A Multi-person Real-virtual Fusion Based Telepresence System

Xilong Zhou[1], Shizhou Shi[1], Liuxin Zhang[1], Chuanmin Jia[2], Siwei Ma[2] and Qianying Wang[1]

[1]*Lenovo (Beijing), Co., Ltd.* [2]*Peking University*

***Abstract*— In this paper, we introduce a novel immersive video conferencing system capable of real-time, multi-person viewing, smooth, high-fidelity, life-sized autostereoscopic display. Compared to the immersive conferencing system we previously proposed in [1], the system established in this paper supports multi-user viewing without the need for additional eye-tracking devices. We employ an adaptive low-complexity view synthesis method based on an accelerated Real-time Intermediate Flow Estimation (RIFE) model for multi-view generation, followed by real-time light field encoding to achieve realistic 3D rendering. To meet the needs of conferencing scenarios, our system also integrates real-virtual fusion functionality.**

Keyword: real-time view synthesis, multiple users, virtual-real fusion

I. INTRODUCTION

In recent years, the demand for high-quality video has steadily increased, with users increasingly seeking more realistic and immersive 3D visual experiences. To meet this demand, researchers have been exploring real-time remote portrait rendering solutions that can provide immersive visual experiences. As intelligent video encoding technology continues to advance, many real-time remote rendering solutions for conference scenarios have emerged. Google's Project Starline is the first real-time remote rendering system to use light fields for auto-stereoscopic display, significantly enhancing the user experience compared to traditional 2D conferencing [2].

Another typical immersive conferencing system is VirtualCube[3], which uses 2D screens in the surrounding environment to create an immersive visual experience, catering to the visual needs of users in multi-user conference scenarios. Additionally, other free-viewpoint video (FVV) systems utilize depth image-based rendering (DIBR) methods for virtual view synthesis [4]-[6]. However, due to the significant computational demands, these processing pipelines usually require substantial hardware resources, limiting the widespread adoption of current 3D conferencing systems.

In our previous work [1], we proposed a low-complexity 3D video conferencing system, which supported high-fidelity autostereoscopic visual presentation. However, since the generation of new views relied on eye-tracking, the system could not meet the needs of multi-user conferencing scenarios.

This paper proposes a novel multi-user immersive conferencing system that utilizes the emerging real-time intermediate flow estimation (RIFE) model, IFNet[7], to achieve view synthesis, offering high-fidelity and low-cost auto-stereoscopic remote presentation for multiple users. We have optimized the low-complexity view synthesis model, IFNet v4.6, to accelerate the generation of multi-viewpoint frames within the system. Additionally, we integrated virtual-real fusion technology into our system, providing users in multi-user conferencing scenarios with more flexible virtual environment options. To our knowledge, this is the first system to achieve auto-stereoscopic presentation in multi-user conferencing scenarios using the RIFE model.

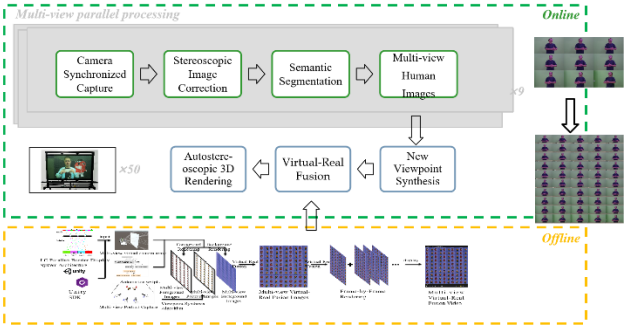

Figure 1. Overview of the proposed immersive conferencing system

II. SYSTEM ARCHITECTURE OVERVIEW

As shown in Figure 1, our system's overall process begins with the camera capturing the original videos. These video frames undergo stereoscopic image correction based on the camera parameters. The corrected images are then semantically segmented to obtain RGBA images for the human figure. Using the RIFE-based optical flow estimation algorithm, new viewpoints are synthesized. After virtual-real fusion, real-time light field encoding is performed, and the output is presented on an autostereoscopic 3D display.

Human Video Capture. Our capturing devices contains nine Basler industrial cameras uniformly along a straight-line track. The spacing of the cameras is determined by the field of view (FOV) of the autostereoscopic display and the distance to the camera convergence point. In our system, the field of view angle for autostereoscopic display is 60° , the camera spacing is 30cm, and the convergence point is 2 meters from the camera array. The cameras capture video at a full resolution of 2464×2056, which is then center cropped with 1440×810 region to eliminate any pixel padding. We use a signal generator to produce a stable transistor-transistor logic (TTL) square wave to synchronize the multi-camera capture process. The camera parameter matrices are obtained using Zhang's calibration method [8]. Stereoscopic image correction

is achieved by aligning the image planes of the camera array according to the perspective transformation.

We deploy the RVM segmentation model [9] to process the corrected images, removing unnecessary background parts and extracting human images with accelerated inference for millisecond-level segmentation. Each module is optimized using FP16 precision and CUDA to better utilize current system performance. Post-processing is done using boundary filtering, optimizing the segmentation effect without increasing extra computational cost.

