Overview of the Volumetric Video Capturing System for Immersive Media

Jaeyeol Choi, Jong-Beom Jeong, Soonbin Lee, Eun-Seok Ryu
Department of Computer Science Education, Sungkyunkwan University, Seoul, Republic of Korea
E-mails: {jaychoi, uof4949, soonbinlee, esryu}@skku.edu

***Abstract*—Immersive videos shown through virtual reality (VR) devices provide users with realistic viewing experiences. This paper focuses on the capturing process in an immersive video producing workflow. First, it introduces depth estimating technology and models of RGB-D cameras, as well as the general components of VR studios. Next, representative VR studios, such as studio of *Microsoft* and *Fraunhofer Heinrich Hertz Institute,* are introduced with a comparison of the used equipment. Furthermore, along with services such as *VCL3D* and *Livescan3D*, which provide a portable and flexible capturing environment, a system that produces videos for VR with only a single device is introduced.**

Keywords—VR studio, 3D video capture, Depth camera, Volumetric Video

I. Introduction

In recent times, companies working in virtual reality (VR) domain have faced difficulties in meeting their customers' demand for immersive videos. To provide high-quality immersive videos, the development of the overall technology, ranging from capturing to displaying, must be accompanied. With improved head mounted displays (HMDs) technology, devices such as *Oculus Quest* or *HTC Vive* can display omnidirectional 3D images fluently [1]. Thus, video-acquisition, compression, transmission, and streaming technologies have become important. In relation to VR technology, effective streaming techniques for immersive videos have been widely researched [2], [3], and the standardization studies on immersive video coding have been conducted [4], [5]. This paper focuses on the process of video acquisition. Broadly speaking, video acquisition can be done through two approaches: the first involves artists creating 3D images through graphic work, and the second involves capturing natural content from the real world. This paper introduces capturing systems for natural VR content.

The volumetric videos differ from typical 360-degree videos in that they provide depth information from all perspectives along with free-viewpoint. Furthermore, volumetric videos can be used in a range of terms such as free-viewpoint video (FVV) and 3D video. In particular, the six-degrees-of-freedom (6DoF) video not only supports the turn of the head, but also provides changes in images according to the movement [6]. Since previous paper reviewing the overall VR video capturing system has not yet been published, this paper summarizes the VR content capturing facilities used in various cases. First, the components of the VR studio are presented focusing on the camera devices, after which the cases of the different uses of the VR studio will be introduced in three stages according to its size.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1F1A1074935).

II. Components of VR Studios

The first part of this section describes the equipment typically used in VR studios. Continually, the depth estimating technologies to capture volumetric video as well as RGB-D camera models are introduced.

A. VR Studio Equipment

The VR video capturing systems commonly consist of cameras (typically capable of capturing color and depth data), microphones, and illumination, as shown in Fig. 1. Various pieces of equipment, including cameras, are mounted on an iron frame or tripod surrounding the main object. In addition, the studio is often encircled by a green background, intended to enhance the discrimination between the subject and the background.

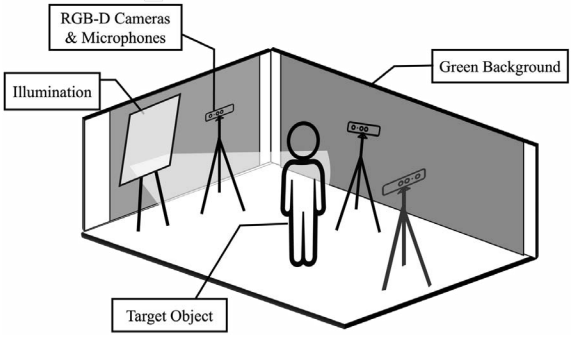

Fig. 1. Example of a virtual reality (VR) studio components

B. Factors required for immersive video acquisition

In addition to color data, depth data acquisition is essential for creating high-quality volumetric videos. Technologies such as *structured light*, *time-of-flight (ToF)*, *Stereo Depth*, and *LiDAR* are utilized to acquire the depth data of the object. The structured light technique estimates depth by using the geometric difference between original pattern and deformed pattern, which is reflected in an object. Besides, ToF estimates depth by measuring the time taken by the infrared light to travel from the sensor to the object and back to the sensor [7]. After the color and depth data are captured, pre-processing tasks, such as calibration and bias correction are mostly performed, and thereafter, a point-cloud is generated [8].

