QoS-Aware Point Cloud Streaming of Wild Animals/Humans for Interactions in Virtual Space

Hiroki Ishimaru[1], Yugo Nakamura[2], Manato Fujimoto[3,4], Hirohiko Suwa[1,4], Keiichi Yasumoto[1,4]
[1]Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan
Email: {ishimaru.hiroki.ie3, h-suwa, yasumoto}@is.naist.jp
[2]Kyushu University, Fukuoka, Fukuoka 819-0395, Japan
Email: y-nakamura@ait.kyushu-u.ac.jp
[3]Osaka Metropolitan University, Osaka, Osaka 599-8531, Japan
Email: manato@osaka-cu.ac.jp
[4]RIKEN, Chuo-ku, Tokyo 103-0027, Japan

***Abstract*—3D applications such as VR and AR are attracting increasing commercial attention, and point cloud video is expected to be one of the most suitable representations for real-time applications due to its simplicity and versatility. However, point cloud data is large in size and difficult to stream in a mobile network environment with limited bandwidth. Therefore, a method for streaming point clouds with low bandwidth consumption while maintaining the quality of the user experience is needed. In this paper, we present a point cloud streaming method of real-space objects such as humans and animals for real-time 3D reconstruction in VR space. The system uses a depth camera to scan a human or animal, divides the point cloud into parts of the body, and then controls the quality of the point cloud (i.e., resolution and frame rate) for each part in real-time according to the object's motion and context. This enables point cloud streaming with limited resources (computational and network resources) and maximizes the user's quality of experience. We exhibit a series of systems that enhance the user experience in remote communication in realistic environments and scenarios while maintaining interactivity between real-space objects and remote users.**

***Index Terms*—Real-time communication, 3D point cloud, QoE, Streaming, VR**

I. Introduction

The impact of COVID-19 has prompted many aspects of our daily lives to go online. On the other hand, it is sometimes difficult to go online in situations where the gap between the real and virtual worlds is large. Tourism is an example of such a situation, but if online access to such situations is achieved, it will be possible to promote safe economic activities while preventing infection and providing new options for our daily life. For this reason, there is demand for a system that allows users to interact with people and animals in a remote place while moving freely within the place, even while participating remotely, in situations such as tourism, where experience and sensation in real space are important.

Many 2D video-mediated methods have been proposed for such a system. For example, there is a virtual tour system that extends an existing online conferencing system by applying 360-degree video, but the user's degree of freedom is low, and the experience is far from realistic [1]. Under these circumstances, 3D video has been attracting attention in recent years as a means of providing content that is closer to reality in applications such as VR/AR. However, the large data size of 3D video makes it impractical to distribute the raw data in the current infrastructure system. Therefore, an efficient streaming method is required and has become a challenge [2].

This paper deals with point clouds, which are a simple and versatile media for representing 3D images. There are many studies and approaches for streaming point clouds, such as encoding/decoding of point cloud data, tiling, and transmission optimization using view angle prediction [3], [4]. However, most of them focus on the compression of point clouds and the prevention of unnecessary data transmission. There are also some studies that evaluate the impact of these quality control mechanisms on user QoE (Quality of Experience), but there are not many of them [5]–[7].

In this paper, we propose an approach that extends point cloud quality control mechanisms to context-aware and adaptive quality control of objects to achieve point cloud streaming over limited resources while maintaining high user QoE.

II. System Overview

A. Approach

To realize online tourism, it is necessary to convert real space with various tourist objects into real-time data and reconstruct it in virtual space. Still, it is difficult to acquire and use all objects in real space in real-time due to computational and network resources in mobile environments. Lee *et al*. proposed an approach in which a portion of the real space is cut out, and only the necessary objects are reflected in real-time, while static data is used for the other objects [8]. We adopted this approach in our system, dividing the real space into the objects that the remote participants wanted to project and the surrounding objects that would serve as their backgrounds. Fig. 1 shows this approach. The necessary objects and the surrounding objects are called OOI (Object Of Interest) and AO (Ambient Object), respectively.

In addition, we considered that using cameras and communication infrastructure pre-installed at the site would be

