 1, ..., n\}$ for the encoder and $\mathbf{\Theta}_s = \{\theta_s^{(i)} | i = 1, ..., n\}$ for the decoder. Additionally, as each bitstream is compressed independently, this requires to fully *exchange the representation* at the client side when transitioning from a low quality to a high quality representation, resulting in repeated requests for the same content and thus high bandwidth utilization. Motivated by these observations, our goal in this work is to achieve an additive decomposition of the latent representation which is reminiscent of training a model with stacked entropy bottlenecks. Note that this stands in contrast to approaches from image compression, that reduce quantization residuals in a fixed manner through nested quantization [24, 25, 27].

3 RELATED WORK

Point Cloud compression for multimedia content is challenging. Standard approaches mainly handle geometry and attributes individually to address their different characteristics.

Geometry Compression While we focus on attribute compression, assuming the availability of a perfect geometry of the point cloud, it is worthwhile to also review methods for learned geometry compression approaches. Early work [16, 32] proposes to use 3D convolutions on voxelized point clouds and leverage learned entropy models [4]. Through introducing elaborate entropy models [29], and more capable architectures for encoding and decoding, rate-distortion performance has subsequently improved as in [34, 44]. Most notably, the introduction of sparse convolutions [9] allows to drastically reduce the latency and memory requirements when operating on voxelized representations [43]. Finally, employing group-based decoding allows to leverage correlations between upsampled voxels in a parallelized manner [42] and to extend this approach to allow inter-frame coding for dynamic point cloud sequences [41].

Attribute Compression Attribute compression is mainly dominated by traditional approaches that rely on Graph Transform [45] or Region Adaptive Hierarchical Transforms (RAHT) [11] to transform coefficients on the irregular geometry. On the other hand, V-PCC [1] leverages projections into 2D patches, which are then compressed using video codecs. Investigating more natural folding techniques, the authors of [33] explored a learned projection method to fold a 2D grid on a point cloud.

With the emergence of learning-based approaches, the work in [14] uses a deep entropy model to capture the probability distribution of coefficients after applying RAHT [11], thus increasing its compression performance. Fully relying on learned transformations, the authors of [37] use a point-based model to compress attributes, relying on computationally expensive multilayer perceptrons, and thus requiring to only process small blocks of the point cloud at a time. Recently, sparse tensor autoencoders from [42] have shown comparable results to traditional attribute compression approaches, which may be related to drastically increasing the availability of training data through synthetically projecting textures on uncolored point clouds.

Point Cloud Codecs Combining selected methods from above, numerous full-fledged compression algorithms have been proposed to jointly handle geometry and attribute compression. Mekuria *et al.* [28] developed a codec utilizing octree partitions and intra-frame prediction while projecting the attributes to 2D grids and compressing them with legacy image codecs, enabling further user-centered studies for immersive video [39]. Recently, two MPEG standard proposals emerged [36], distinguishing between static point cloud compression using Geometry-based PCC (G-PCC) and dynamic point cloud compression with Video-based PCC (V-PCC). While G-PCC relies on octree partitions for geometry compression and methods such as RAHT [11] or lifting schemes for attributes, V-PCC projects geometry and attribute information into video frames

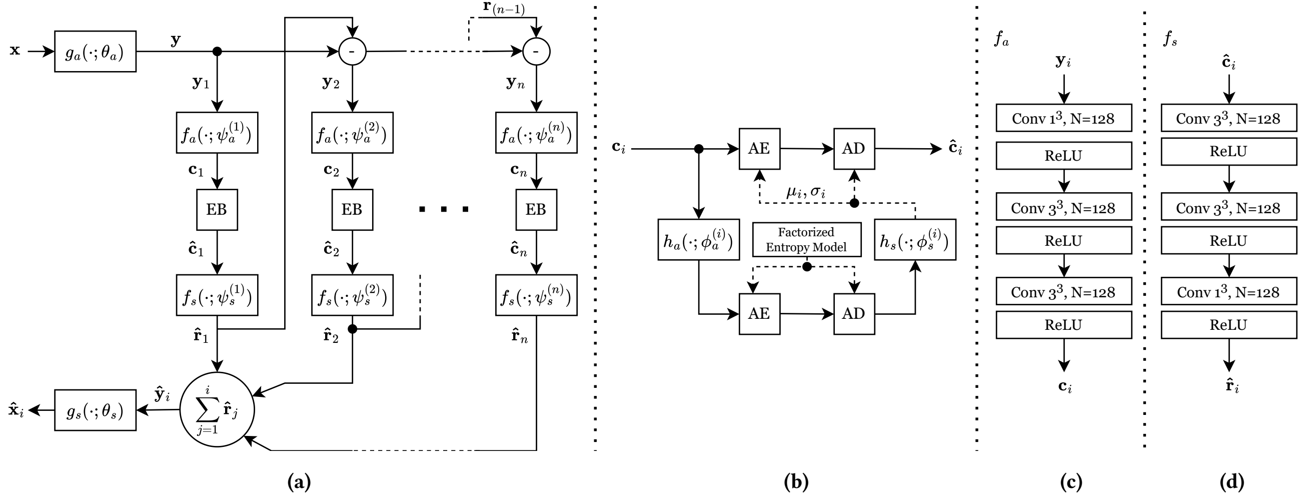


Figure 2: (a) Stacking entropy bottlenecks for layered coding. The representation $\hat{\mathbf{x}}_i$ is reconstructed by decoding the first i layers of the bitstream. (b) Entropy Bottleneck (EB), as proposed by [5, 29], locally models the distribution of latents c_i through a Gaussian distribution with mean μ and variance σ. (c) A non-linearity f_a extracts relevant features for compression at each stage, while (d) f_s reprojects the quantized coefficients to subtract from the latents of the previous stage.

and leverages video codecs for compression to exploit temporal correlations between frames. Most recently, the first learning based approach for geometry and attribute compression, has been proposed [46], allowing rate-control through adaptive quantization.

Progressive Coding To the best of our knowledge, Progressive Coding has not been studied for the domain of point cloud compression, but attracted some interest in learned image compression mainly building on the framework proposed in [4, 5, 29]. Early work partitions the latent representation of an image signal derived through an encoder network into a base and enhancement representation, both being decoded through dedicated decoder networks, allowing a preview representation of the content [8]. Following a similar goal and encoding a set of enhancement layers, multiple approaches utilize Recurrent Neural Networks to iteratively encode the quantization residuals [12, 20, 21], therefore requiring repeated execution of both the encoder and decoder during compression.

Following a different approach, Lu *et al.* [27] propose a nested quantization approach for transforming a trained, fixed-rate model into a progressive one. In detail, multiple quantization grids are proposed, allowing to reduce the quantization error from the coarse to the fine grid using a conditional probability model for each refinement. Additionally, they propose to order elements of the latent representation by their estimated variance, allowing for even more fine-grained rate-control. Progressively dividing quantization bins into 3 segments [24] follows a similar approach, but requires learned post-processing to reduce artifacts in lower-quality results caused by not considering the progressive coding scheme at training time. Similarly, Li *et al.* [25] replace the uniform quantizer through a dead-zone quantizer to counteract symbol redundancy when performing nested quantization.

Extending the ideas of TailDrop [23], a progressive learning scheme is proposed in [17] to order the channels of the latent representation by importance, thus allowing to select arbitrary ranges of the channels for progressive decoding without the need for nested quantization.

Most relevant for our approach are the works on nested quantization [24, 25, 27], which aim at progressively encoding quantization residuals, but rely on a *trained, fixed-rate model.* As a result, the encoder and decoder are not conditioned on variable quantization at training time. Hence, it shows reduced performance when aiming for low-rate representations, which is partially counteracted by learned post-processing models in [24].

4 METHOD

4.1 Model Architecture for additive decomposition of the latent representation

Our model follows the transform coding framework [4], reviewed in Section 2, implementing the encoder g_s, the decoder g_a and the hyperprior models h_a and h_s according to the architecture proposed in [42]. The latent representation $\mathbf{y}$ is then iteratively decomposed using a number of stacked hyperprior models [5, 29] to achieve an additive decomposition of the latent representation, as depicted in Fig. 2a. Specifically, at each level, we encode the compression residual $\mathbf{y}_i$ of the preceding layer, using

$$\mathbf{y}_i = \begin{cases} \mathbf{y} & \text{if } i = 1 \\ \mathbf{y}_i - \hat{\mathbf{r}}_{i-1} & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{y}_1 = \mathbf{y}$ is used for the base layer.

Instead of directly encoding $\mathbf{y}_i$ at each stage, we introduce two neural network blocks f_a and f_s, depicted in Fig. 2c and Fig. 2d to wrap each entropy model. This is motivated by the following goals and observations: i) We want the model to learn which features should be compressed at each stage, allowing to disable certain elements of the current residual, and ii) directly computing the residual after quantization might not result in a representation suiting further entropy coding as quantization errors are not linearly linked to distortions in the reconstruction. Consequently, at each stage, a transformed latent representation $\mathbf{c}_i = f_a(\mathbf{y}_i; \psi_a^{(i)})$ is extracted and entropy coded using the hyperprior-model of the current stage. After decoding, the quantized representation $\hat{\mathbf{c}}_i = Q(\mathbf{c_i})$ is then expanded through $\hat{\mathbf{r}} = f_s(\hat{\mathbf{c}}_i; \psi_s^{(i)})$. This allows to compute the residual of the latent representation of the next layer through Eq. 5 to compress the next refinement layer of the bitstream. During decoding, the latent representation of the i layers available at decoding time is obtained through

$$\hat{\mathbf{y}}_i = \sum_{j=0}^{i} \hat{\mathbf{r}}_j \tag{6}$$

and used to reconstruct the attributes of the point cloud $\hat{\mathbf{x}}_i = g_s(\hat{\mathbf{y}}_i; \phi_s)$ through the synthesis transform. Hereby, the analysis and synthesis transforms' parameters are shared over all levels to reduce the model size.

4.2 Training

For training, we follow common procedure (cf. [4]), using a Lagrangian loss function to balance rate and distortion. However, as we aim for a set of quality representations, which we jointly optimize, we decode all quality levels $\hat{\mathbf{x}}_i$ during training and introduce a set of weighting parameters $\lambda = \{\lambda_1, ..., \lambda_n\}$, one for each level of the decomposition.

As quantization hinders optimization using gradient descent, additive uniform noise is used as a drop-in proxy for quantization in the bottleneck, allowing to back-propagate gradients to the encoder [4]. Note that this hinders training of further layers, effectively rendering the computed residuals of the consecutive level useless through the additive uniform noise. To allow effective training of the ith consecutive layer, we exchange the residuals $\tilde{\mathbf{r}}_j, j = 1, ..., i-1$ through their actual quantized counterparts $\hat{\mathbf{r}}_j, j = 1, ..., i-1$, resulting in the training latent residual to be compressed at the ith stage as

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{y}}_i + \sum_{j=1}^{i-1} \hat{\mathbf{r}}_j. \tag{7}$$

This results in gradients only being back-propagated through the bottleneck belonging to the last stage of the ith reconstruction and ensures that each level is optimized over an equal amount of samples, opposed to [17]. Finally, we formulate the loss function as a weighted sum of all reconstruction errors and approximate the rate of the latents for each level by concatenating all likelihoods of the previous layers into $\tilde{\mathbf{c}}_{\mathbf{1:i}} = [\tilde{c}_0, ..., \tilde{c}_i]$ and $\tilde{\mathbf{z}}_{\mathbf{1:i}} = [\tilde{z}_0, ..., \tilde{z}_i]$ in

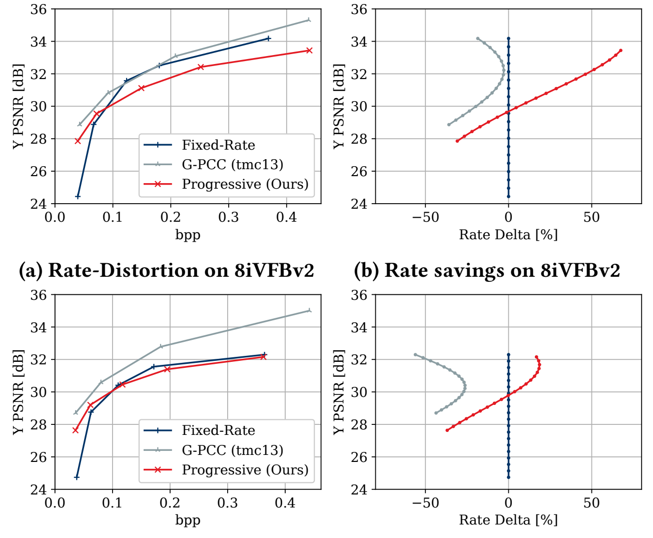
$$\mathcal{L} = \mathbb{E}_{x \sim p(x)} \sum_{i=1}^{n} \lambda_i \|\mathbf{x} - \tilde{\mathbf{x}}_i\|_2^2 - \log_2 p_{\tilde{\mathbf{c}}}(\tilde{\mathbf{c}}_{1:i}) - \log_2 p_{\tilde{\mathbf{z}}}(\tilde{\mathbf{z}}_{1:i}) \tag{8}$$

This allows joint optimization of all levels, ensuring the shared encoder and decoder being conditioned on all latent residuals.

