Multifocal Display System for Near-Eye Device and Optimal Decomposition Algorithm for Video Contents

Wooseup Youm, Yong-Duck Kim*, Kukjoo Kim*, Chan-mo Kang*, Jin-wook Shin*, Hyunsu Cho*, Chul Woong Joo*, Sukyung Choi*, Dae Hyun Ahn*, Byoung-Hwa Kwon*, In Bok Baek*, Nam Sung Cho*, Yongkyu Choi**, Kangmin Kim**, Hee Young Um**, and Chun-Won Byun**

***Reality Display Research Section, ETRI, Daejeon, Republic of Korea**
****Maxlogic, Inc., Gyeonggi-do, Republic of Korea**
Email: cwbyun@etri.re.kr

Abstract

In this paper, we discuss advanced technologies aimed at delivering high-quality stereoscopic images within artificial reality (AR), virtual reality (VR), and extended reality (XR) devices. Providing a realistic stereoscopic effect is crucial when implementing VR through wearable devices, as it significantly enhances the user's sense of immersion. Without accurate depth cues, presenting virtual environments as flat images or simple binocular parallax often leads to discomfort or dizziness due to the vergence and accommodation conflict (VAC). To address this issue, we have simultaneously developed hardware capable of generating multifocal images and software featuring a decomposition algorithm for video content.

Author Keywords

AR/VR/MR devices, VAC (Vergence and Accommodation Conflict), multifocal display system, Decomposition algorithm.

1. Introduction

XR technology first captured significant attention during CES 2016, where the growth potential of the AR/VR/XR market was widely recognized. However, due to the absence of sufficiently advanced devices and services capable of meeting consumer expectations, its adoption remained limited for several years. Recently, rapid advancements in generative AI have reignited interest in XR technology. AI-powered solutions now enable real-time processing of large-scale data, particularly in spatial recognition and high-resolution video rendering—capabilities that were previously challenging to achieve. Furthermore, AI facilitates the generation of images, audio, and haptic feedback within virtual environments, significantly enhancing user interaction in immersive digital spaces. Despite these advancements, delivering stereoscopic visual information that closely mimics human vision remains a major technical challenge. Despite the growing demand for XR solutions, a clear and comprehensive approach to overcoming these challenges has not yet been established.

Unlike traditional television broadcasts, XR video content has to dynamically respond to the user's movements, creating a truly immersive virtual environment. This requirement goes beyond merely displaying fixed video on a screen; instead, the visual output must adjust naturally based on the user's gaze and movement to provide an accurate sense of depth. However, most current XR devices still rely on fixed image projection, failing to deliver sufficient depth perception. This results in the vergence-accommodation conflict (VAC), which can lead to discomfort and dizziness, as illustrated in Figure 1. Additionally, if an XR device operates at a low frame rate or lacks efficient video processing capabilities, motion-to-photon (MTP) latency can further exacerbate motion sickness. Another critical issue is display resolution—low-resolution XR screens can cause the screen door effect, which disrupts immersion by creating visible gaps between pixels. Addressing these challenges necessitates the development of high-speed, high-resolution display panels.

To mitigate these limitations, various XR devices have been introduced to the market, as summarized in Table 1. However, no single device has yet overcome all technological constraints. Recent research presented by Meta and Carnegie Mellon University at SPIE suggests potential solutions to these challenges. Nevertheless, most existing studies primarily focus on stereoscopic imaging techniques and optical technologies. For practical applications, further advancements are required, particularly in reducing the size and weight of XR devices while also developing sophisticated playback technologies capable of supporting diverse visual reproduction methods.

In this paper, we propose a multifocal display system designed to overcome VAC and other technological constraints. Our approach integrates a device system for near-eye displays along with a decomposition algorithm for video content, effectively addressing the shortcomings of current XR technologies. By implementing these innovations, we aim to enhance the realism and comfort of XR experiences, paving the way for more immersive and practical applications.

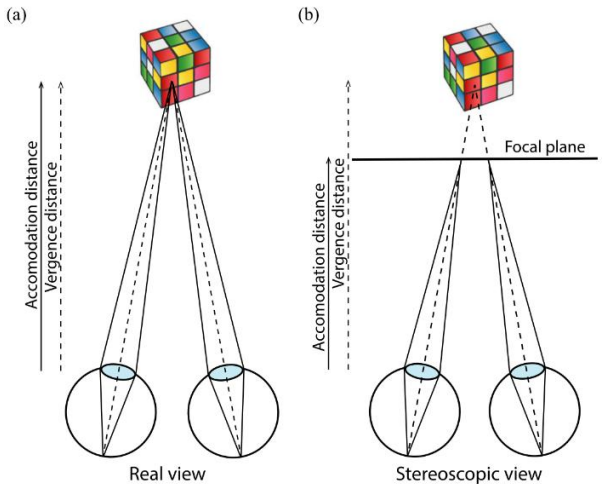


Fig. 1. Comparison of Vergence and Accommodation in (a) real world and (b) stereoscopic displays [1]

