PRJ5-1
Invited

Workflow and Technologies for Immersive XR

Hiroshi Mukawa

Hiroshi.Mukawa@sony.com
R&D Center, Sony Group Corporation
2-10-1 Osaki, Shinagawa-ku, Tokyo, Japan
Keywords: Volumetric capture, Motion capture, 3D audio, Retinal scan display, Motion to photon latency

ABSTRACT

A number of technologies must be highly integrated to deliver immersive XR experiences to users. This applies not only to XR hardware but also to content creation and distribution technologies. In this paper, the author will discuss an immersive XR workflow and introduce some of the key technologies.

1 Introduction

The XR, meaning augmented reality (AR), virtual reality (VR), and mixed reality (MR), technologies are increasing the visibility and importance as applications related to “Metaverse” is getting considered as one of the next growth business domains [1]. In those applications, an immersive XR experience plays an important role. Users can feel the sense of presence or the sense of reality through the immersive XR experience. To deliver the experience to users, reality expression and real-time interaction are the key parameters.

To boost reality expression, visual, audio, and haptic technologies should be advanced. For example, both visual and audio expressions are expected to evolve from a conventional 2-dimensional flat expression to a 3-dimensional volumetric one. The rendering latency would be a critical issue for the real-time interaction as the amount of data of the display signal tends to be large for a higher resolution and dynamic range. These imply that new technologies are required to complete the practical immersive XR workflow.

In this paper, two video and one audio content creation technologies, as well as two display technologies for the immersive XR are introduced.

2 Immersive XR Workflow and Key Technologies

The immersive XR workflow consists of content creation, distribution, sensing & recognition, and reproduction. Several technologies are necessary for each step of the workflow and some important technologies are shown in Fig. 1. In the following sections, five XR-related technologies (video capture/rendering, motion capture/sensing, audio capture/reproduction, display, latency compensation) we have developed are introduced.

2.1 Video Capture/Rendering

We have developed a volumetric video capture technology to capture the physical world as 3D video data and enable an immersive viewing experience from any perspective at the playback [2].

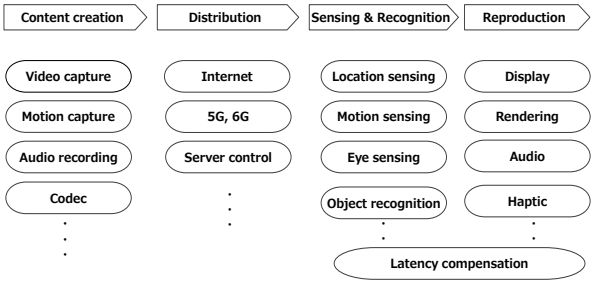

Fig. 1 Immersive XR workflow and technologies

The process flow is shown in Fig. 2. At the capture stage, objects are shot using several cameras arranged around them. Here, all cameras are synchronized with each other. Then, the 3D model is reconstructed by combining captured video data. At this stage, only the geometry without using texture data is constructed. At the rendering stage, the texture and color data of camera images are mapped on top of the objects. Finally, users can enjoy volumetric video through devices such as TVs, light field displays, and head-mounted displays.

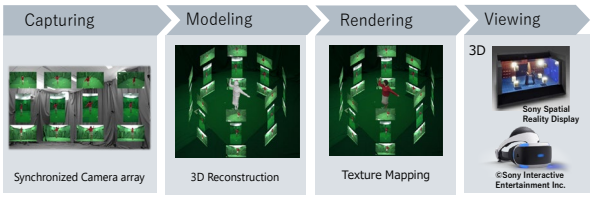

Fig. 2 The workflow of volumetric capture

2.2 Motion Capture/Sensing

Motion capture is a technology for digitizing persons or objects in the real world so that a computer can handle them as data. With this technique, a computer-graphic character motion can be reproduced in a more realistic way. We have been using motion capture technology in film, animation, and game content creation processes.

Most motion capture systems, however, require studio equipment and a special suit to wear. To address the issue, we have developed a small wearable motion sensor that enables to capture or sense the motion anywhere either

Fig. 3 Wearable motion capture

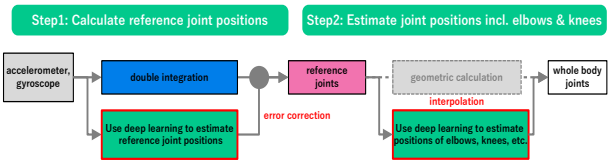

Fig. 4 The estimation process of joint positions using the wearable motion sensors

indoors or outdoors (Fig. 3) [2]. The sensor comprises an accelerometer and gyroscope.

