Scene Graph Driven Hybrid Interactive VR Teleconferencing

Mingyuan Wu*
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Ruifan Ji*
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Haozhen Zheng
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Jiaxi Li
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Beitong Tian
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Bo Chen
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Ruixiao Zhang
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Jacob Chakareski
New Jersey Institute for Technology
Jersey City,, New Jersey,, USA

Michael Zink
University of Massachusetts, Amherst
Amherst, Massachusetts, USA

Ramesh Sitaraman
University of Massachusetts, Amherst
Amherst, Massachusetts, USA

Klara Nahrstedt
University of Illinois at Urbana-Champaign
Champaign-Urbana, Illinois, USA

Abstract

We propose an interactive and intelligent hybrid teleconferencing system compatible with Virtual Reality devices. Our system understands meeting contexts and leverages user interactions to enable better system configurations. By employing interactive scene graphs [11], our system extracts and transmits essential meeting context to users while relaying user interactions back to the streaming systems for user-involved adaptive streaming and foveated rendering. We demonstrate the system's real-time performance and compatibility with commercial VR devices such as the Meta Quest 3.

CCS Concepts

• **Human-centered computing**; • **Computing methodologies** → **Scene understanding**; **Virtual reality**;

Keywords

Virtual Reality, Teleconferencing System, Machine Learning

ACM Reference Format:
Mingyuan Wu*, Ruifan Ji*, Haozhen Zheng, Jiaxi Li, Beitong Tian, Bo Chen, Ruixiao Zhang, Jacob Chakareski, Michael Zink, Ramesh Sitaraman, and Klara Nahrstedt. 2024. Scene Graph Driven Hybrid Interactive VR Teleconferencing. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3664647.3684996

*Mingyuan Wu and Revan Ji contributed equally to the paper. Klara Nahrstedt is the primary advisor

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3684996

1 Introduction and System Features

The development of teleconferencing systems has served a critical role in enhancing social connections. Today, these systems are on the verge of a revolutionary transformation with recent advances in Virtual Reality (VR) technology, including more affordable Head Mounted Devices (HMDs) for immersiveness and better controllers for interactivity. VR products can reshape the teleconferencing landscape, moving beyond the traditional experience of navigating a 2D grid of faces during a meeting.

How can VR benefit teleconferencing? Many studies [3–5, 7] developed volumetric video streaming systems, believing that better 3D experiences enable a sense of immersion. However, these systems are often static and lack interactivity, for which teleconferencing is designed. Some studies [2, 6, 8–10, 12] built systems that passively adapt to user viewing patterns or behavior based on predictions. However, such predictions can be inaccurate, and users cannot actively customize the system to their preferences. Additionally, user interaction spaces cannot be unlimited, as this is not supported by a real system. A constrained interaction space is desired, and this space should be updated in real-time during the meeting.

To solve the above mentioned challenges, we built our teleconferencing system with interactivity insights from the VR community: Meta proposed SceneScript [1] that combines contextualized AI with an HMD to provide seamless access to real-time contextual information and assist with real-world interactions. In this paper, our teleconferencing system integrates contextual information via HMDs and allows for constrained user interactions via VR controllers. The system includes three key components: (a) contextual vision understanding of meetings, (b) constrained interactions between any remote meeting participant and meeting contexts via