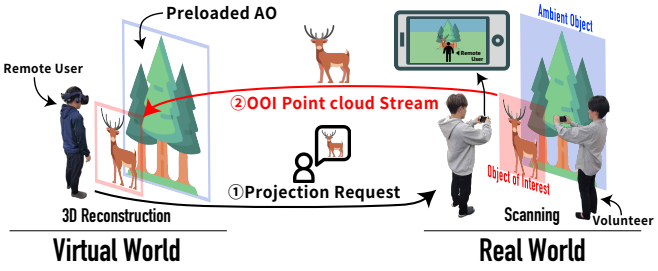

Fig. 1. Our approach in streaming real space objects to virtual space

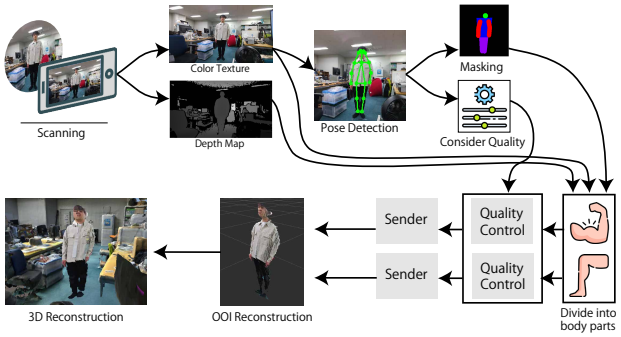

Fig. 2. System Data Flow

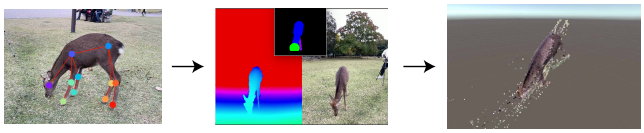

Fig. 3. In case of animals: masks are generated from animal pose estimation results to get a point cloud

inappropriate in terms of ubiquity and versatility. Since there are countless tourist spots, the system must ultimately be flexible enough to accommodate them. Therefore, we assumed that volunteers would be present in the vicinity of the target real space, and OOI point cloud scanning and transmission would be conducted with their cooperation, using their mobile terminals, such as smartphones and tablets. Under these conditions, even when sending only OOI point clouds, there can be delays that degrade the QoE due to limited communication bandwidth and computational resources. Hence, we propose a method to dynamically control the quality of the OOI point cloud to keep the resources used under the limit. Unlike the previous tiling approaches [6] which divide the point cloud object evenly, the proposed method recognizes the object's attribute and divides the point cloud at the body part scale to perform finer-grained quality control of the object to maximize the user experience while reducing bandwidth consumption.

B. System Design

Based on the proposed approach, targeting tourism application, we propose a system that enables remote and local users to share space by scanning and transferring objects in real space and reconstructing them in virtual space in real-time, without requiring fixed infrastructure. As shown in Fig. 1, a volunteer at a sightseeing spot performs scanning and data transmission of local objects (OOI) in the real space in real-time at the request of a remote user. The real space is reconstructed in the remote user's virtual space by using the transmitted point cloud data of the local objects (OOI) together with the background data (AO) prepared in advance, thereby sharing the space of the tourist site is realized.

To reduce bandwidth consumption of the point cloud data (OOI) handled by the system, we propose a dynamic quality control based on body part movements and incorporate it in the system. Based on the techniques used in video compression standards such as H.265 for 2D video and tiling approaches for 3D point clouds, we reduce the total amount of point cloud data by dynamically varying the resolution and frame rate of each body part according to context: high frame rate/low resolution when the object part is moving significantly, and low frame rate/high resolution when the part is nearly stationary. Although the absolute quality of the point cloud data (OOI) is lower than the raw data after applying these processes, bandwidth consumption is optimized by controlling the quality so as not to affect the user's QoE as much as possible.

The following is a description of the prototype system we developed to evaluate the feasibility of the proposed system. Fig. 2 shows a data flow of the system. The system is built on ROS and uses Unity for 3D reconstruction.

First, we explain how to obtain the AO and OOI to reconstruct the real space in a virtual environment. First, we obtain the AO by using tools such as a 3D scanning application that utilizes LiDAR installed in the iPhone, or a camera that can capture 360-degree images. In recent years, 3D scanning of various locations has been progressing, so scanning of AO objects may not be necessary in the future.

Next, we describe the actual method of acquiring OOI point clouds for transmission and real-time reconstruction. As mentioned above, we assume that mobile devices will be used to acquire the OOI point clouds, as Apple's release of LiDAR sensors for the iPhone and iPad has made 3D scanning easier for many people. The depth and texture of the OOI are continuously acquired using a depth camera and treated as a color point cloud. Our system uses Azure Kinect as well as iPhone for evaluation.

After scanning an OOI object, quality control is dynamically performed according to its context. First, as shown in Fig. 2, pose estimation is performed on the 2D color video input from the camera to segment body parts and track their movement. For the pose estimation, a trt-pose estimation model is used for human pose estimation [9], [10], and a model trained by DeepLabCut is used for animals (deer) pose estimation [11], [12], as shown in Fig. 3.