The operating principle is shown in Fig. 4. Six motion sensors are attached to the body, one on the head, one on the waist, two on each of the wrists and ankles. In step 1, the reference joint positions such as wrists and ankles where the sensors are attached are calculated. We used the deep learning technique to correct drift errors which are caused by double integration of sensor data. In step 2, the joint positions such as elbows and knees where no sensors are attached are estimated. We used a deep learning technique again to express natural postures a human can take. By combining signal processing and machine learning techniques, we successfully estimate whole-body joint positions using six small motion sensors.

2.3 Audio Capture/Reproduction

The audio effect has an enormous impact on immersive experiences. To realize the sense of presence, we have developed our own object-based 360 Spatial Sound technologies. We named the new experience “360 Reality Audio” [3].

There is a 3D surround audio technology that can express the location of sound on horizontal planes around the user’s head within a certain vertical range. However, the 360 Reality Audio gives artists and creators a new way to express their creativity from individual instruments to an audience with dynamic control of every sound within a 360 spherical sound field.

Two core technologies enable the immersive audio experience. One is the coding technology of sound objects. It is called 360 Reality Audio music format that can maintain the sound quality even with the location data. MPEG-H 3D Audio playback devices play 360 Reality Audio formatted music contents. The other is the personalized Head-Related Transfer Function (HRTF) technology. The HRTF is a response that characterizes how our ear receives a sound from a certain point in space. As everyone has a unique size and shape of head, ears, ear canals, each one of us has a unique HRTF. We developed the technology to generate personalized HRTF

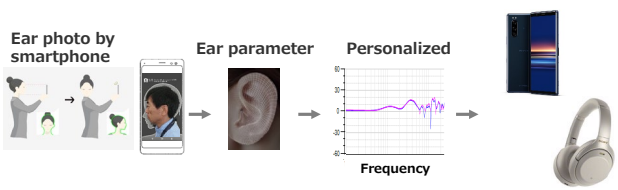

Fig. 5 Personalized HRTF technology

coefficients for headphone listening by analyzing photos of individual ears. (Fig. 5)

The workflow from the content creation to playback is shown in Fig. 6. Content creation consists of recording, editing, and encoding processes. The recording process is almost the same as the current process and archive audio files can also be utilized for the 360 Reality Audio.

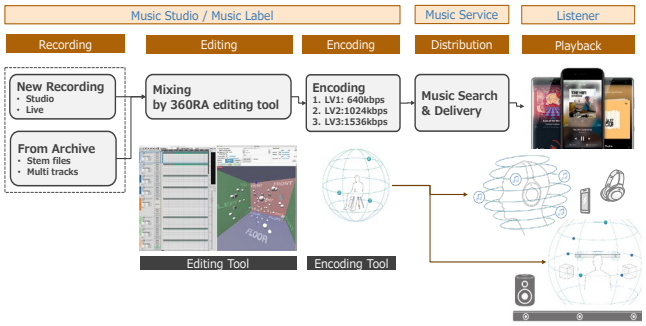

Fig. 6 The workflow of 360 Reality Audio

2.4 Display

Several XR displays have been proposed. For example, the CAVE system [4] provides an immersive panoramic view that surrounds users with three to six display walls. While the Cave system provides XR experience, it is limited to panoramic XR experience at pre-determined places. To overcome the issue, we have developed AR/MR near-eye displays which provide both panoramic and volumetric XR experience in a variety of

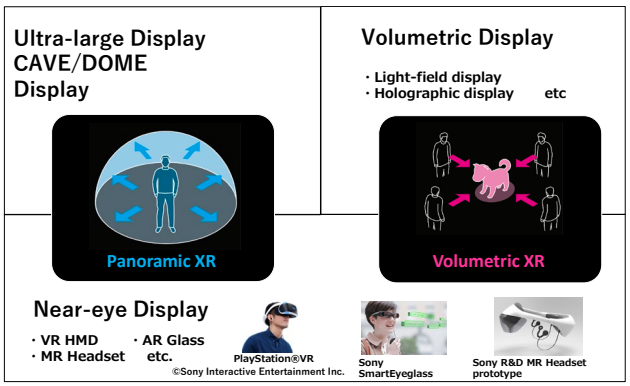

Fig. 7 Panoramic XR and volumetric XR

places [5]. (Fig.7) We employed the retinal scan approach as it has a high luminance capability thanks to

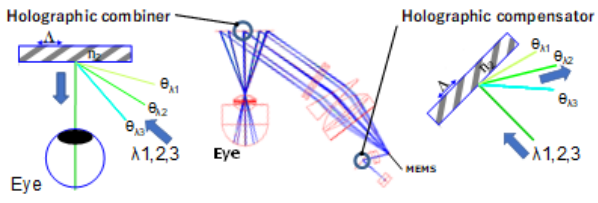

Fig. 8 Optical configuration of the retinal scan display

its high optical efficiency using laser light sources and the vergence-accommodation conflict (VAC) free nature. Those advantages are especially important for optical see-through (OST) displays for visual comfort.