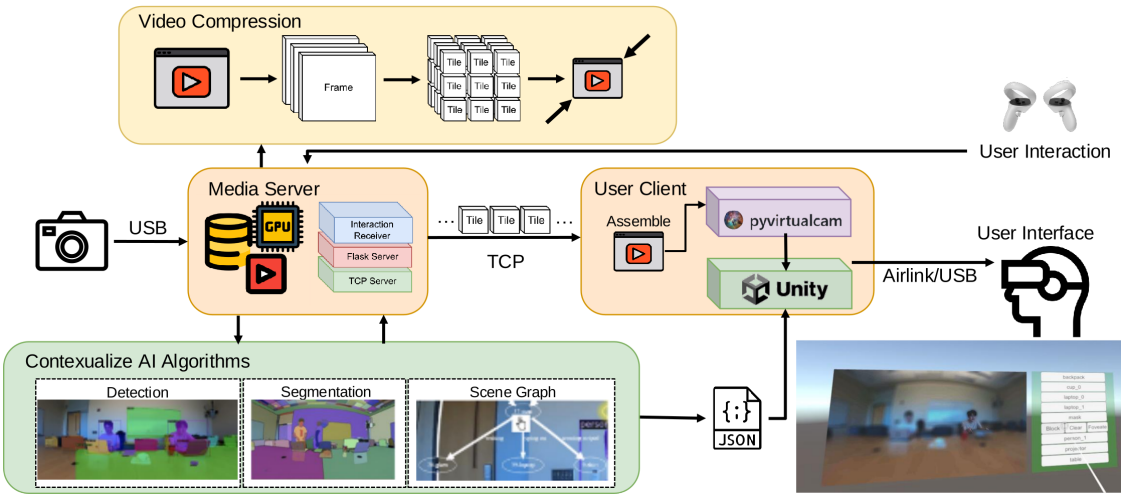

Figure 1: Model Architecture. Contextual AI, System Components, User Interface.

VR devices and (c) a tile-based video compression and streaming pipeline. An interactive scene graph [11] is used to build contextual understandings from a room camera stream in the meeting room. The model extracts important meeting contextual information, including objects, segmentation masks, and object relationships. The scene graph representation is stored in the JSON format and transmitted to the remote meeting participant. With the streamed scene graph, the remote participant can see the list of objects and interact with the interface to identify objects of interest. When the user interaction information is received, a tile-based adaptive streaming system uses this information to determine the importance of each tile. Unimportant scene objects will be blurred. Since each title is separately compressed and transmitted with different bitrates, unimportant tiles can be drastically compressed, reducing latency and saving bandwidth for a better system configuration.

Our contributions can be summarized as follows:

Design Novelty. Our teleconferencing system is the first to incorporate the concept of scene graph (contextual AI algorithms), interaction feedback (user) and tile-based adaptive streaming (system) and form a positive feedback loop between them. The system fundamentally differs from other passively adaptive streaming methods based on predictions of user's viewing patterns or salient regions.

System Implementation. The teleconferencing system we built is fully compatible with the commercial PC VR. It is also optimized for real-time streaming and user interactions.

2 System Overview

In this demo, we show our scene graph driven hybrid interactive VR teleconferencing between a camera, a media server, a user client and a Meta Quest 3, as shown in Figure 1. In our hybrid setup, an in-person conference is happening in a physical meeting room, while a remote participant is joining from VR. The video of the conference is being streamed from the meeting room to the remote participant's headset in real time. The remote participant can interact with the stream using VR controllers.

Our streaming setup consists of client and server parts. The server is responsible for capturing the video, dividing the frame into tiles, compressing each tile into different qualities, and sending them via TCP. It also produces a scene graph and segmentation masks of important objects [11] in real-time and serves the JSON file on a Flask server. The client, taking the compressed video streams, reassembles them into a full video and exposes it as a virtual camera stream. A Unity application then displays the camera stream to the remote participant. The application also captures how the user interacts with the meeting context, and relays it to the server. Upon receiving these interactions, the server blurs out certain areas and adjusts the compression aggressiveness of each tile based on the user's interactions, ensuring high quality for regions of interest and lower quality for less important areas, thereby optimizing the streaming efficiency.

Unity User Interface As shown in Figure 1: The video stream is displayed in front of the user, while a list of scene objects, shown as buttons, appears on the right and is updated with contextual information. Using the VR controller, the participant can click on a button to access options to Block, Clear, or Foveate the objects. "Foveate" blurs out everything except the selected object, "Clear" shows the objects, and "Block" blurs out the objects.

System Performance. We report the latency and bandwidth for the main components of our teleconferencing system. When a user foveates on a person in the meeting, the bandwidth of tile-based frames decreases from 67295 bytes per frame to 51444 bytes per frame on average, resulting in a 16.1% bandwidth savings. Additionally, the average end-to-end latency of the user interaction feedback loop is 367 milliseconds, while the average end-to-end latency for tile-based video streaming and compression is 295 milliseconds. The contextual AI, running on an Nvidia RTX 4090, achieves 3.4 FPS for scene graph generation and 31.3 FPS for real-time efficient segmentation. Overall, the system operates in real time and meets the requirements of teleconferencing.