The quality of the point cloud is determined by the value of the displacement of the corresponding body part. The resolution of the point cloud is adjusted by resizing the original input (color image, depth map: 720P) using nearest neighbor

TABLE I
SPECIFICATIONS OF POINT CLOUD QUALITY OUTPUT BY THE SYSTEM

Compression Rate (original:100%)	Head	Body	Arms	Legs	All
25%	0.28Mbps (13FPS)	2Mbps (13FPS)	2.3Mbps (14FPS)	1.5Mbps (14FPS)	6.1Mbps
50%	0.9Mbps (12FPS)	5.5Mbps (9FPS)	6.1Mbps (9FPS)	4.8Mbps (12FPS)	17.3Mbps
75%	1.8Mbps (9FPS)	6.5Mbps (5FPS)	6.4Mbps (5FPS)	5.6Mbps (6.5FPS)	20.2Mbps

completion (three steps: 25%, 50%, and 75%), and the frame rate is switched by adjusting the transmission timing. This process is performed for each body part. Table I shows the bandwidth consumption of a sample human video sequence when our system controls the quality. In this sample, the bandwidth consumption (i.e., the maximum value) is 20.2 Mbps when all body parts are set to 75% quality, which is within the possible transmission range with the existing infrastructure.

The resulting point cloud data for each body part is then sent to Unity as a ROS topic and reconstructed in 3D together with the AO prepared in advance. To ensure water tightness in the representation of the point cloud, the size of the points is dynamically changed according to the quality of the point cloud. The reconstructed tourist attraction space is presented to the user through a VR-HMD. Our evaluation system is designed to operate in a mobile environment using edge devices and smartphones, but power consumption is not a priority at this stage. We believe that the effective use of GPUs and the optimization of software will reduce power consumption.

III. SYSTEM DEMONSTRATION

In the demonstration, our prototype system will be set up at the venue to show real-time quality control according to the context of objects and the reconstructed real space after the transmission in VR space. We will also show an example without QoS control for comparison. We will also show a scenario in which interaction with people and animals is assumed.

The hardware configuration for the real-time processing demonstration is shown in Fig. 4. Azure Kinect is connected to NVIDIA Jetson to capture object scan data and perform the dynamic quality control described above. The resulting point cloud data of each body part is sent to another PC connected via LAN, and 3D reconstruction/rendering is performed on a VR headset (HTC-Vive) connected to that PC. Each of these hardware devices mimics the mobile device in the proposed system, and we have chosen these hardware configurations for the purpose of system evaluation. With the above system demonstration, we show that our system is feasible.

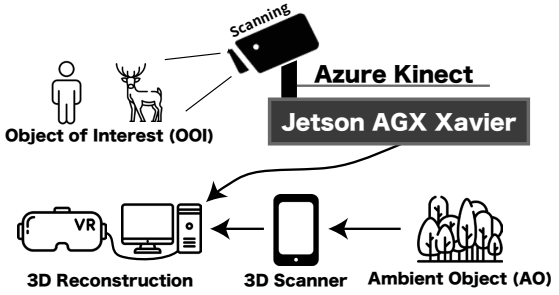

Fig. 4. Hardware Configuration (for demonstration)

ACKNOWLEDGMENT

This work is supported in part by the Commissioned Research of National Institute of Information and Communications Technology (NICT) in Japan, Contract No. 222C03 and JSPS KAKENHI Grant Number JP21H03431.

REFERENCES

[1] A. Nassani, L. Zhang, H. Bai, and M. Billinghurst, "Showmearound: Giving virtual tours using live 360 video," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–4.

[2] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li, and Y. Ji, "Point cloud video streaming: Challenges and solutions," *IEEE Network*, vol. 35, no. 5, pp. 202–209, 2021.

[3] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li *et al.*, "Emerging mpeg standards for point cloud compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2018.

[4] M. Hosseini and C. Timmerer, "Dynamic adaptive point cloud streaming," in *Proceedings of the 23rd Packet Video Workshop*, 2018, pp. 25–30.

[5] L. Wang, C. Li, W. Dai, S. Li, J. Zou, and H. Xiong, "Qoe-driven adaptive streaming for point clouds," *IEEE Transactions on Multimedia*, 2022.

[6] J. Li, C. Zhang, Z. Liu, W. Sun, and Q. Li, "Joint communication and computational resource allocation for qoe-driven point cloud video streaming," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[7] J. Van den Bulck and S. Eggermont, "Media use as a reason for meal skipping and fast eating in secondary school children," *Journal of human nutrition and dietetics*, vol. 19, no. 2, pp. 91–100, 2006.

[8] Y. Lee, B. Yoo, and S.-H. Lee, "Sharing ambient objects using real-time point cloud streaming in web-based xr remote collaboration," in *The 26th International Conference on 3D Web Technology*, 2021, pp. 1–9.

[9] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, 2018. [Online]. Available: https://www.nature.com/articles/s41593-018-0209-y

[10] T. Nath*, A. Mathis*, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis, "Using deeplabcut for 3d markerless pose estimation across species and behaviors," *Nature Protocols*, 2019. [Online]. Available: https://doi.org/10.1038/s41596-019-0176-0

[11] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.