Fig. 2. Intel RealSense L515

C. Cameras used for capturing immersive video

Primarily, RGB-D cameras are used to record color and depth data simultaneously, or RGB and depth measurement cameras are used separately. The former method is mainly used to reduce the number of cameras for the sake of convenience, whereas the latter is used to create high-resolution content. The frequently used or recently released RGB-D cameras are listed in Table 1 (created using details provided in [9]) and are explained in more detail in [10] and [11]. Fig. 2 is a picture of *Intel RealSense L515*, which uses the LiDAR technology for measuring depth.

TABLE 1. SPECIFICATIONS OF COMMONLY USED RGB-D CAMERA MODELS

Model	RGB Resolution (30fps)	Depth Resolution (30fps)	Size (mm)	Depth Estimate
Intel RealSense D435	1920 × 1080	1280 × 720	90 × 25 × 25	Stereo Depth
Intel RealSense D455	1280 × 800	1280 × 720	124 × 26 × 29	Stereo Depth
Intel RealSense L515	1920 × 1080	1024 × 768	61 × 61 × 26	LiDAR
MS Kinect v2	1920 × 1080	512 × 424	249 × 66 × 67	ToF
MS Kinect Azure	3840 × 2160	640 × 576	103 × 39 × 126	ToF
Azure Kinect DK	3840 × 2160	1024 × 1024	103 × 39 × 126	ToF
Astra S+	1920 × 1080	640 × 480	149 × 28 × 29	Structured Light
Astra Stereo S U3	1920 × 1080	1280 × 800	65 × 23 × 12	Stereo Depth

III. CASES OF IMMERSIVE VIDEO CAPTURING SYSTEM

This section introduces the specific case of a volumetric video capturing system from previous research in three parts, categorized according to the scale of the system.

A. Cases of High-cost and High-quality Facilitiy

Microsoft introduced an end-to-end solution to create high-quality FVV [12]. The studio consists of 106 cameras. Half of them are *STC-CMC4MCL* RGB cameras, and the others are *STC-CMB4MCL* IR cameras, which are used to sense the volume of the objects. With the exception of 10 cameras with attached overhead, most cameras are evenly mounted on eight wheeled structures. The green background and illumination (including the infrared light source), which is shown in Fig. 3, surround the target object to enhance the video quality of the reconstructed form. The outputs of the six cameras are delivered to one computer and then combined into the alignment of disks. After pre-processing, the point-cloud indicating depth data is generated, after which the meshing process is fulfilled using hull-constrained *Poisson surface reconstruction (PSR)*, which is main contribution of the study, that reduces the artifacts of surface reconstruction.

Fig. 3. The volumetric capturing system used by Microsoft

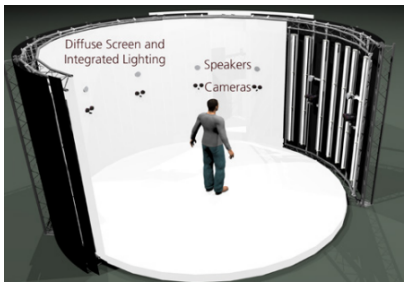

Fig. 4. The structure of volumetric capturing system used by Fraunhofer

Fig. 5. The capturing stage used in the production of *HUMAN4D* dataset

The next example is the VR studio of the Fraunhofer Heinrich Hertz Institute, which is presented in [1]. Fig. 4 shows the concept of the overall system. The cylindrical structure is surrounded by 32 cameras and LED panels organized with semitransparent tissue that provides equal light in all directions. The 20-megapixel cameras consist of 16 stereo pairs that help capture high-definition stereoscopic images for both eyes. Moreover, depth estimation is performed by comparing 3D patches between the images from the two cameras in pairs. A remarkable feature of the system is the white light-emitting background which replaces the commonly used green background. The reason for this is to provide better conditions for the re-lighting of 3D models. Using a volumetric video capturing system and the utilization of GPU resources, this research offers a workflow that produces high-quality free viewpoint volumetric video in relatively lesser time.

The following case is *HUMAN4D*, a service that provides diverse 4D datasets of human activities [13]. The peculiarity of this example is that the volumetric capturing is performed with motion capturing. The human motion data sample was captured in the in the studio of *Artanim Foundation*, as shown in Fig. 5. A total of 24 *Vicon MTX40S* motion cameras were used to obtain motion data, and four *Intel RealSense D415* cameras were used to obtain depth data to create mesh information. One or more actors performed their movements in an approximately 4m×4m space for each activity recording (e.g., running, talking, basketball dribbling, etc.).