3 Conclusion

In our work, we introduced a scene graph-driven hybrid teleconferencing system designed for better interactivity and system efficiency. Our system features vision-based contextual understanding, constrained user interactions, and a tile-based compression and streaming pipeline. The system creates a positive feedback loop between the three components, paving the way for future diverse research aimed at enhancing the interactivity of teleconferencing.

4 Acknowledgement

This work was supported by the National Science Foundation grants NSF CNS 19-00875, NSF CNS 21-06592, NSF OAC 18-35834 KN, NSF CCF 22-17144, NSF CNS-1901137, and NSF CNS-2106463. Jacob Chakareski has been supported in part by NSF CCF-2031881, NSF ECCS-2032387, NSF CNS-2040088, NSF CNS-2032033, and NSF CNS-2106150; NIH R01EY030470; and by the Panasonic Chair of Sustainability at the New Jersey Institute for Technology. Any results and opinions are our own and do not represent views of National Science Foundation.

References

[1] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. 2024. SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model. arXiv:2403.13064 [cs.CV]

[2] Bo Chen, Zhisheng Yan, Haiming Jin, and Klara Nahrstedt. 2019. Event-Driven Stitching for Tile-Based Live 360 Video Streaming. In *Proceedings of the 10th ACM Multimedia Systems Conference*. Association for Computing Machinery, 1–12.

[3] Matthias De Fré, Jeroen van der Hooft, Tim Wauters, and Filip De Turck. 2024. Scalable MDC-Based Volumetric Video Delivery for Real-Time One-to-Many WebRTC Conferencing. In *Proceedings of the 15th ACM Multimedia Systems Conference* (, Bari, Italy,) *(MMSys '24)*. Association for Computing Machinery, New York, NY, USA, 121–131. https://doi.org/10.1145/3625468.3647617

[4] Simon N.B. Gunkel, Rick Hindriks, Yonatan Shiferaw, Sylvie Dijkstra-Soudarissanane, and Omar Niamut. 2024. VP9 bitstream-based Tiled Multipoint Control Unit: Scaling simultaneous RGBD user streams in an immersive 3D communication system. In *Proceedings of the 15th ACM Multimedia Systems Conference* (, Bari, Italy,) *(MMSys '24)*. Association for Computing Machinery, New York, NY, USA, 23–33. https://doi.org/10.1145/3625468.3647608

[5] Simon N. B. 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* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 65–79. https://doi.org/10.1145/3458305.3459595

[6] Anh Nguyen and Zhisheng Yan. 2023. Enhancing 360 Video Streaming through Salient Content in Head-Mounted Displays. *Sensors* 23, 8 (2023).

[7] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[8] Jounsup Park, Mingyuan Wu, Kuan-Ying Lee, Bo Chen, Klara Nahrstedt, Michael Zink, and Ramesh Sitaraman. 2020. SEAWARE: Semantic Aware View Prediction System for 360-degree Video Streaming. In *2020 IEEE International Symposium on Multimedia (ISM)*. 57–64. https://doi.org/10.1109/ISM.2020.00016

[9] Jangwoo Son and Eun-Seok Ryu. 2018. Tile-based 360-degree video streaming for mobile virtual reality in cyber physical system. *Computers & Electrical Engineering* 72 (2018), 361–368.

[10] Atsushi Tagami, Kazuaki Ueda, Rikisenia Lukita, Jacopo De Benedetto, Mayutan Arumaithurai, Giulio Rossi, Andrea Detti, and Toru Hasegawa. 2019. Tile-Based Panoramic Live Video Streaming on ICN. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. 1–6.

[11] Mingyuan Wu, Yuhan Lu, Shiv Trivedi, Bo Chen, Qian Zhou, Lingdong Wang, Simran Singh, Michael Zink, Ramesh Sitaraman, Jacob Chakareski, and Klara Nahrstedt. 2023. Interactive Scene Graph Analysis for Future Intelligent Teleconferencing Systems. In *2023 IEEE International Symposium on Multimedia (ISM)*. 251–255. https://doi.org/10.1109/ISM59092.2023.00048

[12] Lan Xie, Zhimin Xu, Yixuan Ban, Xinggong Zhang, and Zongming Guo. 2017. 360ProbDASH: Improving QoE of 360 Video Streaming Using Tile-Based HTTP Adaptive Streaming. In *Proceedings of the 25th ACM International Conference on Multimedia*. 315–323.