5 EVALUATION

5.1 Implementation

We implement our approach in PyTorch, using MinkowskiEngine [9] to leverage sparse convolutions in our model and CompressAI [6] for the implementation of entropy bottlenecks. For training, 240 point clouds from the UVG dataset [15] in 10 bit resolution are sampled and sliced into cubes of size 128^3. As the variation of textures in the dataset is limited, images from the Describable Texture Dataset [10] are projected on the point clouds, similar to the approach in [42].



(a) Rate-Distortion on 8iVFBv2 (b) Rate savings on 8iVFBv2

(c) Rate-Distortion on MVUB (d) Rate savings on MVUB

Figure 3: Rate-Distortion curves, averaged on the test point clouds in Table 1, comparing our approach with the fixed-rate model [42] and G-PCC [3]. Rate is calculated per configuration.

We select Adam [22] as an optimizer, using an initial learning rate of 10^{-3} for the model, which is reduced by factor 0.5 after 30 epochs. The auxiliary loss is optimized with learning rate 0.01. Gradient norm clipping with threshold 0.1 is used to stabilize the training. Training and inference is conducted on a NVIDIA GeForce 4090. The fixed-rate model is derived from [42] and retrained under the same conditions as our progressive approach.

The code for training and testing our model is made available on GitHub[1].

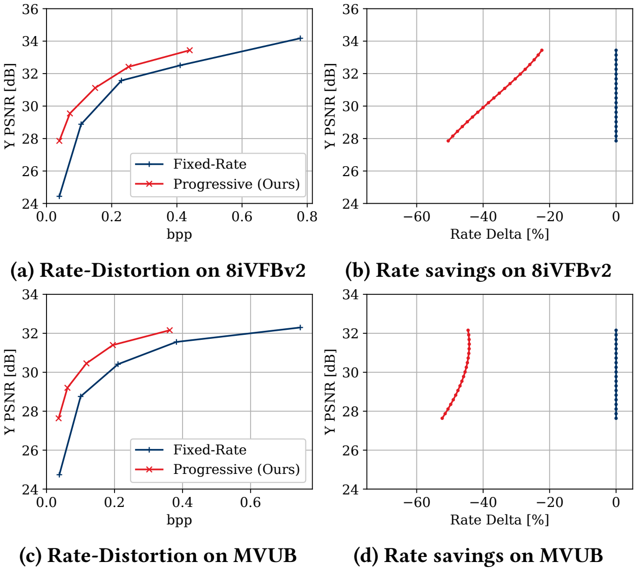
5.2 Results

For the rate-distortion evaluation, test frames from the 8iVFBv2 [13] and MVUB [26] dataset are selected. The resulting rate-distortion curves are depicted in the left column of Fig. 3, comparing our progressive scheme against G-PCC [3] and the fixed-rate approach [42] on both datasets. The right-hand side of the figure depicts potential rate-savings according to the Bjøntegaard rate delta [7], using the fixed-rate method as a reference. Hereby, negative values indicate rate savings while positive values indicate the increased bandwidth requirement to deliver the same quality. Note that the Bjøntegaard model used to compute the right-hand column of Fig. 3 is fitted to the averaged rate-distortion points over all selected frames in the dataset using the least squares method. Additionally, per-frame results are reported in Table 1 for Y-PSNR and weighted YUV-PSNR

[1] https://github.com/mic-rud/ProgressivePCAC

Table 1: Rate-Distortion Performance of our progressive model, using the fixed-rate model as reference (left) and the progressive approach, accumulating bits over five qualities from low to high quality (right). While the fixed-rate model requires less bits when requesting a specific quality, our model offers substantial reduction in bandwidth when progressively transitioning from low to high quality.

			Rate-Distortion Performance				Transition from low to high			
			Y		YUV		Y		YUV	
Dataset	Sequence	Frame	$\Delta_r \downarrow$	$\Delta_{PSNR} \uparrow$	$\Delta_r \downarrow$	$\Delta_{PSNR} \uparrow$	$\Delta_r \downarrow$	$\Delta_{PSNR} \uparrow$	$\Delta_r \downarrow$	$\Delta_{PSNR} \uparrow$
8iVFBv2 [13]	Longdress	1300	4.71%	-0.07 dB	5.87%	-0.08 dB	-41.03%	1.06 dB	-39.30%	1.05 dB
	Soldier	690	23.31%	-0.46 dB	8.58%	-0.48 dB	-32.47%	1.22 dB	-40.82%	1.77 dB
	Loot	1200	23.82%	-0.12 dB	2.50%	0.29 dB	-36.28%	2.01 dB	-46.53%	2.62 dB
	Redandblack	1550	13.45%	-0.02 dB	7.87%	0.03 dB	-34.88%	1.50 dB	-37.17%	1.52 dB
MVUB [26]	Andrew	1	-2.04%	0.05 dB	-16.17%	0.18 dB	-45.82%	0.62 dB	-52.40%	0.88 dB
	David	1	8.99%	0.30 dB	-3.30%	0.40 dB	-41.43%	2.02 dB	-46.41%	2.09 dB
	Phil	1	7.95%	-0.15dB	5.05%	0.02 dB	-38.98%	1.42 dB	-40.25%	1.41 dB
	Sarah	1	44.45%	1.09 dB	12.23%	1.05 dB	-18.85%	2.32 dB	-22.82%	2.32 dB



(a) Rate-Distortion on 8iVFBv2

(b) Rate savings on 8iVFBv2

(c) Rate-Distortion on MVUB

(d) Rate savings on MVUB

Figure 4: Progressive Rate-Distortion curves, assuming the transition from low to high quality with cumulative bits for the fixed-rate model [42].

using weights $(\frac{6}{8}, \frac{1}{8}, \frac{1}{8})$, showing a comparison between our approach with the fixed-rate model as a reference. Over all frames in both datasets, we notice a decrease in compression performance when aiming for higher quality representations using the progressive model compared to both G-PCC and the fixed-rate counterpart. However, when aiming for higher compression ratio, the progressive model shows potential for rate reduction, i.e. a more attractive rate-distortion trade-off at lower rates compared to it's fixed-rate counterpart. Overall, the progressive model requires on average 16.32% more bits on the 8iVFBv2 [13] dataset and 14.85% more bits on the MVUB [26] dataset. Considering the results in Table 1, a significant, data-dependent differences between the sequences becomes apparent: While the point clouds *Longdress* and *Andrew* with their demanding colors show very little differences between the performance of the fixed-rate baseline and the progressive approach, the point cloud *Sarah* with attributes that can be considered simple, results in strong deterioration for the performance of the progressive approach compared to the fixed-rate model. Similarly, the point clouds *Loot* and *Soldier* can be compressed more efficiently using the fixed-rate model.

While the capabilities of delivering higher rate representations using a progressive model compared to the fixed-rate baseline is reduced, progressive coding shows its strength when gradually transitioning from a low to a high quality representation. Given this assumption, we accumulate the bits per point (bpp) required for the transition over five quality levels in Fig 4 and the right-hand columns of table 1. As a result, the fixed-rate model requires substantially more bandwidth, as it requires to fully exchange the representation of a point cloud, while the progressive model only requires an additional layer. This contributes to substantial rate-savings considering the Y-PSNR and YUV-PSNR quality over all tested frames as shown in Table 1. Similarly, the averaged rate-distortion points in Fig. 4 and the resulting rate-delta curves confirm this observation. However, the presented results for the potential of progressive transitioning have to be interpreted with caution: If the transition in the fixed-rate model is performed by skipping a rate-level, the consecutive rate-distortion points are shifted to the left as the bandwidth for the omitted representation is saved. In contrast, the progressive model does not allow skipping quality levels, i.e. to decode a specific quality, it always requires all preceding layers.

Finally, the resulting reconstructions are rendered in Fig. 5, selecting the first, second and fifth layer of the progressive model and the corresponding fixed-rate reconstructions at similar quality. The last row shows the error in attribute reconstruction in the luminance channel (Y-PSNR) for each layer of the progressive model, showcasing how high-frequency details are omitted in the first layers of the bitstream to be compressed in later enhancement layers. For the fifth layer, the progressive model achieves 0.75 dB

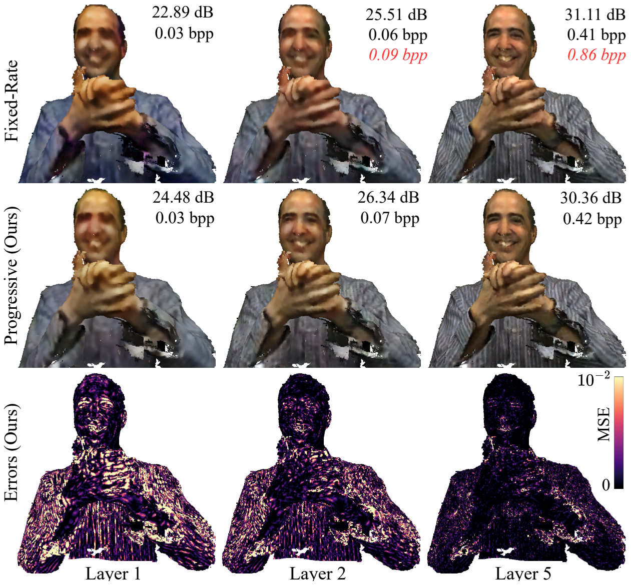


Figure 5: Renders of *Phil* [26] at multiple rate configurations. Y-PNSR and bpp for the reconstruction reported on the top. For the fixed-rate model [42], cumulative bit for transitioning through all qualities are given in red.

less Y-PSNR quality using the same bandwidth when directly requesting the respective quality, but allows saving 51% bandwidth when transitioning over all five quality levels.

6 DISCUSSION AND FUTURE WORK

In this work, we presented an approach for progressive point cloud attribute compression, decomposing the latent representation from the encoder into a set of additive residuals, which are consecutively coded through a number of stacked entropy bottlenecks. This allows progressive coding in a low-to-high quality approach, delivering comparable rate-distortion to its fixed-rate counterpart conditioned for a low-rate encoding, but a decreased rate-distortion performance for higher rate encodings. Assuming an immersive client approaching an object and hence gradually exchanging the representation of the object from low to high quality over multiple levels, we find that the progressive coding approach allows for considerable bandwidth savings. This can be attributed to only requiring to request enhancement layers, while fixed-rate models force users to request an independent representation for each quality switch.

The high quality for low rate encodings of our approach stands in contrast to the observations made by authors of progressive coding techniques proposed for image compression [24, 25, 27], who notice a drop in rate-distortion performance when aiming for lower rates. However, for a fair comparison this requires transferring these methods to the application of point cloud compression.

Finally, applying progressive coding techniques to point cloud geometry compression remains an open topic. While the presented approach is restricted to static point clouds, extending progressive schemes to dynamic point clouds is a promising direction. We anticipate that this allows for more flexibility when optimizing the user QoE during playout, reducing the cost for increasing the quality of a segment *after* the base layer has already been requested.

REFERENCES

[1] ISO/IEC JTC 1/SC 29. 2021. *ISO/IEC 23090-5:2021, Information technology — Coded representation of immersive media — Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC).* ISO/IEC.

[2] ISO/IEC JTC 1/SC 29. 2022. *ISO/IEC 23009-1:2022, Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats.* ISO/IEC.

[3] ISO/IEC JTC 1/SC 29. 2023. *ISO/IEC 23090-9:2023, Information technology — Coded representation of immersive media — Part 9: Geometry-based point cloud compression.* ISO/IEC.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. 2016. End-to-end optimization of nonlinear transform codes for perceptual quality. In *Picture Coding Symposium (PCS).* IEEE, 1–5.

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436* (2018).

[6] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. 2020. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029* (2020).

[7] Gisle Bjontegaard. 2001. Calculation of average PSNR differences between RD-curves. *ITU-T SG16/Q.16, 33th VCEG Meeting* (2001).

[8] Chunlei Cai, Li Chen, Xiaoyun Zhang, Guo Lu, and Zhiyong Gao. 2019. A novel deep progressive image compression framework. In *Picture Coding Symposium (PCS).* IEEE, 1–5.

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)).* 3075–3084.

[10] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing Textures in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[11] Ricardo L De Queiroz and Philip A Chou. 2016. Compression of 3D point clouds using a region-adaptive hierarchical transform. *IEEE Transactions on Image Processing* 25, 8 (2016), 3947–3956.

[12] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Drasic: Distributed recurrent autoencoder for scalable image compression. In *Data Compression Conference (DCC).* IEEE, 3–12.

[13] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 2017. 8i voxelized full bodies-A voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006* 7, 8 (2017), 11.

[14] Guangchi Fang, Qingyong Hu, Hanyun Wang, Yiling Xu, and Yulan Guo. 2022. 3dac: Learning attribute compression for point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 14819–14828.

[15] Guillaume Gautier, Alexandre Mercat, Louis Fréneau, Mikko Pitkänen, and Jarno Vanne. 2023. UVG-VPC: Voxelized Point Cloud Dataset for Visual Volumetric Video-based Coding. In *International Conference on Quality of Multimedia Experience (QoMEX).* IEEE, 244–247.