B. Cases of Mid-cost and Flexible Facility

The volumetric video acquisition system, which is not complicated and can be easily utilized by non-experts, was introduced in [14]. The *VCL3D* service simplifies the setup and post-processing procedure after obtaining image data by providing customized software. As shown in Fig. 6, the proposed system mainly consists of multiple tripod modules conceptually called *eye* and central control computer called *orchestrator*. *Intel RealSense L415* is used as RGB-D sensor mounted on the *eye*. (*Microsoft Kinect Azure* is also supported according to the *VCL3D* website). Optionally, an edge device (such as *Intel NUC*) is also mounted on the *eye* with a camera

and carries out encoding and serializing operations. The acquired data from multiple *eye*s are transmitted to the *orchestrator* through the *broker* (also optional), and storing or visualizing operations occur at the main workstation. The infrastructure offers flexibility to users by simplifying the intricate capturing and post-processing procedure into a brief way of just setting up an adjustable number of devices and using the provided software.

An open-source service that includes a model for creating real-time 3D reconstruction (colored point-cloud) from multiple *Kinect v2* or *Azure Kinect* devices was introduced in [15]. The system called *client* consists of Kinect device allocated to an individual computer. Each *client* is connected to the main *workstation*. The *client* produces its own 3D data by the main workstation's control that synchronizes each frame transmission. After filtering each point cloud in each client, the main workstation merges the data delivered from the clients.

C. Cases of Single-device Capturing System

VR studios can be made up only with a single 360-degree personal device, at no great cost. In the case of individuals who are filming while moving outside or those who are not in a situation to set up multiple sensors, a camera capable of capturing depth and 360-degree video can be used. According to [16], 360-degree cameras can be categorized into three types: cameras with fish-eye lens, cameras with dual fish-eye lenses to cover opposite positions (e.g., *Insta360 ONE*), and cameras with more than two lenses to capture more accurate images (e.g., *GoPro Omni*). The most advanced form is a device that consists of 16 cameras and supports 6DoF capture and view synthesis, which was introduced in [17]. Examples of the actual use of these 360-degree capturing devices were presented in [18] and [19].

A system that can produce volumetric 6DoF video with only a common 360-degree camera that does not support depth estimation was introduced in [20]. The overall architecture consists of an off-line part that constructs 3D geometry information of the picture and real-time playback that renders an appropriate angle of video according to the user's view. For the off-line part, the input value is a plain 360-degree monoscopic video. As shown in Fig. 7, two main algorithms are used to convert the acquired 360-degree video into a 6DoF video. The structure-from-motion (SfM) algorithm leads to the restoration of the camera parameters and sparse 3D geometry. The next step, called dense reconstruction algorithm, computes the depth map and merges it into a 3D point-cloud. During playback, the frame is warped for both eyes according to the user's view to provide real-time 6DoF stereoscopic streaming.

IV. Conclusion

This paper introduces various types of VR contents capturing systems. Despite the requirement of high cost and high degree of human expertise, the studios introduced and discussed in III-A are efficient at creating high-quality content that provides high video resolution. The studios introduced in III-B are not expensive, provides flexibility, and available for non-experts, but they have disadvantage of being dependent on particular software service provided. Single machine systems introduced in III-C provide a capturing environment that is not limited to a specific location; however, they have limitation in the quality of the generated images or are accompanied by many post-processing stages. This paper helps users to choose appropriate volumetric video capturing system according to their practical need. In future, research on establishing VR studio that supports 6DoF immersive video capturing will be conducted.

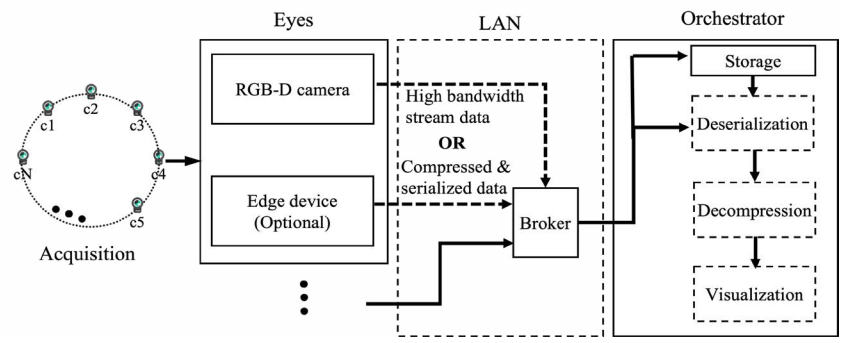

Fig. 6. The overall architecture of *VCL3D* system introduced in [14]