ISSN 0097-996X/25/5602-1020-$1.00 © 2025

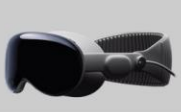
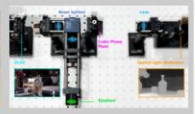
Company	Apple	Meta	Magic Leap	Creal	Meta	CMU (Carnegie Mellon Univ.)
Model						
Display	OLEDoS	LCD	LCoS	DMD	LCD	OLED & SLM
Optic	Pancake	Pancake	Waveguides	Holographic Film	Pancake	Lohmann & Alvaez
FoV	110˚	110˚	45˚	36˚	50˚	25˚
3D Expression	Binocular	Binocular	Binocular	Light Field	Light Field	Light Field
Platform	VisionOS	Customized Android	Customized Android	-	-	-
Status	Release	Release	Release	Conference	Prototype	Prototype

Table 1. Recent Trends in XR Devices: Display Technologies, Optical Systems, and Development Status

2. Proposed Multi-focal Display Device: Hardware

In previous research aimed at reducing the VAC, vari-focal, multi-focal, and light field displays have been proposed [2]. The vari-focal display measures the user's gaze position and depth through eye-tracking and adjusts the display screen to match the user's focal depth, while rendering objects outside the focal depth as blurred. However, the realism of images rendered through simulation is lower compared to that of the multi-focal display [3]. The light field display generates a light field that controls not only the intensity but also the direction of light rays in space, delivering a stereoscopic image to the user. The light field display requires a high computational load for rendering, and there are technical challenges in maintaining display resolution as the number of implementation viewpoints (positions and focal points) increases. The multi-focal display, on the other hand, divides space or time to provide the user with images across a limited number of focal planes, allowing for a natural optical blur effect outside the focal depth without requiring eye-tracking. Although the multi-focal display demands higher driving speeds for both the variable optical system and display, proportional to the number of focal planes, it is comparatively easier to implement and delivers effective results over the other two technologies.

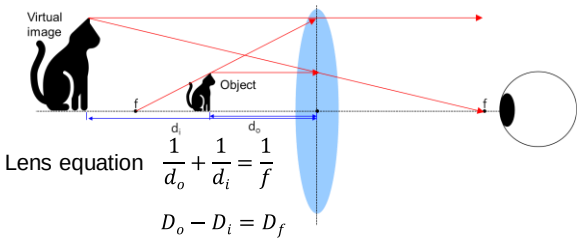


Fig. 2. Focus tunable lens equation.

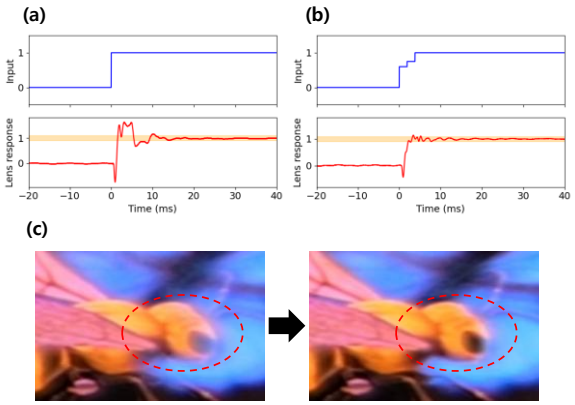


Fig. 3. Optimization of the input signal profile for FTL

As shown in Fig. 2, although the position of the object (OLEDoS panel) remains fixed, this approach is used to vary the position of the virtual image. For the focus tuning device, a variable-focus optical system capable of controlling focal planes is constructed using a deformable lens driven by a voice coil motor (VCM) mechanism [4]. For multifocal image rendering, this paper employed a focus-tunable lens (FTL) (EL-10-30-C-VIS-LD, Optotune) based on the VCM mechanism, enabling the implementation of three focal planes. The FTL consists of a polymer membrane enclosing an optical fluid, with an actuator coil surrounding the container to modulate internal pressure. By adjusting the driving current supplied to the actuator, the curvature of the membrane changes, thereby dynamically altering the focal length of the lens. This approach enables continuous optical power tuning over a broad range. However, since the lens relies on physical deformation for focal length adjustment, its response speed is inherently limited.

In this study, the display system operates at 90 frames per second (fps), sequentially presenting distinct images across three focal planes. Consequently, the FTL also has to function at a synchronization rate of 90 Hz with a frame duration of approximately 11.1 ms. If the stabilization time of the FTL is comparable to or exceeds this duration, achieving accurate multifocal image representation becomes challenging.