[16] André FR Guarda, Nuno MM Rodrigues, and Fernando Pereira. 2019. Point cloud coding: Adopting a deep learning-based approach. In *Picture Coding Symposium (PCS).* IEEE, 1–5.

[17] Ali Hojjat, Janek Haberer, and Olaf Landsiedel. 2023. ProgDTD: Progressive Learned Image Compression With Double-Tail-Drop Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.* 1130–1139.

[18] Jeroen Van der Hooft, Maria Torres Vega, Stefano Petrangeli, Tim Wauters, and Filip De Turck. 2019. Tile-based adaptive streaming for virtual reality video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 4 (2019), 1–24.

[19] Mohammad Hosseini and Christian Timmerer. 2018. Dynamic adaptive point cloud streaming. In *23rd Packet Video Workshop.* 25–30.

[20] Khawar Islam, L Minh Dang, Sujin Lee, and Hyeonjoon Moon. 2021. Image compression with recurrent neural network and generalized divisive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 1875–1879.

[21] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. 2018. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 4385–4393.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Toshiaki Koike-Akino and Ye Wang. 2020. Stochastic bottleneck: Rateless autoencoder for flexible dimensionality reduction. In *IEEE International Symposium on Information Theory (ISIT).* IEEE, 2735–2740.

[24] Jae-Han Lee, Seungmin Jeon, Kwang Pyo Choi, Youngo Park, and Chang-Su Kim. 2022. DPICT: Deep progressive image compression using trit-planes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16113–16122.

[25] Shaohui Li, Han Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. 2022. Learned Progressive Image Compression With Dead-Zone Quantizers. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[26] Charles Loop, Qin Cai, S Orts Escolano, and Philip A Chou. 2016. Microsoft voxelized upper bodies-A voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m38673 M* 72012 (2016), 2016.

[27] Yadong Lu, Yinhao Zhu, Yang Yang, Amir Said, and Taco S Cohen. 2021. Progressive neural image compression with nested quantization and latent ordering. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 539–543.

[28] Rufael Mekuria, Kees Blom, and Pablo Cesar. 2016. Design, implementation, and evaluation of a point cloud codec for tele-immersive video. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (2016), 828–842.

[29] David Minnen, Johannes Ballé, and George D Toderici. 2018. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems* 31 (2018).

[30] Minh Nguyen, Shivi Vats, Sam Van Damme, Jeroen Van Der Hooft, Maria Torres Vega, Tim Wauters, Christian Timmerer, and Hermann Hellwagner. 2023. Impact of Quality and Distance on the Perception of Point Clouds in Mixed Reality. In *International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 87–90.

[31] Roger Pantos and William May. 2017. *HTTP live streaming*. Technical Report.

[32] Maurice Quach, Giuseppe Valenzise, and Frederic Dufaux. 2019. Learning convolutional transforms for lossy point cloud geometry compression. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 4320–4324.

[33] Maurice Quach, Giuseppe Valenzise, and Frederic Dufaux. 2020. Folding-based compression of point cloud attributes. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 3309–3313.

[34] Maurice Quach, Giuseppe Valenzise, and Frederic Dufaux. 2020. Improved deep point cloud geometry compression. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[35] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. 2007. Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Transactions on circuits and systems for video technology* 17, 9 (2007), 1103–1120.

[36] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. 2018. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2018), 133–148.

[37] Xihua Sheng, Li Li, Dong Liu, Zhiwei Xiong, Zhu Li, and Feng Wu. 2021. Deep-pcac: An end-to-end deep lossy compression framework for point cloud attributes. *IEEE Transactions on Multimedia (TOMM)* 24 (2021), 2617–2632.

[38] Shishir Subramanyam, Irene Viola, Alan Hanjalic, and Pablo Cesar. 2020. User centered adaptive streaming of dynamic point clouds with low complexity tiling. In *ACM International Conference on Multimedia (MM)*. 3669–3677.

[39] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. Vr2gather: A collaborative social vr system for adaptive multi-party real-time communication. *IEEE MultiMedia* (2023).

[40] Gregory K Wallace. 1991. The JPEG still picture compression standard. *Commun. ACM* 34, 4 (1991), 30–44.

[41] Jianqiang Wang, Dandan Ding, Hao Chen, and Zhan Ma. 2023. Dynamic Point Cloud Geometry Compression Using Multiscale Inter Conditional Coding. *arXiv preprint arXiv:2301.12165* (2023).

[42] Jianqiang Wang, Dandan Ding, Zhu Li, Xiaoxing Feng, Chuntong Cao, and Zhan Ma. 2022. Sparse tensor-based multiscale representation for point cloud geometry compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[43] Jianqiang Wang, Dandan Ding, Zhu Li, and Zhan Ma. 2021. Multiscale point cloud geometry compression. In *2021 Data Compression Conference (DCC)*. IEEE, 73–82.

[44] Jianqiang Wang, Hao Zhu, Haojie Liu, and Zhan Ma. 2021. Lossy point cloud geometry compression via end-to-end learning. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 12 (2021), 4909–4923.

[45] Cha Zhang, Dinei Florencio, and Charles Loop. 2014. Point cloud attribute compression with graph transform. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2066–2070.

[46] Junteng Zhang, Tong Chen, Dandan Ding, and Zhan Ma. 2023. YOGA: Yet Another Geometry-based Point Cloud Compressor. In *ACM International Conference on Multimedia (MM)*. 9070–9081.

Volumetric Video Compression Through Neural-based Representation

Yuang Shi
National University of Singapore
Singapore
yuangshi@u.nus.edu

Ruoyu Zhao
Tsinghua University
China
zhao-ry20@mails.tsinghua.edu.cn

Simone Gasparini
IRIT - University of Toulouse
France
simone.gasparini@toulouse-inp.fr

Géraldine Morin
IRIT - University of Toulouse
France
geraldine.morin@toulouse-inp.fr

Wei Tsang Ooi
National University of Singapore
Singapore
ooiwt@comp.nus.edu.sg

ABSTRACT

Volumetric video offers immersive exploration and interaction in 3D space, revolutionizing visual storytelling. Recently, Neural Radiance Fields (NeRF) have emerged as a powerful neural-based technique for generating high-fidelity images from 3D scenes. Building upon NeRF advancements, recent works have explored NeRF-based compression for static 3D scenes, in particular point cloud geometry. In this paper, we propose an end-to-end pipeline for volumetric video compression using neural-based representation. We represent 3D dynamic content as a sequence of NeRFs, converting the explicit representation to neural representation. Building on the insight of significant similarity between successive NeRFs, we propose to benefit from this temporal coherence: we encode the differences between consecutive NeRFs, achieving substantial bitrate reduction without noticeable quality loss. Experimental results demonstrate the superiority of our method for dynamic point cloud compression over geometry-based PCC codecs and comparable performance with state-of-the-art PCC codecs for high-bitrate volumetric videos. Moreover, our proposed compression based on NeRF generalizes to arbitrary dynamic 3D content.

CCS CONCEPTS

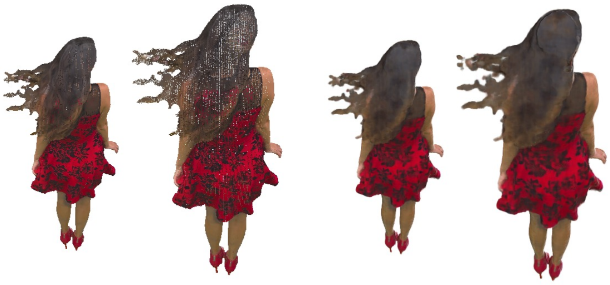
• **Information systems → Multimedia streaming**; • **Computing methodologies → Animation**.

KEYWORDS

Volumetric video; Point cloud compression; Neural radiance fields; Temporal coherence

ACM Reference Format:
Yuang Shi, Ruoyu Zhao, Simone Gasparini, Géraldine Morin, and Wei Tsang Ooi. 2024. Volumetric Video Compression Through Neural-based Representation. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652220

MMVE '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0618-9/24/04.
https://doi.org/10.1145/3652212.3652220



(a) Point cloud representation.

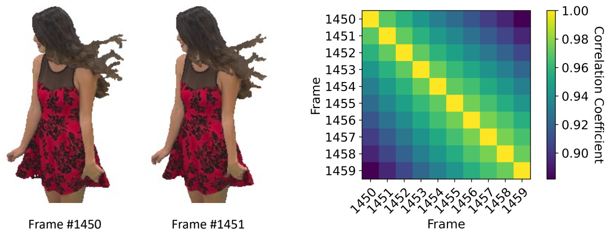
(b) NeRF reresentation.

Figure 1: Sample rendered images from a point cloud (left) showing visual artifacts due to its discrete nature which does not affect images rendered with NeRF (right).

1 INTRODUCTION

Volumetric video captures a 3D representation of a real-world scene or subject, allowing viewers to explore and interact with the captured content in six degrees of freedom (6DoF). Volumetric video is likely to play an increasingly important role in various industries, enabling new forms of visual storytelling and immersive experiences. In contrast to traditional 2D video, which has standard and mature forms of representation, 3D volumetric video has a plethora of representation formats. The representations of volumetric video can be categorized into explicit and implicit representations.

Most existing works are based on explicit 3D representations because they are easy to process through classical rendering pipelines. Textured mesh is the most classical 3D model, but 3D point clouds, which consist of a set of 3D points with coordinates and color, have gained much popularity as the choice for high-quality representation for volumetric video as they are more adapted to dynamic acquisition. In order to provide a high-quality immersive experience with limited network conditions and computational resources, point cloud compression (PCC) techniques are paramount for volumetric video streaming. For example, Google's Draco [8], MPEG's video-based PCC (V-PCC), and MPEG's geometry-based PCC (G-PCC) [22] are three typical PCC codecs. Nevertheless, because of their discrete nature, point clouds can easily present visual artifacts that affect the visual quality [15]. For instance, point clouds may



(a) Neighboring frames.

(b) Similarity matrix of NeRFs.

Figure 2: Example of temporal redundancy in (a) point clouds and (b) NeRFs.

cause holes when being projected to screen space [16], which can be seen in Figure 1(a).

Given that explicit representations fail to achieve photo-realistic rendering quality, the latest advancements in implicit neural representations, especially neural radiance fields (NeRF) [18], have gained more popularity. NeRF [18] is a neural-based novel view synthesis technique. Given a set of 2D RGB images of a 3D scene, NeRF can model it as a neural radiance field with multilayer perceptrons (MLPs), and render immersive and high-fidelity novel views from this representation. Figure 1(b) shows an example of the rendered images from NeRF. Overall, NeRF has gained recognition as an effective approach [14, 27] for accurately representing the dynamic interactions between light and color in three-dimensional space. Because of its ability to generate highly realistic images from 3D scenes, utilizing the recent advancements in NeRF for 3D content compression has become an attractive avenue. For instance, Bird *et al.* [3] adopt NeRF to represent 3D static scenes and apply an entropy penalty for model compression. Hu *et al.* [10] leverage NeRF to represent the geometry of 3D point clouds. Quantization and entropy encoding are then applied to compress neural networks, achieving comparable rate-distortion (R-D) performance with G-PCC. Although previous works make an effective step toward NeRF-based 3D content compression, there is still a big gap into practical volumetric video compression. The key challenge is to maintain the high-quality representation of volumetric videos while reducing the size of the representation itself [15].

In this paper, we present an end-to-end pipeline for volumetric video compression utilizing neural-based representation. We represent each frame of volumetric video as a NeRF, constructing a sequence of NeRFs. By representing volumetric video with neural networks, the problem of volumetric video compression becomes neural model compression.

Our work builds upon the key insight that there is significant similarity between successive NeRFs, which suggests the temporal redundancy in latent neural space. Figure 2(a) gives an example of temporal redundancy in explicit representation (*i.e.*, dynamic point clouds), where consecutive point cloud frames contain similar visual content. We find that such temporal redundancy still exists in latent neural space. For better illustration, we train ten NeRFs to represent ten consecutive point cloud frames and then measure their correlation. We present the similarity matrix of those neural representations in Figure 2(b). As shown, neighboring NeRFs share significantly high similarities, with over 0.98 correlation coefficient. Based on this observation, temporal compression is proposed for model compression. Specifically, instead of encoding each NeRF separately, we only encode the differences between consecutive NeRFs. This way, we achieve a significant reduction in bitrate without a noticeable loss of rendering quality. We apply an exponent-based non-uniform quantization scheme [5] to our temporal compression.

We propose an efficient, NeRFs-based representation for 3D dynamic scenes; the compression ratio benefits from the temporal coherence of the model. Here, we consider dynamic point cloud compression as a possible application scenario and thus compare the proposed method with state-of-the-art PCC codecs. We conduct extensive experiments on 8iVFBv2 and 8iVSLF Dataset [6, 13] with the original NeRF [18]. Experimental results demonstrate the superiority of our method compared with geometry-based PCC (*i.e.*, G-PCC and Draco). We also show that the proposed method can achieve comparable R-D performance w.r.t. V-PCC which is regarded as the state-of-the-art PCC codec, when dealing with high-bitrate volumetric videos.