Novel Viewpoint Synthesis. We devised a pre-trained accelerated RIFE model to interpolate 50 equally spaced new views within the field of view of a 9-camera array. To achieve rapid stereoscopic vision generation, we made the following improvements:

(1) Optimized the network structure of the viewpoint synthesis part to make it more lightweight: We completely re-implemented the architecture using the TensorRT library in C++ code instead of running the original Python code with the PyTorch library [10]. All network parameters in the RIFE model were quantized from 32-bit floating-point format to 16-bit floating-point format.

(2) Added a feature extraction module to extract features from the source viewpoints.

With these improvements, the system synthesizes 50 new views in 35ms for each image.

Autostereoscopic Vision. We built a virtual capture end based on Unity and the LC parallax barrier display principle, simultaneously completing the 50-viewpoint virtual scene capture. The script allows for adjusting the scene depth, dynamic effects, and foreground-background settings (setting the focal plane, frame-by-frame rendering, layer settings).

After capturing the virtual foreground, we use a synthesis algorithm to fuse the multi-viewpoint portrait images, multi-viewpoint virtual foreground images, and multi-viewpoint background images into multi-viewpoint virtual-real fusion images (50-grid).

Finally, we use a real-time light field encoding algorithm to render the multi-view images into interleaved images, which are output to a 65-inch Lenovo autostereoscopic light field display to achieve autostereoscopic vision. Our system enables clear and smooth stereoscopic display within a 60° field of view, allowing viewers to see without wearing any auxiliary devices and supporting simultaneous viewing by multiple people. The system's frame rate can reach 20fps (50 ms/50 viewpoints).

III. DEMONSTRATION

A camera array is employed consisting of nine uniformly distributed Basler industrial cameras with 8mm lenses. The captured videos are transmitted to the host via two video graphic cards and synchronized using a signal generator. The entire processing pipeline runs on the GPU memory. Viewpoint synthesis is handled by two RTX 4090 GPUs, while other tasks are executed on an RTX 5000 Ada GPU. The display end uses a 65-inch 8K autostereoscopic light field screen from Lenovo for autostereoscopic presentation. The workflow is divided into three modules: multi-viewpoint preprocessing (24ms), new viewpoint synthesis (35ms), and virtual-real fusion with 3D rendering (8ms). The system's total runtime is determined by the longest module, which is new viewpoint synthesis (35ms). This parallel processing architecture significantly enhances performance.

Figure 2. Viewing autostereoscopic content from different angles

As shown in Figure 2, our final presentation effect is a combination of the real and the virtual. The real person, the virtual character, and the virtual background are all presented autostereoscopically by the system. Viewers can see different content from different angles.

IV. CONCLUSION

In this paper, we present a novel immersive multi-user conferencing system designed to meet future users' pursuit of immersive 3D visual experiences. We meticulously designed and optimized the RIFE model to achieve real-time multi-view generation and developed a cross-view rendering method for virtual-real fusion that supports simultaneous viewing by multiple users. Next step we will continues improve our system using higher resolution (16K) bigger size (110) autostereoscopic multi-person display to achieve more immersive 3D experiences.

REFERENCES

[1] H. Huang et al., "Low-Complexity 3D-Vision Conferencing System based on Accelerated RIFE Model," IEEE Picture Coding Symposium (PCS), Taichung, Taiwan, 2024, pp. 1-5.

[2] J. Laurence, D. B. Goldman, A. Supreeth, et al., "Project Starline: A High-Fidelity Telepresence System," ACM Trans. Graph., vol. 40, no.6, pp. 1 – 16, 2021.

[3] Y. Zhang, J. Yang, Z. Liu, et al., "VirtualCube: An Immersive 3D Video Communication System," IEEE Trans. Vis. Comput. Graph., vol. 28, no.5, pp. 2146 – 2156, 2022.

[4] O. Stankiewicz, M. Domanski, A. Dziembowski, et al., "A Free-´ Viewpoint Television System for Horizontal Virtual Navigation," IEEE Trans. Multimedia, vol. 20, no. 8, pp. 2182 – 2195, 2018.

[5] Y. Dong, L. Song, R. Xie, and W. Zhang, "An Elastic System Architecture for Edge Based Low Latency Interactive Video Applications," IEEE Trans. Broadcast., vol. 67, no. 4, pp. 824 – 836, 2021.

[6] P. Carballeira, C. Carmona, C. D´ıaz, et al., "FVV Live: A Real-Time Free-Viewpoint Video System with Consumer Electronics Hardware," IEEE Trans. Multimedia, vol. 24, pp. 2378 – 2391, 2022.

[7] Z. Huang, T. Zhang, W. Heng, et al., "Real-Time Intermediate Flow Estimation for Video Frame Interpolation," in Eur. Conf. Comput. Vis.,Tel Aviv, Israel, 2022, pp. 624 – 642.

[8] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1330 – 1334, 2000.

[9] S. Lin, L. Yang, I. Saleemi and S. Sengupta, "Robust High-Resolution Video Matting with Temporal Guidance," IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 3132-3141.

[10] Z. Huang, "Practical-RIFE," Accessed: Oct. 22, 2023. [Online]. Available: https://github.com/hzwer/Practical-RIF.