To assess the response characteristics of the FTL, a photodiode-based position sensor (PDP90A, Thorlabs) was employed to monitor the positional variation of a laser beam passing through the lens. As shown in Fig. 3(a), when a rectangular step input was applied to change the optical power from 5 to 7 diopters,

significant fluctuations were observed before stabilization, with a stabilization time of approximately 10.1 ms. This duration is nearly equivalent to the frame time (11.1 ms), posing a critical limitation in achieving stable focal plane rendering within a single frame.

To mitigate this issue and enhance focal plane stability, an input shaping technique was applied, modifying the input waveform to reduce the stabilization time. Given the inherent nonlinearity of the FTL, an analytical modeling approach was deemed impractical. Instead, a trial-and-error method was adopted to refine the input waveform. As illustrated in Fig. 3(b), the optimized driving waveform significantly reduced the stabilization time to approximately 4.6 ms. As a result, applying input shaping to the 5-6-7 diopter transitions resulted in improved multifocal image quality compared to the step input, as shown in Fig. 3(c).

3. Proposed Multi-focal Display Devices: Software

For enhancing the immersion of multi-focal displays, the content implemented on the device is also of critical importance. The multifocal MPEG Immersive Video (MIV) content play system seamlessly integrates an MIV player with a multi-focal OLED display module. MIV is a standard established by ISO/IEC JTC1/SC29/WG 04 [5], which defines the methods for generating, transmitting, and rendering stereoscopic video content.

MIV compresses visual data through pruning, which removes redundant information from multiple omnidirectional viewpoints in 3D space, and patch packing, which optimizes video transmission bandwidth. For stereoscopic visualization, the visual information consists of texture data, representing the image displayed on the screen, and depth maps, containing depth information for 3D representation. These are separately encoded and transmitted. The multifocal MIV content player decodes MIV content, compressed and stored using high efficiency video coding (HEVC) on a demonstration laptop, to demonstrate multi-focal stereoscopic images as illustrated in Fig. 4 and Fig. 5. Based on the user's posture information measured by an inertial measurement unit (IMU) sensor, the player synthesizes viewpoint-dependent textures and depth maps through MIV rendering. Using the depth information, a linear decomposition is performed to split the content into three focal planes, and the resulting images are outputted to the handheld demonstration device according to each focal plane. To handle the complex image processing algorithms designed to minimize MIV content size, most of the player's processing functions are executed on an NVIDIA RTX4090 GPU utilizing the CUDA API, ensuring real-time playback.

However, the linear decomposition method, chosen for its ease of implementation, has the drawback of generating defects at the boundaries between images of different focal planes. To address this image quality degradation, ongoing research is being conducted on developing a deep learning-based multifocal decomposition algorithm and applying it to the content player. As shown in Fig. 6 and Fig.7, the hand-held type multi-focal device