The focus of our work lies in leveraging the high temporal coherence in neural models, independent of specific 3D representations and neural architectures. Specifically, our work is not limited to processing point clouds as the input source, but it can be applied to any dynamic content for which a sequence of images may be generated, and used as input of the NeRF sequence. Our approach inherently represents the scene using an implicit representation, thus allowing for the application of our NeRF model to a wide range of volumetric video-related tasks and scenarios. Meanwhile, it is worth highlighting that recent studies [4, 14] have shown that state-of-the-art NeRF variations, despite their improved rendering quality and speed, often sacrifice model size, hindering their suitability for streaming applications. Given our research's specific focus on the streaming context and the challenges associated with achieving a favorable rate-distortion trade-off, we chose to evaluate the original NeRF model as a baseline. Nonetheless, our methodology remains adaptable to incorporate other models, enabling further exploration of compression and trade-offs while considering specific application requirements.

2 NEURAL-BASED VOLUMETRIC VIDEO COMPRESSION

We represent the volumetric video with a sequence of NeRFs [18] and achieve volumetric video compression by compressing the neural representations themselves. Figure 3 shows the overall architecture of our framework. Our framework consists of two key components: model training and temporal compression, which are elaborated in Section 2.1 and Section 2.2, respectively.

The system design stems from the insight that neighboring NeRF models share considerable similarities. In the model training stage, each NeRF is initialized based on the previous frame's NeRF, except for the first frame's NeRF, which is trained from the beginning as the starting point. This training strategy not only achieves significant time and resource savings but also further encourages temporal redundancy between subsequent NeRFs. Then, the temporal compression idea is applied to model compression. That is, for

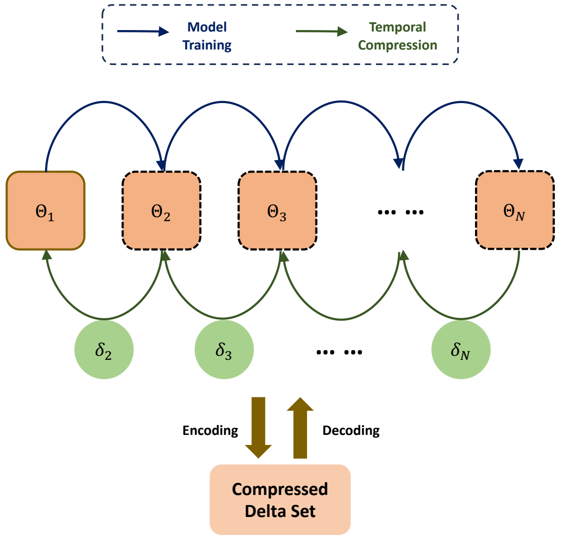


Figure 3: Overview of the proposed NeRF-based volumetric video compression.

a sequence of trained NeRF models, we regard the first model as a "key-frame" or "I-frame" [2] and compress only the deltas between successive models. The models can be restored by accumulating the decoded deltas to the reference model.

2.1 NeRF-based Representation

Given a volumetric video with N frames, we can generate M views for each frame i, and construct a multi-view image set

$$\mathbf{D}_i = \left\{ \left(V_m^i, X_m^i \right) \right\}_{m=1}^{M}, \tag{1}$$

where V_m^i is the camera pose and X_m^i is the corresponding image captured from this pose. A NeRF F_i is trained based on $\mathbf{D}_i$, and we simply use its weights $\mathbf{\Theta}_i$ to represent the model, to avoid cluttering the notation. Therefore, we represent a volumetric video with N frames as a NeRF sequence $\{\mathbf{\Theta}_i\}_{i=1}^{N}$.

We adopt a transfer-learning-like strategy for efficient model training. To be specific, each frame of the volumetric video is modeled by a NeRF initialized using the previous frame's NeRF. The NeRF representing the model at the starting time is trained from scratch and is then taken as a starting point for training subsequent times. By taking advantage of the learned knowledge from a previous model, considerable time and resources, which would have been required to train a model from scratch, can be saved. Meanwhile, such a training strategy further forces successive NeRFs to be closer and thus enhance the compression.

Formally, a NeRF $\mathbf{\Theta}_i$, which is trained to represent the i-th volumetric video frame, is initialized with the weights of its previous model $\mathbf{\Theta}_{i-1}$ and optimized by minimizing the distance from their renderings to the ground truth images:

$$\mathcal{L}_{\mathbf{\Theta}_i} = \sum_{m=1}^{M} \|\hat{X}_m^i - X_m^i\|_2^2, \tag{2}$$

where $\| \cdot \|_2^2$ is the Euclidean norm, $\hat{X}_m^i$ is the predicted image, and X_m^i is the ground truth.

2.2 Model Compression

In order to depict complex geometry and appearance, NeRF requires huge neural networks with billions of parameters, which poses a great challenge for the transmission with limited bandwidth. In this section, we introduce the proposed model compression techniques to considerably reduce the size of NeRF models while keeping good quality of the rendered images, to achieve good R-D performance.

Temporal Compression. To achieve efficient and scalable model compression, we propose to leverage high similarity between adjacent models and only encode and store the difference between them. Each model can be restored by applying the delta values to its previous model. Formally, we can represent a set of models $\{\mathbf{\Theta}_i\}_{i=1}^{N}$ as $\{\mathbf{\Theta}_1, \{\delta_i\}_{i=2}^{N}\}$, where $\mathbf{\Theta}_1$ is the first frame model and $\delta_i = \mathbf{\Theta}_i - \mathbf{\Theta}_{i-1}$ which is the delta values between $\mathbf{\Theta}_i$ and $\mathbf{\Theta}_{i-1}$. Our compression task is to compress the delta values while keeping good rendered image quality of the restored model $\hat{\mathbf{\Theta}}_i$, where $\hat{\mathbf{\Theta}}_i = \mathbf{\Theta}_1 + \sum_{t \leq i} \hat{\delta}_t$ and $\hat{\delta}_t$ is decoded from the compressed δ_t.

An efficient and scalable model compression scheme, called LC-Checkpoint [5], is adopted in our proposed compression pipeline. The compression pipeline consists of two components. First, *exponent-based quantization* and then *priority promotion* are performed for lossy compression. The core idea of exponent-based quantization comes from the representation of floating points. Specifically, a floating point v is represented by $v = (-1)^s \times m \times 2^e$, where s is the sign, m is the mantissa, and e is the exponent. Exponent-based quantization partitions the floating-point numbers in δ_i into multiple buckets, based on their exponent e and sign s. Consequently, the elements with the same exponent and sign will be assigned to the same bucket. Then, the elements in each bucket are represented by the average of maximum and minimum values in the bucket. The number of buckets can be further limited with a priority promotion approach by keeping $2^{N_b} - 1$ buckets with larger exponent e only, where N_b is the number of bits for bucket indexing. The rest buckets are merged into one bucket, which is represented by 0. By doing so, only N_b bits are required to index buckets. Secondly, the quantized values are further compressed using *Huffman coding* [26]. We can trace the performance of the compressed model at different bitrates by changing the number of bits N_b to plot the R-D curves.

Neural Architecture Search. The ability of 3D representation of NeRF is determined by its model architecture. Generally speaking, with a larger number of parameters, NeRF can represent more complex detailed scenes. Meanwhile, as reported by previous works [11, 19, 29], not all parameters are crucial for accurate rendering and one can significantly reduce the model size with a limited impact on performance by properly adapting the number of network layers.

Therefore, inspired by such observation, we explore the effect of model architecture on the rendering performance of NeRF, and analyze the R-D performance by tracking the change of model architecture. Specifically, we choose the network depth (number of layers) and the network width (number of neurons in each layer) as our neural architecture search space, which are set to $\{1, 2, 4, 6, 8\}$ and $\{64, 128, 256\}$, respectively. We first conducted experiments to

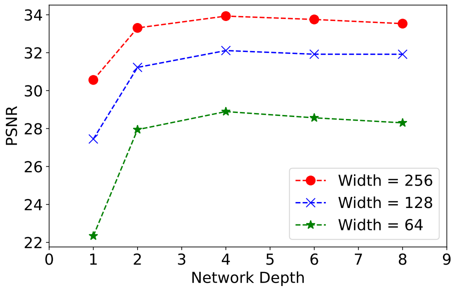


Figure 4: The performance of NeRFs with different neural architectures.

narrow down the search space. We trained NeRFs with different combinations of network depth and width on a volumetric video frame (*i.e.* RedAndBlack Frame #1450) and measured the quality of images rendered from NeRFs using the peak signal-to-noise ratio (PSNR), as shown in Figure 4. Our crucial observation from Figure 4 is that the rendering quality of NeRF monotonically decreases when we reduce the network width, while with the decrease of the network depth, the rendering quality first improves and then drops when the network depth is smaller than 2. These results can be explained by previous works [9, 24, 30], which suggest that network depth provides the model with the ability to learn hierarchical representations, while width provides the model with the capacity to "memorize" the training data. Hence, we narrow down the search space of network depth and network width to $\{2, 4\}$ and $\{128, 256\}$, respectively, for the following evaluation.

3 EVALUATION

3.1 Experimental Settings

Dataset. We use four dynamic point cloud sequences for evaluation: RedAndBlack, Loot, Soldier, and Thaidancer. The first three sequences are from the 8iVFBv2 Dataset [6], and Thaidancer is from the 8iVSLF Dataset [13] which has a much greater number of points and higher bitrate. We select the first 30 frames of each sequence for the experiment. The average number of points per frame and corresponding bitrates (in Gbps) of the uncompressed volumetric videos are summarized in Table 1.

Table 1: Dynamic 3D Point Cloud Dataset

	RedAndBlack	Loot	Soldier	Thaidancer
Points ($\times 10^6$)	0.7	0.8	1.1	3.1
Bitrate (Gbps)	3.6	3.9	5.5	20.7

Rendering and Evaluation Parameters. Open3D [31] version 0.15.1 [1] is used for 2D rendering, where the width and height of the rendered images are 600 and 600, respectively. For NeRF training and testing, we generate 100 views as the training set and 200 views as the testing set for each frame of each point cloud sequence, where the camera settings are the same as [18]. Similarly, to evaluate the performance of other PCC codecs, we generate 200 rendered images for each decoded point cloud using the same camera settings of the testing set. PCC Arena [28] is used to measure the 2D quality of volumetric videos. PSNR and structural similarity index (SSIM) are used to quantify the 2D quality of rendered images. We calculate the average quality among the testing set for every frame.

Comparison Methods. Three PCC codecs are introduced:

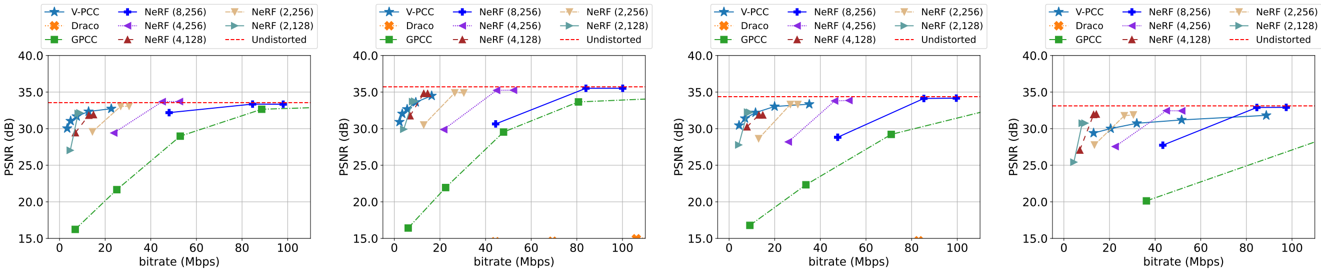
i. *V-PCC* stores point cloud frames into 2D video frames and passes the 2D videos to 2D video codecs for compression. As defined in the V-PCC common test condition (CTC) [7], five compression rates controlled by the geometry and texture quantization parameter are used to generate the R-D curve.
ii. *G-PCC* utilizes an octree [17] or spatial data structures and applies arithmetical encoding to attributes. G-PCC quantizes the coordinates from floating-point numbers to integers with the parameter *positionQuantizationScale*. As defined in PCC Arena [28], eight compression rates, which are controlled by the quantization parameter, are used in our experiment.
iii. *Draco* adopts the K-D tree [1] data structure to compress point clouds. It employs quantization to reduce the number of bits, controlled by the quantization bit and compression level. The quantization bit determines the level of precision for the data. The compression level strikes a balance between the rate of compression and the computational complexity involved. According to PCC Arena [28], we use eight compression rates to track the R-D performance.

NeRF Settings. We keep the same settings for the NeRF model as [18], except for the model architecture. The Adam optimizer [12] is used for optimization. The learning rate begins at 5×10^{-4} and decays exponentially to 5×10^{-5}. ReLU [20] is used as the activation function. The first NeRF corresponding to the first frame is trained from scratch, with 300k iterations. The following NeRFs are initialized with the previous NeRF, and we find the optimization typically only takes 20k to converge.