The part of a configuration of the display is shown in Fig. 8. It consists of a MEMS scanner, a holographic combiner, and a holographic compensator. By applying a holographic compensator, the diffraction color dispersion caused by the holographic combiner can be canceled. The R&D prototype of the retinal scan display is shown in Fig. 9. It has a 1280 x 720 resolution, a 47° field of view, 85% see-through transparency, and up to 10,000 cd/m^2 luminance.

Fig. 9 Retinal scan display prototype

2.5 Latency compensation

For AR/MR near-eye displays such as OST retinal scan displays, accurate spatial registration between virtual objects and the physical world is crucial for a sense of presence. The registration error is mainly due to the system latency introduced by a user's head motion. In OST systems, it is more difficult to align virtual and physical images than in video see-through systems because users see the physical world with no delay. Therefore, the latency compensation is inevitable to achieve an acceptable registration error.

We have developed the OST near-eye display prototype shown in Fig. 10 using two inside-out cameras and inertial measurement units (IMUs) for sensors. The custom SoC was developed for low latency signal processing. The entire system data flow is shown in Fig. 11. We employed the time-warp technique to minimize the motion to rendering latency. The time-warp is a technique to generate the latest virtual image by transforming a pre-rendered image based on the latest head pose. The image transformation is done in a 2D plane by shifting, skewing, expanding, and shrinking an original image according to the latest head pose. This simplifies the signal process and minimizes the latency.

Fig. 10 Photo of the OST MR near-eye display prototype

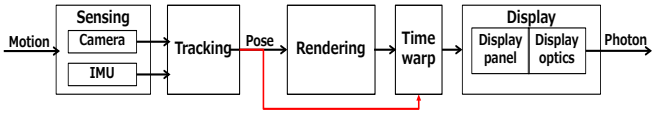

Fig. 11 System data flow for latency compensation using a time-warp technique

We measured registration errors of virtual objects while rotating the near-eye display along the roll axis with a sine wave whose period is 5 Hz and amplitude is ±1°. The measured result is shown in Fig. 12. The blue line is the real object position, and the red cross is the virtual object position. When latency compensation is enabled, the registration error improves from 0.832° to 0.074°, which would be small enough for many use cases.[6]

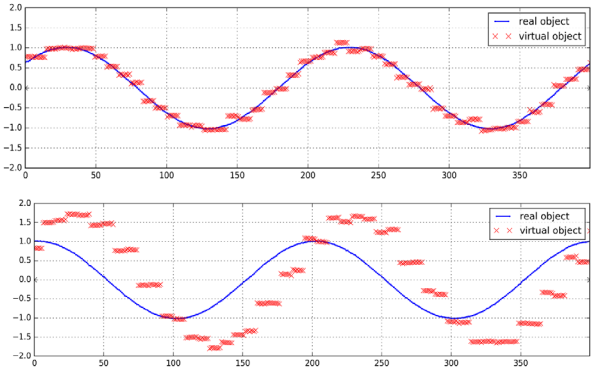

Fig.12 Registration errors with and without the latency compensation

3 Conclusion

The immersive XR has a good potential for the next growth industry. To deliver the immersive XR experience, technologies boosting the reality expression and real-time interaction need to be developed. Moreover, content creation, distribution, sensing & rendering, and reproduction technologies need to be highly aligned with each other. Moreover, as new content and technologies have to be integrated to deliver unmatched experiences to users, it is expected that the XR business ecosystem is created among content companies and technology companies joining this challenging but exciting field.

References

[1] M. Ball, "The Metaverse: What It is, Where to Find it, Who Will Build it, and Fortnite," https://www.matthewball.vc/all/themetaverse

[2] "Empower creators' creativity", Sony Technology Day Report Vol. 3 https://www.sony.com/en/SonyInfo/technology/activities/SonyTechnologyDay2019_demo2/

[3] Journal of the ITU Association of Japan, ITU Journal, Vol. 51 No. 8, pp. 11-14 (2021)

[4] K. Akutsu, et al., "A compact retinal scan near-eye display", ACM SIGGRAPH 2019 Emerging Technologies, Article No. 2, pp. 1-2 (2019)

[5] C. Cruz-Neira, et al., "Surround-screen projection-based virtual reality: the design and implementation of the CAVE," Proceedings of SIGGRAPH 1993, pp. 135–142 (1993)

[6] H. Mukawa, et al., "Optical See-Trough AR HMD with Spatial Tracking," Proceedings of SPIE - The International Society for Optical Engineering 2020, Vol. 11520, pp. 35-36 (2020).