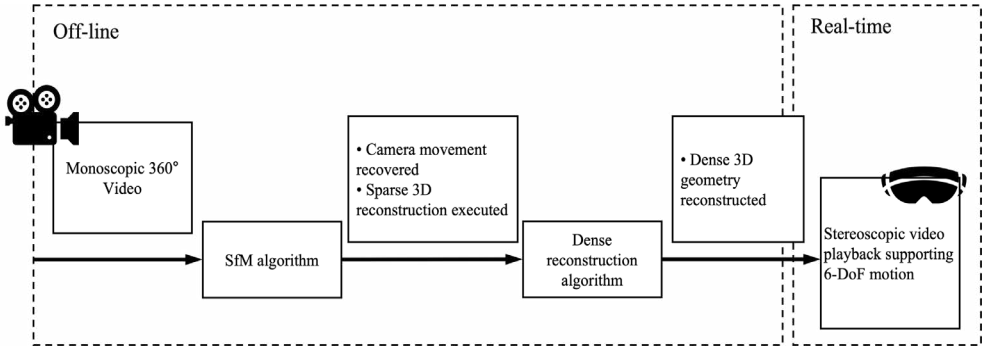

Fig. 7. Summary of the video conversion process described in [20]

REFERENCES

[1] O. Schreer et al, “Capture and 3D Video Processing of Volumetric Video,” in *2019 IEEE International Conference on Image Processing(ICIP)*, pp. 4310-4314, 2019.

[2] T. T. Le, J. B Jeong, S. Lee, J. Kim, and E. S Ryu, “An Efficient Viewport-Dependent 360 VR System Based on Adaptive Tiled Streaming,” 2021.

[3] J. B. Jeong, S. Lee, D. Jang, and E. S. Ryu, “Towards 3dof+ 360 video streaming system for immersive media,” in *IEEE Access, 7, 136399-136408*, 2019.

[4] S. Lee, J. B. Jeong, and E. S. Ryu, “Atlas level rate distortion optimization for 6DoF immersive video compression,” in *Proceedings of the 32nd Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 78-84, 2022.

[5] J. B. Jeong, S. Lee, and E. S. Ryu, “Sub-bitstream packing based lightweight tiled streaming for 6 degree of freedom immersive video,” in Electronics Letters, 57(25), pp. 973-976, 2021.

[6] P. Atsikpasi, and E. Fokides, "A scoping review of the educational uses of 6DoF HMDs," in *Virtual Reality, 1-18*, 2021.

[7] A. Kadambi, A. Bhandari, and R. Raskar, " 3d depth cameras in vision: Benefits and limitations of the hardware," in *Computer vision and machine learning with RGB-D sensors*, pp. 3-26, 2014.

[8] E. Zell et al, " Volumetric Video-Acquisition, Compression, Interaction and Perception," 2021.

[9] H. Yoon, M. Jang, J. Huh, J. Kang, and S. Lee, “Multiple Sensor Synchronization with theRealSense RGB-D Camera,” in *Sensors, 21(18), 6276*, 2021.

[10] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1-10, 2017.

[11] J. Han, L. Shao, D. Xu, and J. Shotton, " Enhanced computer vision with microsoft kinect sensor: A review," in *IEEE transactions on cybernetics, 43(5):1318-1334*, 2013.

[12] A. Collet et al, “High-Quality Streamable Free-Viewpoint Video,” in *ACM Transactions on Graphics (ToG), 34(4), 1-13*, 2015.

[13] A. Chatzitofis et al, “Human4d: A human-centric multimodal dataset for motions and immersive media,” in *IEEE Access, 8*, pp. 176241-176262, 2020.

[14] V. Sterzentsenko et al, “A low-cost, flexible and portable volumetric capturing system,” in *2018 14th international conference on signal-image technology & internet-based systems (SITIS)*, pp. 200-207, 2018.

[15] M. Kowalski, J. Naruniec, and M. Daniluk, "Livescan3d: A fast and inexpensive 3d data acquisition system for multiple kinect v2 sensors," in *2015 international conference on 3D vision*, pp. 318-325, 2015.

[16] H. Ai et al, " Deep Learning for Omnidirectional Vision: A Survey and New Perspectives," in *arXiv preprint arXiv:2205.10468*, 2022.

[17] A. P. Pozo et al, "An integrated 6DoF video camera and system design," in *ACM Transactions on Graphics (TOG), 38(6), 1-16*, 2019.

[18] W.T. Neale, T. Terpstra, N. Mckelvey, and T. Owens, " Visualization of Driver and Pedestrian Visibility in Virtual Reality Environments," in *SAE Technical Paper, 01-0856*, 2021.

[19] Z.Yang, D. Xiang, and Y. Cheng, " VR Panoramic Technology in Urban Rail Transit Vehicle Engineering Simulation System," in Ieee Access 8 (2020): 140673-14068, 2020.

[20] J. Huang, Z. Chen, D. Ceylan, and H. Jin, " 6-DOF VR Videos with a Single 360-Camera," in *2017 IEEE Virtual Reality (VR)*, pp 37-44, 2017.