As discussed in Section 2.2, we compressed the NeRF-based representation by simplifying the neural architecture to further improve its performance. Hence, we train five NeRF sequences for every dynamic point cloud sequence by changing the model architecture. Specifically, according to the experiments, we narrowed down the search space of network depth d and width w to $\{2, 4\}$ and $\{128, 256\}$ so that four model architectures are considered. We additionally train the NeRF with default model architecture ($d = 8$, $w = 256$) for sanity check. We denote these five model settings as $NeRF(8, 256)$, $NeRF(4, 256)$, $NeRF(4, 128)$, $NeRF(2, 256)$, and $NeRF(2, 128)$.

Encoder Settings. As mentioned in Section 2.2, two trade-off parameters are utilized to balance between bitrate and distortion in the proposed method: (i) number of bits N_b, which determines the quantization level for temporal compression, and (ii) model architecture which controls the size of neural networks. Based on our observations, we found that the rendered quality of the compressed NeRF remained stable when the number of bits N_b exceeded 5, while dropping significantly below 3. This observation is reasonable since using less than $2^2 - 1$ buckets to store the weights can result in substantial distortion. Therefore, we limited the number

[1] https://github.com/isl-org/Open3D/releases/tag/v0.15.1



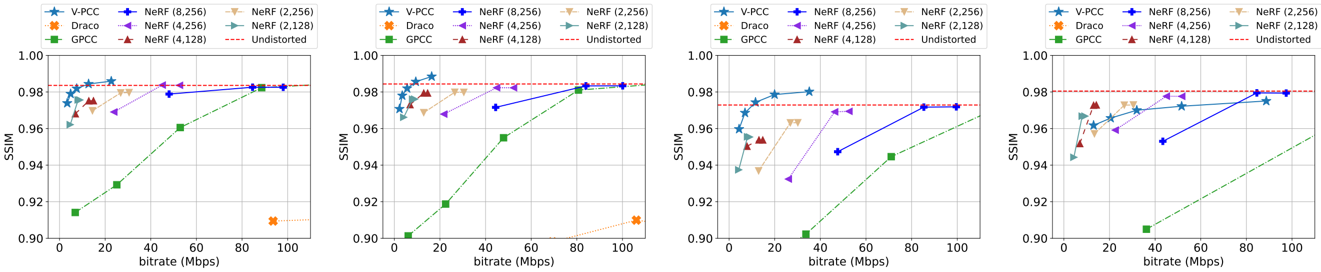
(a) RedAndBlack.

(b) Loot.

(c) Soldier.

(d) Thaidancer.

Figure 5: R-D curves for PSNR: (a) RedAndBlack, (b) Loot, (c) Soldier, and (d) Thaidancer.



(a) RedAndBlack.

(b) Loot.

(c) Soldier.

(d) Thaidancer.

Figure 6: R-D curves for SSIM: (a) RedAndBlack, (b) Loot, (c) Soldier, and (d) Thaidancer.

of bits to a range of 3 to 5 to ensure a reasonable trade-off between bitrate and distortion. For five NeRF architectures, we perform temporal compression with different number of bits to trace the R-D performance, which gives us a total of five R-D curves for each point cloud sequence.

3.2 Experimental Results

We report the R-D curves in Figure 5 and Figure 6 showing how the quality of the four point cloud sequences changes w.r.t. the encoded bit-rate. To have better insights on how much quality is lost during the quantization process, we also plot the quality of rendered images from $NeRF(8, 256)$ without compression and denote it as Undistorted, which serves as the upper bound of visual quality. These figures specifically focus on bit-rates below 100 Mbps, so not all the R-D curves are displayed. By truncating the curves, we can have a clearer visualization of the R-D performance of the proposed method and thus make a better comparison with other PCC codecs. Notably, most of the points in R-D curves for Draco are not within the shown bit-rate range due to its significantly lower compression ratios. This indicates that Draco falls behind another geometry-based PCC, *i.e.* G-PCC, in terms of R-D performance. Consequently, we have chosen not to include Draco in the quantitative comparison to focus on the codecs that are more relevant and competitive in the displayed bit-rate range.

As observed in Figure 5 and Figure 6, the proposed method clearly outperforms G-PCC, always achieving better quality at the same bitrate for all the point cloud sequences. Specifically, our method outperforms G-PCC by at most 15.92 dB in PSNR and 6.15% in SSIM on RedAndBlack, 17.33 dB in PSNR and 7.48% in SSIM on Loot, 15.45 dB in PSNR and 7.41% in SSIM on Soldier, and 11.76 dB in PSNR and 7.94% in SSIM on Thaidancer. Furthermore, the R-D curves of our method demonstrate that, by utilizing the proposed temporal compression technique, the NeRF size can be effectively reduced with only a minor increase in rendering distortion. This finding highlights the effectiveness of our method in achieving substantial compression gains while maintaining acceptable quality.

Particularly, in the case of high-bitrate sequences, *i.e.* Thaidancer with a bitrate exceeding 20 Gbps, our method achieves better R-D performance compared to V-PCC, with up to 2.59 dB improvement in PSNR and 1.11% improvement in SSIM. The advantages of neural-based representation in this context are well-founded. Neural networks empower NeRF with the capability to capture fine-grained details by learning a compact implicit representation. Unlike explicit representations like point clouds, the size of the implicit representation in NeRF is determined by the neural networks and is not directly proportional to the complexity of geometry and attribute of the volumetric video. This decoupling allows for more efficient storage and transmission of the video data. In contrast, as an explicit representation, point cloud requires larger data size and potentially higher bitrate requirements.

We also employ the Soldier sequence as an example, presenting the compression ratio vs. SSIM in Figure 7. The compression ratio denotes the ratio of the compressed model's bitrate to that of the uncompressed baseline model ($NeRF(8, 256)$, labeled as Undistorted).

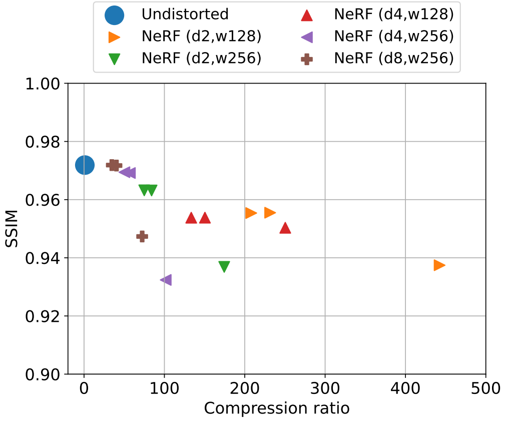


Figure 7: Compression ratio vs. SSIM for Soldier. Each NeRF architecture has three points corresponding to three different N_b: 3, 4, and 5.

Notably, our method achieves a wide range of compression ratios, spanning from 35 to 442, while exhibiting minimal degradation in quality, with quality drop ranging from 0.1% to 3.6% in SSIM. These compelling results underscore the exceptional performance of our proposed method in effectively balancing efficient compression and preservation of visual quality.

Besides the quantitative analysis above, we also show the sample rendered images of RedAndBlack and Thaidancer using compressed NeRF models under different bitrates in Figure 8, for qualitative analysis. The figure shows a view of Frame #1454 of RedAndBlack and Frame #6487 of Thaidancer. As can be found, the proposed method well restores the details of cloth texture. Moreover, the visual quality of the renderings remains relatively intact even when compressing the NeRF model from 84 Mbps to 26 Mbps, which suggests that our method can provide similar levels of detail and visual quality even at low bitrate.

In summary, the objective results reported in Figures 5, 6, and 7, and the sample rendered images from compressed NeRFs shown in Figure 8 demonstrate that our proposed method achieves high compression for volumetric video with minimal loss of detail.

4 CONCLUSION AND DISCUSSION

In this paper, we introduce an extendable and general pipeline for compressing volumetric video using a neural-based representation, which leverages the similarities between consecutive NeRFs and exploits temporal coherence and neural architecture to achieve effective and efficient model compression. We primarily tested our method on point cloud compression. Through experimental evaluations, we demonstrate the superiority of our method compared to geometry-based PCC codecs. Moreover, our approach achieves comparable results with state-of-the-art PCC codecs for high-bitrate volumetric videos. However, it is important to note that our proposed approach has wider applicability and is not restricted to point clouds as the sole input source. The inherent nature of our method, which models the scene using an implicit representation, enables its usage with any volumetric video data. The advantage of NeRF representation of offering a joint, realistic geometry and appearance model holds for our proposed solution. Therefore, our NeRF-based compression framework applies to any dynamic 3D content that may be rendered, expanding its compression performance to potential applications including a diverse range of tasks and scenarios.

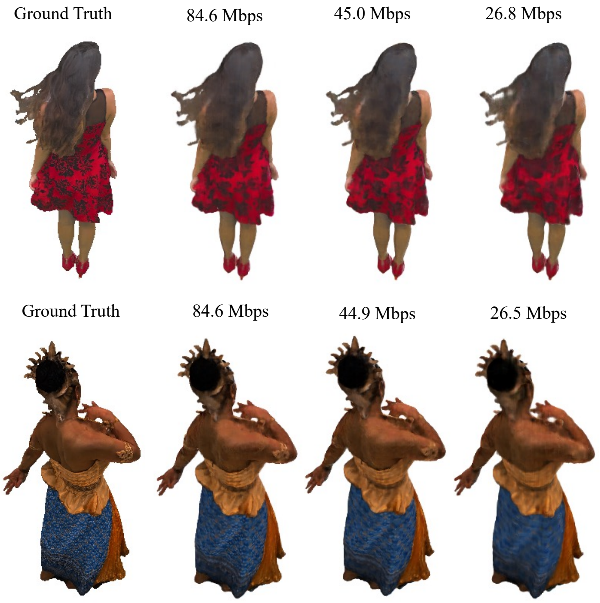


Figure 8: Sample images of RedAndBlack and Thaidancer using the compressed NeRF models at different bitrates.

There are several directions for future research based on the limitations and opportunities identified in our work. Firstly, although the PSNR and SSIM results offer valuable insights into the visual quality, conducting user studies would provide a more comprehensive understanding of the effectiveness of our method. Secondly, our current approach models volumetric videos frame by frame, which essentially represents the scene as a set of static NeRFs. Recent research efforts [21, 23, 25] have extended static NeRF to dynamic NeRF, enabling the representation of dynamic scenes with a single model. However, in the context of NeRF-based volumetric video streaming, modeling a dynamic scene with one model can make rate and viewport adaptation impractical [15]. One potential solution could be to split the video into several equal-size groups of frames (GOF) and train dynamic NeRFs for each group. Then, exploring the temporal redundancy among consecutive dynamic NeRFs and the relationship between the size of GOF and the level of temporal redundancy would be an interesting avenue for future research.

ACKNOWLEDGMENTS

This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (T1 251RES2038). The authors kindly thank Axel Carlier for his initial inspirational idea.

REFERENCES

[1] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept 1975.

[2] Vasudev Bhaskaran and Konstantinos Konstantinides. *Image and video compression standards: algorithms and architectures*. Springer Science & Business Media, 1997.

[3] Thomas Bird, Johannes Ballé, Saurabh Singh, and Philip A. Chou. 3D scene compression through entropy penalized neural representation functions. In *Proceedings of the 2021 Picture Coding Symposium, PCS 2021, Bristol, United Kingdom, June 29 - July 2, 2021*, pages 1–5. IEEE, 2021.

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *Proceesdings of the 17th European Conference on Computer Vision, ECCV 2022, Tel Aviv, Israel, October 23-27, 2022,*, volume 13692, pages 333–350. Springer, 2022.

[5] Yu Chen, Zhenming Liu, Bin Ren, and Xin Jin. On efficient constructions of checkpoints. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 1627–1636. PMLR, 2020.

[6] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 8i voxelized full bodies-a voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006*, 7(8):11, 2017.

[7] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. Common test conditions for point cloud compression. *ISO/IEC JTC1/SC29/WG11 w17766*, 2018.

[8] Google. Draco 3D data compression. https://github.com/google/draco. Accessed: 2023-11-28.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, , Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[10] Yueyu Hu and Yao Wang. Learning neural volumetric field for point cloud geometry compression. In *Proceedings of the 2022 Picture Coding Symposium, PCS 2022, San Jose, CA, USA, December 7-9, 2022*, pages 127–131. IEEE, 2022.

[11] Yongdong Huang, Yuanzhan Li, Xulong Cao, Siyu Zhang, Shen Cai, Ting Lu, Jie Wang, and Yuqi Liu. An efficient end-to-end 3D voxel reconstruction based on neural architecture search. In *Proceedings of the 26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 3801–3807. IEEE, 2022.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Maja Krivokuća, Philip A Chou, and Patrick Savill. 8i voxelized surface light field (8ivslf) dataset. *ISO/IEC JTC1/SC29 WG11 (MPEG) input document m42914*, 2018.

[14] Junhua Liu, Yuanyuan Wang, Yan Wang, Yufeng Wang, Shuguang Cui, and Fangxin Wang. Mobile volumetric video streaming system through implicit neural representation. In *Proceedings of the 2023 Workshop on Emerging Multimedia Systems, EMS 2023, New York, NY, USA, 10 September 2023*, pages 1–7. ACM, 2023.

[15] Kaiyan Liu, Ruizhi Cheng, Nan Wu, and Bo Han. Toward next-generation volumetric video streaming with neural-based content representations. In *Proceedings of the 1st ACM Workshop on Mobile Immersive Computing, Networking, and Systems, ImmerCom 2023, Madrid, Spain, 6 October 2023*, pages 199–207. ACM, 2023.