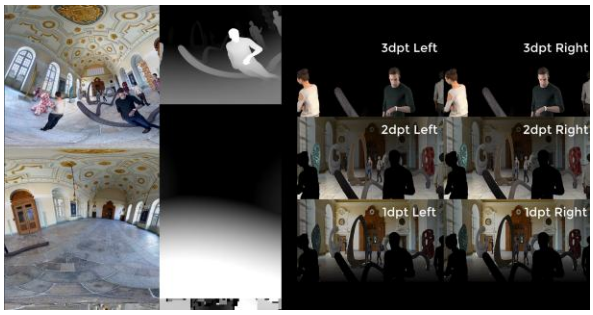


Fig. 4. MIV Decoding and play structure

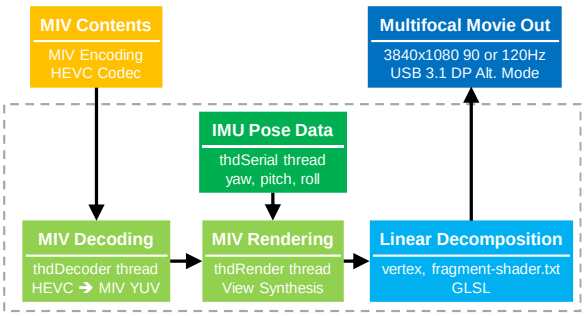


Fig. 5. The multifocal MIV content player

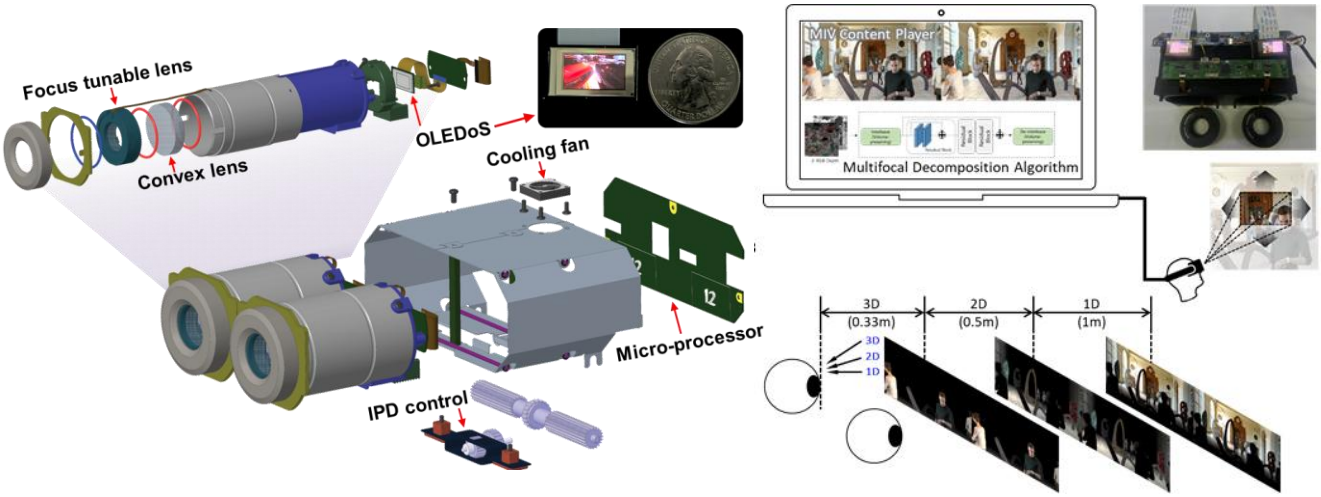


Fig. 6. Proposed hand-held type multi-focal device with decomposition algorithm for video contents.

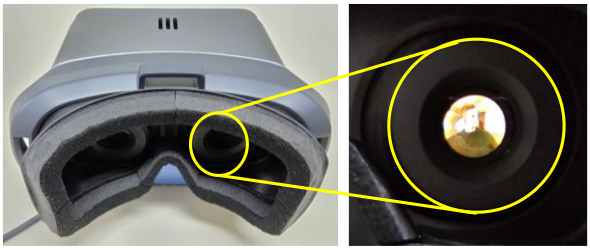


Fig. 7. A photograph of the proposed hand-held type multi-focal device.

with a decomposition algorithm for video content was implemented to operate at 90 Hz, with three sub-frames for each focal plane.

4. Conclusions

To realize a high-quality multi-focal stereoscopic display system, several technical challenges in both device and player technologies must be addressed. On the device side, it is essential to develop multi-focal lenses capable of rapid operation while covering large display areas, all while maintaining low power consumption for practical usability. Additionally, the integration of high-resolution, high-density, high-speed, and high-brightness self-emissive microdisplay is crucial to achieving immersive and realistic stereoscopic visualization.

From the player technology perspective, ensuring sufficient computing power to handle complex rendering and processing requirements is a primary concern, particularly in mobile environments. With the continuous evolution of mobile devices, achieving a balance between real-time performance, power efficiency, and portability becomes increasingly critical. Future research should focus on optimizing the computational pipeline and exploring advanced decomposition algorithms, such as deep learning-based approaches, to enhance visual quality while meeting the constraints of mobile platforms.

5. Acknowledgements

Wooseup Youm and Yong-Duck Kim contributed equally to this work. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00026, Near-eye light field device technology development for hyper-realistic metaverse service, 70%) and internal fund of Electronics and Telecommunications Research Institute (ETRI). [24BC1800, Development of key technology elements for the construction of real-virtual convergence spatial media metaverse, 30%]

6. References

[1] Zhan, T., Xiong, J., Zou, J. et al. Multifocal displays: review and prospect. PhotoniX 1, 10 (2020). https://doi.org/10.1186/s43074-020-00010-0

[2] Xiao, L., Kaplanyan, A., Fix, A., Chapman, M., & Lanman, D. (2018). Deepfocus: Learned image synthesis for computational display. In ACM SIGGRAPH 2018 Talks (pp. 1-2). https://doi.org/10.1145/3214745.3214769

[3] March, Joseph, et al. "Impact of correct and simulated focus cues on perceived realism." SIGGRAPH Asia (2022) Conference Papers. https://doi.org/10.1145/3550469.3555405

[4] Rathinavel, Kishore, et al. "An extended depth-at-field volumetric near-eye augmented reality display." IEEE transactions on visualization and computer graphics 24.11 (2018): 2857-2866. https://doi.org/10.1109/TVCG.2018.2868570

[5] J. M. Boyce et al., "MPEG Immersive Video Coding Standard," in Proceedings of the IEEE, vol. 109, no. 9, pp. 1521-1536, Sept. 2021, https://doi.org/10.1109/JPROC.2021.3062590.