[16] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics*, 40(4):59:1–59:13, Jul 2021.

[17] Donald Meagher. Geometric modeling using octree encoding. *Computer Graphics and Image Processing*, 19(2):129–147, Jun 1982.

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, Dec 2021.

[19] Saeejith Nair, Yuhao Chen, Mohammad Javad Shafiee, and Alexander Wong. NAS-NeRF: Generative neural architecture search for neural radiance fields. *CoRR*, abs/2309.14293, 2023.

[20] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning, ICML 2010, June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress, 2010.

[21] Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Representing volumetric videos as dynamic MLP maps. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4252–4262. IEEE, 2023.

[22] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo César, Philip A. Chou, Robert A. Cohen, Maja Krivokuca, Sebastien Lasserre, Zhu Li, Joan Llach, Khaled Mammou, Rufael Mekuria, Ohji Nakagami, Ernestasia Siahaan, Ali J. Tabatabai, Alexis M. Tourapis, and Vladyslav Zakharchenko. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, Mar 2019.

[23] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4D: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16632–16642. IEEE, 2023.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. NeRFPlayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, May 2023.

[26] Jan van Leeuwen. On the construction of huffman trees. In *Proceedings of the 3rd International Colloquium on Automata, Languages and Programming, ICALP 1976, University of Edinburgh, UK, July 20-23, 1976*, pages 382–410. Edinburgh University Press, 1976.

[27] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9762–9772. IEEE, 2021.

[28] Cheng-Hao Wu, Chih-Fan Hsu, Tzu-Kuan Hung, Carsten Griwodz, Wei Tsang Ooi, and Cheng-Hsin Hsu. Quantitative comparison of point cloud compression algorithms with PCC arena. *IEEE Transactions on Multimedia*, 25:3073–3088, Feb 2023.

[29] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-NeRF: An efficient and dynamically growing nerf. *IEEE Transactions on Visualization and Computer Graphics*, 29(12):5124–5136, Dec 2023.

[30] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the 2016 British Machine Vision Conference, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[31] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing, 2018.

A Comparative Study of K-Planes vs. V-PCC for 6-DoF Volumetric Video Representation

Na Li
Rutgers University
Piscataway, NJ, USA
na.li@rutgers.edu

Mufeng Zhu
Rutgers University
Piscataway, NJ, USA
mz526@rutgers.edu

Shuoqian Wang
SUNY Binghamton
Binghamton, NY, USA
swang130@binghamton.edu

Yao Liu
Rutgers University
Piscataway, NJ, USA
yao.liu@rutgers.edu

ABSTRACT

With NeRF, neural scene representations have gained increased popularity in recent years. To date, many models have been designed to represent dynamic scenes that can be explored in 6 degrees-of-freedom (6-DoF) in immersive applications such as virtual reality (VR), augmented reality (AR), and mixed reality (MR). In this paper, we aim to evaluate how newer neural representations of 6-DoF video compare with more-traditional point cloud-based representations in terms of their representation and transmission efficiency. We design a new methodology for fair comparison between `K-Planes`, a new dynamic neural scene representation model, and video-based point cloud compression (`V-PCC`). We conduct extensive experiments using three datasets with a total of 11 sequences with different characteristics. Results show that the current `K-Planes` models excel for moderately dynamic content, but struggle with highly dynamic scenes. In addition, in emulated volumetric data capture scenarios, the recorded point cloud data can be highly noisy, and the visual quality of views rendered by trained `K-Planes` models are significantly better than `V-PCC`.

CCS CONCEPTS

• **Information systems → Multimedia streaming**; • **Computing methodologies → Point-based models**; **Volumetric models**.

KEYWORDS

6-DoF, volumetric videos, point cloud, neural scene representations

ACM Reference Format:
Na Li, Mufeng Zhu, Shuoqian Wang, and Yao Liu. 2024. A Comparative Study of K-Planes vs. V-PCC for 6-DoF Volumetric Video Representation. In *16th International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652212.3652227

1 INTRODUCTION

In recent years, both the ubiquity and sophistication of devices for video collection have grown. Concurrently, the capabilities of neural network models for fusing information from multiple video signals have seen substantial growth. These two general developments set the stage for more-immersive multimedia streaming applications aimed at enhancing user experiences. Among immersive applications, 6-DoF volumetric video, which captures a real-world scene from a multitude of perspectives over time, enables the greatest level of immersion. Traditional 6-DoF representations include 3D triangular meshes with texture and point clouds. These representations rely on more-direct storage and rendering of 3D scenes. On the other hand, neural representations are often implicit representations of a scene: the stored data powering the representation is often not interpretable by humans. For example, NeRF [23] uses a single multi-layer perceptron (MLP) for representing a scene. Both generating such representations by training from collected imagery and rendering these representations can require significant computational resources.

Although many lines of research have explored 6-DoF representations for individual scenes, addressing the additional temporal dimension in 6-DoF videos presents more challenges. Traditional 6-DoF representations can be adapted for video transmissions by transmitting standard video-encoded streams of RGB-Depth data, as demonstrated in prior work such as [17]. Alternatively, representation-specific codecs [22, 27, 29] such as video-based point cloud compression (`V-PCC`) [26, 28] and Draco [9, 18] can be employed. However, neural representations for 6-DoF video are less-well explored. These representations must simultaneously capture temporal and spatial characteristics and also allow for space-efficient network transmission and compute-efficient rendering.

In this paper, we set to evaluate how newer neural representations of 6-DoF videos compare with more-traditional point cloud-based representations in terms of their representation and transmission efficiency. For our comparison, we use the `K-Planes` model [14] as the state-of-the-art approach for future neural 6-DoF video representation that can be efficiently transmitted. This model strikes a balance between space and computational efficiency, using a representation that factors over both space and time dimensions coupled with a small neural network. For traditional point-cloud-based representations, we select `V-PCC` as it is an emerging standard for compressing dynamic point cloud data.

MMVE '24, April 15–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0618-9/24/04…$15.00
https://doi.org/10.1145/3652212.3652227

To perform fair comparison, we design a new methodology including the generation of training data for K-Planes and testing data for both K-Planes and V-PCC as well as the implementation of experiment procedures. Our study uses three different datasets of dynamic 6-DoF scenes. Among them, two are derived from existing datasets, while the third has been created by our team. We have conducted extensive experiments across these three datasets with 11 dynamic sequences with different characteristics. To the best of our knowledge, we are both the first to propose such a comparison methodology and the first to present results from such a comparison study of 6-DoF video representations. The configuration files used for K-Planes training in our experiments along with the trained models are available at: https://github.com/symmru/MMVE-2024.

Results show that for dynamic 6-DoF content with little to moderate motion, using K-Planes models for representation can save the storage size and improve visual quality of rendered views compared to using V-PCC -based encoding. However, the current K-Planes models cannot represent highly dynamic content very well. Moreover, in a emulated real-world scenario where point cloud data is derived from recorded RGB and depth information, we find that the derived point cloud data is very noisy. This confirms the insights from previous studies, e.g., [19, 20]. The visual quality of V-PCC suffers significantly. On the other hand, neural-based solution K-Planes performs substantially better compared to V-PCC in such emulated scene capture scenario.

2 BACKGROUND AND RELATED WORK

Traditional 6-DoF representations. Volumetric videos capture frame sequences in a 3D space, allowing users to view in 6 degree-of-freedom (6-DoF): from arbitrary positions, (x, y, z), in 3D space and arbitrary orientations, (ϕ, θ, ρ). 6-DoF content is widely employed in today's computer gaming and virtual reality platforms. In these platforms, objects and scenes are represented as synthetic models using 3D triangular meshes with texture information that describes how faces of the mesh should appear. Besides trianglular meshes, another volumetric video representation, **point clouds**, has received increased interests in recent years. Point clouds associate color information with 3D pixel/point positions. They can be captured from real-world scenes using RGB-Depth cameras. Typical point cloud scenes contain millions of points and are infeasible to store in raw formats. Point cloud compression (PCC) is currently under active development under Moving Picture Experts Group (MPEG). Among the efforts, video-based point cloud compression (**V-PCC**) [15, 26] aims to leverage existing 2D video codecs for compressing dense point cloud data.

NeRF-based neural representations. Neural radiance field (NeRF) is an emerging representation of 3D scenes. It uses the volume rendering technique for rendering color of pixels on an image. To render a view of a scene, rays are traced from the camera origin through each pixel in the rendered image. The original NeRF [23] proposes to use a simple multi-layer perceptron (MLP) to estimate the volume density σ_i and color $\mathbf{c}_i$ of sample i on a ray as a function of its position $\mathbf{x}_i$ and direction of the ray $\mathbf{d}_i$. The training time of the original NeRF is known to be very long. The authors described in their paper that a typical training can take 1 to 2 days on a Nvidia V100 GPU.

TensoRF [11] is a more recent work that represents the radiance field as a 4D tensor. The main idea of TensoRF is to use tensor decomposition to represent the 4D tensor as the sum of vector-matrix outer products. Compared to the original NeRF, TensoRF models can be trained substantially faster (more than 100x improvement) and with better rendering quality.

Neural representations for dynamic 6-DoF content. For modeling dynamic scenes, many NeRF-variants exist, e.g., D-NeRF [24] and DyNeRF [21]. Among them, K-Planes [14] is a novel approach that represents dynamic volumetric content as a 4D volume (as opposed to the static 3D volume). K-Planes factorizes a 4D volume into 6 planes: 3 space-only planes and 3 space-time planes. Given $q = (i, j, k, t)$ on the 4D volume, it is projected onto each of the 6 planes and bilinearly interpolated to obtain 6 feature vectors. Features from all 6 planes are combined using the Hadamard product (elementwise multiplication). Additionally, K-Planes uses multi-scale planes with different resolutions. Features obtained from different scales $s \in S$ are concatenated. To determine the density and color, it uses two MLPs. The first MLP $\mathcal{F}_\sigma$ is for mapping the feature into volume density σ and an additional feature $\hat{f}(q)$. The second MLP $\mathcal{F}_c$ estimates the color using the additional feature $\hat{f}(q)$ and input of ray direction $\mathbf{d}$.

3 METHODOLOGY

To compare the performance of V-PCC and K-Planes for 6-DoF video representation, we use three datasets with a total of 11 dynamic volumetric sequences. We next describe details of our methodology for conducting this comparative analysis, including the generation of datasets generation as well as the selection of metrics used for comparison.

3.1 Datasets

We use three datasets in this study. The front-facing images of all 11 sequences in the three datasets are shown in Figure 1. The 8iVFB dataset [13] consists of four dynamic point cloud sequences, Longdress , Loot, Soldier, and Redandblack as shown in Figure 1(a)-(d). It is a voxelized full body dataset, with the spatial resolution of each sequence being 1024x1024x1024. The vsenseVVDB2 dataset [31] also includes four dynamic point cloud sequences, AxeGuy, LubnaFriends, Rafa2, and Matis. Similar to 8iVFB , the spatial resolution of sequences in vsenseVVDB2 is also 1024x1024x1024.

Both 8iVFB and vsenseVVDB2 are datasets created for evaluating the performance of V-PCC . They only contain raw points data for each sequence. To use these datasets for comparable K-Planes evaluation, we must generate camera views from different perspectives with known camera extrinsic parameters. Unfortunately, neither of these datasets provide the raw camera-captured video frame data. To obtain data for training the K-Planes models, we use Blender 3.5.1 [5] to render the raw point clouds and use them as the groundtruth data. We describe how we generate training data for K-Planes in Section 3.1.1.

The third dataset Blender is created by our team. It includes three animated Blender 3D models, Lego [23], Pig downloaded from Blender Market [7], Amily downloaded from Blender Demo Files [6]. For the "Lego" model, we created animation raising and lowering the bulldozer's bucket by moving the control panel built

(a) Longdress (b) Loot (c) Soldier (d) Red&Black (e) AxeGuy (f) Lubna (g) Rafa2 (h) Matis (i) Lego (j) Pig (k) Amily

Figure 1: Front-facing images all 11 testing traces. (a)-(d): `8iVFB` dataset; (e)-(h): `vsenseVVDB2` dataset; (i)-(k): `Blender` dataset. All traces contain dynamic sequences that can be explored in 6-DoF.

in original `.blend` file. For "Pig" and "Amily" models, we used the animations in the downloaded `.blend` file. Since `V-PCC` only takes point cloud data `.ply` as input, for fair comparisons, we also need to convert these three models in the `Blender` dataset into point clouds. We emulate the process of capturing real-world point clouds via RGB-Depth cameras within the Blender environment. We must also note, however, that despite accurate camera intrinsics and extrinsics data provided by Blender, due to other factors such as depth data quantization, the recorded depth data is inherently noisy, as with real-world LiDAR sensors [19, 20].

3.1.1 `K-Planes` *Training Data Generation.* For all three datasets, we place the center of the model (`.blend` model or the imported point cloud) at the origin (0,0,0). We follow the original NeRF work [23] and generate `K-Planes` training data by placing virtual cameras in the scene at 80 different positions, starting at position (0,4.0,0.5), which is approximately 4.03 units away from the origin. All 80 positions are obtained by rotation around the origin by a set of randomly generated Euler angles. At any position, the virtual camera is set to "look at" the origin. Since 9 out of 11 sequences are persons, we limit the camera positions to the upper hemisphere only. Note that with this setup, regardless of the camera positions, the distance from the origin is not changed. For each frame in a dynamic sequence, we render 80 views as recorded by 80 virtual cameras with a resolution of 800x800 and save them as `.png` files.

3.1.2 Testing Data Generation. For comparing the visual quality of rendered views, we generate ground truth testing data in a similar way as `K-Planes` training data generation. For the `8iVFB` and `Blender` datasets, we use 20 views from 20 differently-positioned cameras for testing. For the `vsenseVVDB2` dataset, since we use all 300 frames of the dataset, we use views from 10 different perspectives for evaluation. Besides, we set the same random seed for each dataset to make sure all consecutive frames of each model have the same camera parameters.

3.2 Comparison Metrics

To characterize the performance of different codecs, the video compression community commonly uses the rate-distortion (RD) curve e.g., [16, 30]. Here, "rate" represents the bitrate of the encoded media content. "Distortion" represents the visual quality of the compressed representation compared to the ground truth, uncompressed, representation. In this work, we focus on two distortion metrics: peak signal-to-noise ratio (PSNR) and video multi-method assessment fusion (VMAF) proposed by Netflix [1].

To plot the RD-curve for `V-PCC` , we use five different `qp` combinations described in common test conditions (CTC) by MPEG [25]. Details of the five settings are shown in Table 1. Among the five configurations, `r1` and `r5` result in the lowest and highest bitrates, respectively.

Table 1: `qp` combinations used in `V-PCC` common test conditions (CTC) [25]

`qp` settings	r1	r2	r3	r4	r5
Q_g: `qp` for geometry map	32	28	24	20	16
Q_c: `qp` for attribute (color) map	42	37	32	27	22

Table 2: `K-Planes` overall settings

Multi-scale	S=1,2; S=1,2,4; S=1,2,4,8
Time dimension	30; 60; 75
Feature length	F=4; F=8; F=16; F=32

`K-Planes` uses multi-scale planes with different resolutions for storing parameters. Following the setup in the `K-Planes` paper [14], we consider four spatial scale settings, {1, 2, 4, 8}. With different spatial scale settings, the resolution of the feature plane differs. For example, with $S = 1$, each spatial feature plane has the resolution of 64×64, and the scene contains 64^3 voxels. With $S = 8$, the resolution of the spatial feature plane is 512 × 512. To inference the density and color of a sample on a ray, features obtained from multi-scale planes are concatenated before being passed into the MLPs. In our experiments, we consider 3 different multi-scale settings, as shown in Table 2. In addition, we consider the impact of setting the time dimension to different values for representing the 4D volume. For feature vector at a plane position, we consider four different feature length settings: 4, 8, 16, and 32.

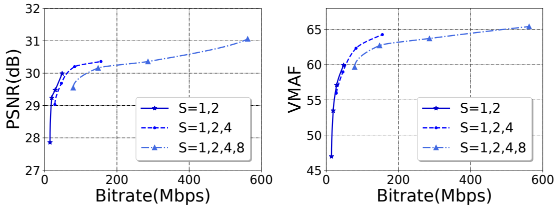
The RD-curve allows us to calculate the average difference in bitrates among different encoding mechanisms under the same distortion. This metric is called the Bjøntegaard-Delta bitrate (BD-rate) [3, 4, 10]. A negative BD-rate represent bitrate/bandwidth savings while achieving the same visual quality and is thus considered better. Similarly, a BD-PSNR metric can be calculated, where a positive number represents the improvement in PSNR while using the same bitrate/bandwidth. We report numerical results of the following metrics: BD-PSNR, BD-Rate$_p$ calculated using PSNR as the visual quality metric; BD-VMAF, and BD-Rate$_v$ calculated using VMAF.

4 K-PLANES RESULTS

In this section, we first characterize the performance of K-Planes for dynamic 6-DoF video representation under different model configurations. Specifically, we compare three multi-scale settings as listed in Table 2 and two time dimension settings.

4.1 Multi-Scale Settings

Figure 2 shows the the RD-curve results, using PSNR and VMAF as the visual quality metric, for the "Lego" sequence in the `Blender`



(a) PSNR

(b) VMAF

Figure 2: RD-curve result for the “Lego” sequence when using different numbers of spatial plane scales.

Table 3: BD-Rate⇓, BD-PSNR⇑, BD-VMAF⇑ results on scale performance, using S=1,2 as the anchor for calculation.

Settings	BD-Rate$_p$⇓	BD-PSNR⇑	BD-Rate$_v$⇓	BD-VMAF⇑
S=1,2	0%	0	0%	0
S=1,2,4	37.6%	-0.29	15.7%	-0.94
S=1,2,4,8	122.4%	-1.09	68.1%	-2.32

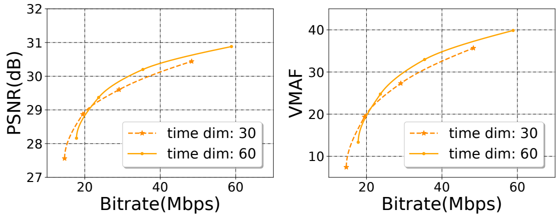
dataset. This sequence has 60 frames. We set the time dimension of the model to 30 (half of the number of frames as used in K-Planes [14]). The figure shows three curves, representing the RD-curve for multi-scale settings S=1,2; S=1,2,4; S=1,2,4,8, respectively. For each curve, we vary the “rate” by using different feature length $F \in \{4, 8, 16, 32\}$ for the model. Overall, 12 K-Planes models are trained, each using 40 training videos with a learning rate of 0.001. Following the configuration file used for K-Planes dynamic scene training [14], we set the number of the training epochs to 120, 000. For each K-Planes model, 20 testing videos are used for visual quality evaluation. We calculate the “rate” by considering the frame rate of the sequence to be 30 frames-per-second (fps). That is, 2 seconds to playback 60 frames in the sequence.

The RD-curves confirm that with larger feature length (thus more parameters), the visual quality consistently improves. We further analyzed the BD-Rate, BD-PSNR, and BD-VMAF results in Table 3. Note here that BD-Rate$_p$ is calculated using PSNR as visual quality, while BD-Rate$_v$ is calculated using VMAF. We use S=1,2, the smallest multi-scale setting as the anchor setting for calculating BD-* results. Results show that neither S=1,2,4 nor S=1,2,4,8 can outperform the smallest multi-scale setting, S=1,2. Their BD-Rates with respect to S=1,2 are positive, indicating more bitrates are needed to reach the same visual quality; and their BD-PSNR and BD-VMAF results are negative, indicating worse visual quality under the same bitrates. Based on these findings, we focus the remaining experiments of K-Planes on the S=1,2 multi-scale setting.

4.2 Time Dimension Settings

For representing a dynamic 3D scene, K-Planes includes six planes, three space-only planes and three space-time planes. While the space dimension is determined by the multi-scale settings, the time dimension is typically set to half of the number of frames in the dynamic scene [14]. We set to examine if by using larger time dimension setting (and thus larger space-time planes) can help to further improve the visual quality of highly dynamic scenes.

For this evaluation, we use the “Longdress” sequence from the 8iVFB dataset. We use the first 60 frames from this sequence, and we use two different time dimension settings: 30 and 60.



(a) PSNR

(b) VMAF

Figure 3: RD-curve result for the “Longdress” sequence when using different time dimension.

Table 4: BD-Rate⇓, BD-PSNR⇑, BD-VMAF⇑ results on time dimension, using time dimension=30 as the anchor. Results show that the improvement is limited.

Settings	BD-Rate$_p$⇓	BD-PSNR⇑	BD-Rate$_v$⇓	BD-VMAF⇑
time_dim= 30	0%	0	0%	0
time_dim= 60	-4.0%	0.12	-4.4%	1.18

The RD-curve results for PSNR and VMAF are shown in Figure 3. The “rate” in the figure is varied by using different feature lengths $F \in \{4, 8, 16, 32\}$. The figure shows that these two different time dimension settings are comparable when compressing the “Longdress” sequence. At lower bitrates, i.e., shorter feature lengths, setting the time dimension to 30 gives better results; while at higher bitrates, using larger time dimension helps. However, the improvement is very limited. Table 4 shows the BD-Rate, BD-PSNR, and BD-VMAF results. These results are obtained using time dimension 30 as the anchor. Results show that by using the longer time dimension, the bitrate can be reduced by approximately 4% while achieving the same visual quality, and that the PSNR can improve by 0.12 dB with the same bitrate.

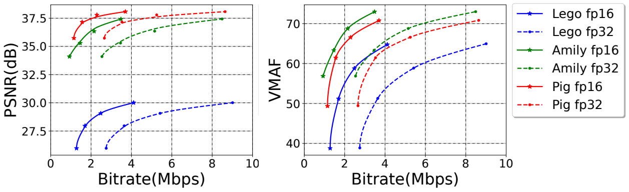
Given that “Longdress” is among the sequences with the most motion in our datasets, and that the improvement by using longer time dimension is very limited, we choose to use shorter time dimension (e.g., half of the number of frames) in the remaining experiments.

4.3 Model Precision

Mixed precision. PyTorch provides an automatic mixed precision package called TORCH.AMP [8]. It allows operations to use a mixture of float32 and float16 precision. This allows us to explore the feasibility of compacting the model by saving the model in float16 and load the parameters later for inference. We note that K-Planes also use float16 to speed up model training.

Further model compression. We explore lossless data compression via the zip tool [2] for further reducing the stored K-Planes representation size. While the trained models vary for different parameter settings and content, we observe that lossless data compression can further reduce the saved model size by 25% to 54%.

We conduct the experiment using all 11 sequences from all three datasets. For 8iVFB and Blender datasets with 60 frames, we set the time dimension to 30. For vsenseVVDB2 with 300 frames, we set the time dimension to 75. The trained models are further losslessly compressed via zip, and we use the .zip file size for calculating the “rate” for the rate-distortion curve. Figures 4 shows the RD-curve



(a) PSNR (b) VMAF

Figure 4: RD-curve result for the `Blender` dataset: `float16` vs. `float32`.

Table 5: BD-Rate⇓, BD-PSNR⇑, BD-VMAF⇑ results of `float32` on the `8iVFB` dataset, using `float16` as the anchor. (When using `float16` as the anchor, results of `float16` will be all 0s and are thus omitted in the table. Negative BD-PSNR and BD-VMAF results indicate that with the same model size, the quality of views rendered from `float16` models is better compared to `float32` models.)

Sequence	BD-Rate$_p$⇓	BD-PSNR⇑	BD-Rate$_v$⇓	BD-VMAF⇑
Longdress	120.3%	-1.67	120.2%	-17.05
Loot	128.7%	-2.74	128.9%	-19.45
Soldier	166.7%	-1.98	166.2%	-5.73
Redandblack	119.5%	-2.08	119.3%	-18.99

Table 6: BD-Rate⇓, BD-PSNR⇑, BD-VMAF⇑ results of `float32` on the `vsenseVVDB2` dataset, using `float16` as the anchor.

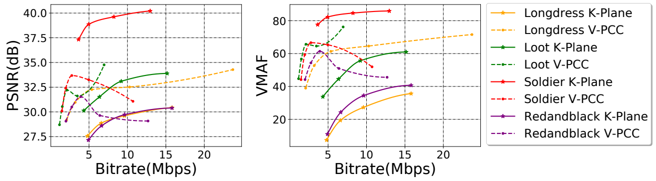
Sequence	BD-Rate$_p$⇓	BD-PSNR⇑	BD-Rate$_v$⇓	BD-VMAF⇑
Rafa	170.9%	-2.15	171.3%	-7.34
Lubna	163.6%	-2.75	162.8%	-10.44
Matis	132.7%	-1.99	132.4%	-15.15
Axeguy	174.1%	-1.82	174.6%	-7.44

Table 7: BD-Rate⇓, BD-PSNR⇑, BD-VMAF⇑ results of `float32` on the `Blender` dataset, using `float16` as the anchor.

Sequence	BD-Rate$_p$⇓	BD-PSNR⇑	BD-Rate$_v$⇓	BD-VMAF⇑
Lego	115.7%	-2.28	114.7%	-14.92
Amily	142.6%	-2.35	143.3%	-11.47
Pig	128.9%	-1.39	128.7%	-12.45

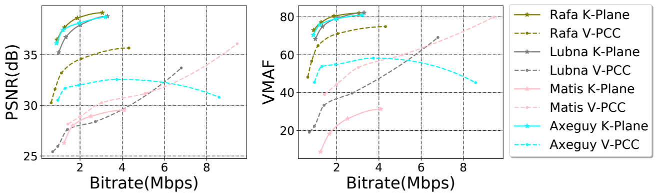
results for both PSNR and VMAF for the `Blender` datasets. The BD-Rate, BD-PSNR, and BD-VMAF results of the three datasets are shown in Tables 5, 6, and 7. In these tables, the `float16` results are used as an anchor for calculating the BD-* results of using float32 for storing the trained model.

The RD-curve results show that the visual quality results of `float16` are comparable to `float32` while `float16` saves more than half of the saved model size. The BD-PSNR results show that when using the same bitrate for representing the dynamic scene, the visual quality of `float32` is 1.39 dB to 2.75 dB worse than `float16`, and the BD-VMAF results show that the VMAF results of `float32` is 5.73 to 19.45 worse than `float16`. Thus, we will use `float16` results for `K-Planes` vs. `V-PCC` comparison in the next section.



(a) PSNR (b) VMAF

Figure 5: RD-curve result for the `8iVFB` dataset.



(a) PSNR (b) VMAF

Figure 6: RD-curve result for the `vsenseVVDB2` dataset.

5 V-PCC VS. K-PLANES

In this section, we report our findings comparing `V-PCC` with `K-Planes` for dynamic 6-DoF volumetric video representation. For `V-PCC`, we compress the raw `.ply` files using five different quantization parameter settings in Table 1. We obtain the `V-PCC` compressed binary file sizes and use them for calculating the "rate" in the RD-curve. We then decode and reconstruct the point cloud `.ply` files, render them, and compare them with the groundtruth test views. For `K-Planes` , we use multi-scale setting S=1,2, train the models for 150, 000 epochs with a learning rate of 0.001, and save the model in `float16`. We further losslessly compress the saved `K-Planes` models using the `.zip` tool and use the compressed `.zip` file for calculating the "rate". We do not use `zip` for compressed `V-PCC` binary files as no data deflation can be achieved.

5.1 Results of `8iVFB` and `vsenseVVDB2` Datasets

We discuss the results of the `8iVFB` and `vsenseVVDB2` datasets first since both datasets are carefully curated raw point cloud data and are made for `V-PCC` . The RD-curve results are shown in Figures 5 and 6. We notice that in a few cases, for `V-PCC`, with increased rate, e.g., the `r5` setting, the visual quality can become worse than lower rate, e.g., the `r3` setting. This finding is consistent with the subjective study performed by Cox et al. [12]. We have also checked our experiments and made sure it is the correct results. We thus report these results in the paper.

We report the BD-PSNR and BD-VMAF results in Tables 8 and 9. (We do not report the BD-Rate results in these tables because the RD-curves of comparative setups are very far away with no overlap in their "distortion" coordinates.) The results show that for the `8iVFB` dataset, `V-PCC` outperforms `K-Planes` in all but one ("Soldier") sequence; while for the `vsenseVVDB2` dataset, `K-Planes` outperforms `V-PCC` in all but one ("Matis").

We find that the performance of `K-Planes` and `V-PCC` appear to be correlated with the amount of motion in the dataset. For example, the four sequences that `K-Planes` perform well in (i.e., "Soldier", "Rafa", "Lubna", and "AxeGuy") are with little to moderate motion. For the remaining four highly dynamic sequences, however, `K-Planes` struggles to achieve a good performance, and `V-PCC` can compress them better.

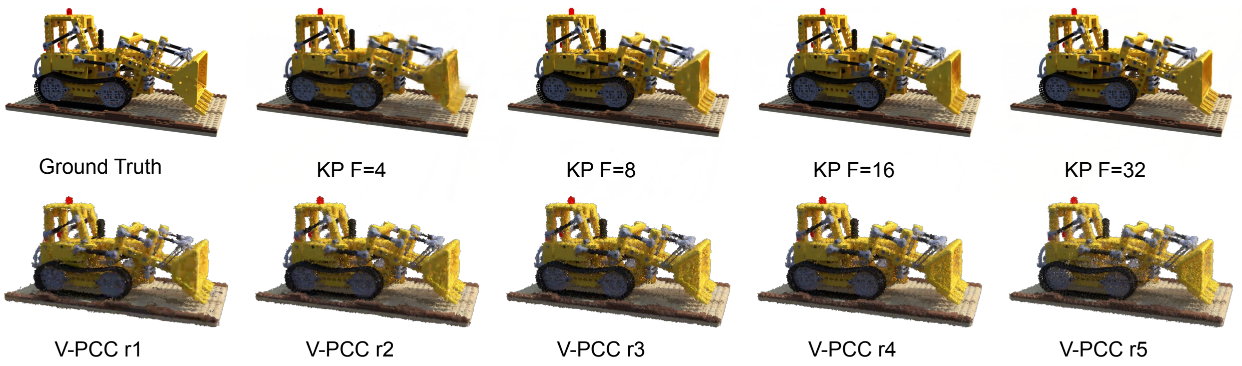


Figure 7: Visual quality comparison of "Lego" in the Blender dataset created by our team. The top row shows the groundtruth view and views rendered by trained K-Planes models with different feature length $F \in \{4, 8, 16, 32\}$. The bottom row shows views rendered by point clouds compressed by V-PCC in one of the five settings r1 (lowest bitrate), r2, r3, r4, and r5 (highest bitrate).

Table 8: 8iVFB dataset: BD-PSNR⇑ and BD-VMAF⇑ results of V-PCC , using K-Planes as the anchor.

Sequence	BD-PSNR⇑	BD-VMAF⇑
Longdress	3.20	39.58
Loot	2.20	30.11
Soldier	-6.45	-21.29
Redandblack	0.49	21.20

Table 9: vsenseVVDB2 dataset: BD-PSNR⇑ and BD-VMAF⇑ results of V-PCC , using K-Planes as the anchor.

Sequence	BD-PSNR⇑	BD-VMAF⇑
Rafa	-4.12	-10.10
Lubna	-9.69	-45.45
Matis	0.76	23.68
Axeguy	-6.06	-23.67

Table 10: Blender dataset: V-PCC results

Sequence	Metric	r1	r2	r3	r4	r5
Lego	PSNR (dB)	20.67	20.96	21.02	20.92	20.80
	VMAF	1.39	1.38	0.80	0.35	0.98
	Size (MB)	5.23	11.13	22.94	43.00	75.81
Amily	PSNR (dB)	26.55	26.29	25.63	24.97	24.92
	VMAF	8.65	5.50	0.95	0.01	0.01
	Size (MB)	0.84	1.31	2.86	8.25	20.21
Pig	PSNR (dB)	29.94	29.94	29.39	28.67	28.68
	VMAF	21.47	21.57	14.60	6.60	8.78
	Size (MB)	1.37	2.01	4.46	12.86	31.70

Table 11: Blender dataset: K-Planes results

Sequence	Metric	F=4	F=8	F=16	F=32
Lego	PSNR (dB)	25.97	27.96	29.07	30.01
	VMAF	38.78	51.19	58.78	64.76
	Size (MB)	1.28	1.70	2.48	4.11
Amily	PSNR (dB)	34.10	35.29	36.34	37.42
	VMAF	56.85	63.32	68.76	72.97
	Size (MB)	0.93	1.46	2.15	3.48
Pig	PSNR (dB)	35.73	37.16	37.79	38.08
	VMAF	49.34	61.45	66.56	70.78
	Size (MB)	1.15	1.56	2.29	3.70

5.2 Results of the Blender Dataset

For the three models in the Blender dataset, we first generate point clouds using a procedure that emulates real-world point cloud capture via RGB-Depth images. For this evaluation, we use the original textured mesh model for generating groundtruth test views.

We report our results in Tables 10 and 11. For V-PCC, while the structural information of the scene is correct, the visual quality is very low. For the "Lego" sequence, the PSNR is only about 20 dB; for "Amily", and "Pig", the PSNR results are lower than 30 dB. The obtained VMAF scores are also very low. We conjecture that the poor visual quality of V-PCC is partially caused by the point clouds recorded via RGB-Depth data, which is inherently noisy. In comparison, the 8iVFB and vsenseVVDB2 datasets are carefully curated. Additionally, another possible cause is that for 8iVFB and vsenseVVDB2 experiments, the groundtruth is rendered using the raw, uncompressed point cloud; while for the Blender experiments, the groundtruth is photorealistic rendering of the model. This results in significantly lower visual quality of V-PCC.

For K-Planes, the results are substantially better. For "Lego", the PSNR can be as high as more than 30 dB; while for "Amily" and "Pig", the PSNR can reach over 37 dB, with a K-Planes representation size of only about 4 MB (or 16 Mbps for the 2-second long sequence.) We further present visual results of four sets of rendered views of the "Lego" sequences in the Blender dataset in Figure 7.

6 CONCLUSION

In this paper, we performed a comparative study of a new dynamic neural scene representation model, K-Planes, and V-PCC for representing and efficiently transmitting 6-DoF volumetric video data. We find that for K-Planes, increasing the length of feature vectors can improve the visual quality faster than increasing the number of multi-scale planes. Results show that the current K-Planes models can outperform V-PCC when there is little to moderate amount of motion in the 6-DoF video sequence. We also find that in a volumetric data capturing scenario emulated by Blender, the visual quality of views rendered from K-Planes is significantly better than V-PCC.

ACKNOWLEDGMENTS

We appreciate constructive comments from anonymous referees. This work is partially supported by NSF under grants CNS-2200042 and CNS-2200048.

REFERENCES

[1] 2018. VMAF: The Journey Continues. https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12.

[2] 2019. zip - package and compress (archive) files. https://manpages.ubuntu.com/manpages/focal/man1/zip.1.html.

[3] 2020. Bjontegaard_metric. https://github.com/Anserw/Bjontegaard_metric.

[4] 2023. bjontegaard. https://pypi.org/project/bjontegaard/.

[5] 2023. Blender 3.5. https://www.blender.org/.

[6] 2023. Free | Amily Animations | Blender Demo. https://www.blender.org/download/demo-files/.

[7] 2023. Free | Piggy Animations | Vfx Grace. https://blendermarket.com/products/piggy-animations-vfx-grace.

[8] 2023. TORCH.AMP. https://pytorch.org/docs/stable/amp.html.

[9] 2024. Draco. https://google.github.io/draco/.

[10] Gisle Bjontegaard. 2001. Calculation of average PSNR differences between RD-curves. *VCEG-M33* (2001).

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision (ECCV)*.

[12] Samuel Rhys Cox, May Lim, and Wei Tsang Ooi. 2023. VOLVQAD: An MPEG V-PCC Volumetric Video Quality Assessment Dataset. In *Proceedings of the 14th Conference on ACM Multimedia Systems*. 357–362.

[13] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A Chou. 2017. 8i voxelized full bodies-a voxelized point cloud dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006* 7, 8 (2017), 11.

[14] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes for Radiance Fields in Space, Time, and Appearance. arXiv:2301.10241 [cs.CV]

[15] D Graziosi, O Nakagami, S Kuma, A Zaghetto, T Suzuki, and A Tabatabai. 2020. An overview of ongoing point cloud compression standardization activities: Video-based (V-PCC) and geometry-based (G-PCC). *APSIPA Transactions on Signal and Information Processing* 9 (2020), e13.

[16] Dan Grois, Detlev Marpe, Amit Mulayoff, Benaya Itzhaky, and Ofer Hadar. 2013. Performance comparison of h. 265/mpeg-hevc, vp9, and h. 264/mpeg-avc encoders. In *2013 Picture Coding Symposium (PCS)*. IEEE, 394–397.

[17] Simon NB Gunkel, Rick Hindriks, Karim M El Assal, Hans M Stokking, Sylvie Dijkstra-Soudarissanane, Frank ter Haar, and Omar Niamut. 2021. VRComm: an end-to-end web system for real-time photorealistic social VR communication. In *Proceedings of the 12th ACM Multimedia Systems Conference*. 65–79.

[18] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware mobile volumetric video streaming. In *Proceedings of the 26th annual international conference on mobile computing and networking*. 1–13.

[19] Henry Haugsten Hansen, Sayed Muchallil, Carsten Griwodz, Vetle Sillerud, and Fredrik Johanssen. 2020. Dense lidar point clouds from room-scale scans. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 88–98.

[20] Branislav Jenco. 2022. *Virtual LiDAR error models in point cloud compression*. Master's thesis.

[21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5521–5531.

[22] Rufael Mekuria, Kees Blom, and Pablo Cesar. 2016. Design, implementation, and evaluation of a point cloud codec for tele-immersive video. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (2016), 828–842.

[23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

[24] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

[25] Sebastian Schwarz, Gaëlle Martin-Cocher, David Flynn, and Madhukar Budagavi. 2018. Common test conditions for point cloud compression. *Document ISO/IEC JTC1/SC29/WG11 w17766, Ljubljana, Slovenia* (2018).

[26] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. 2018. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2018), 133–148.

[27] Shishir Subramanyam, Irene Viola, Jack Jansen, Evangelos Alexiou, Alan Hanjalic, and Pablo Cesar. 2022. Evaluating the impact of tiled user-adaptive real-time point cloud streaming on vr remote communication. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3094–3103.

[28] Jeroen Van Der Hooft, Tim Wauters, Filip De Turck, Christian Timmerer, and Hermann Hellwagner. 2019. Towards 6dof http adaptive streaming through point cloud compression. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2405–2413.

[29] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia* 30, 2 (2023), 48–59.

[30] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. 2003. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology* 13, 7 (2003), 560–576.

[31] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. 2020. Textured mesh vs coloured point cloud: A subjective study for volumetric video compression. In *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*.