A Dual Rig Approach for Multi-View Video and Spatialized Audio Capture in Medical Training

Joshua Maraval, Bangning Wei, David Pesce, Yann Gayral, Meriem Outtas, Nicolas Ramin, Lu Zhang

A Dual Rig Approach for Multi-View Video and Spatialized Audio Capture in Medical Training

Joshua Maraval
IRT b<>com
Rennes, France
joshua.maraval@b-com.com

David Pesce
IRT b<>com
Rennes, France
david.pesce@b-com.com

Yann Gayral
IRT b<>com
Rennes, France
yann.gayral@orange.fr

Bangning Wei
IRT b<>com
Univ Rennes
Rennes, France
bangning.wei@b-com.com

Meriem Outtas
Univ Rennes, INSA Rennes
CNRS, IETR-UMR 6164
Rennes, FRANCE
meriem.outtas@insa-rennes.fr

Nicolas Ramin
IRT b<>com
Rennes, France
nicolas.ramin@b-com.com

Lu Zhang
Univ Rennes, INSA Rennes
CNRS, IETR-UMR 6164
Rennes, FRANCE
lu.zhang@insa-rennes.fr

***Abstract*—We present a multi-view camera and spatialized audio microphone capture system designed for computer vision applications in free navigation immersive experiences. We propose a dataset of two long and complex in-situ training situations in the medical field. The scenarios in the dataset feature precise gestures for the learner to reproduce during complex situations with multiple simultaneous visual and auditory cues important for training. 3D computer vision techniques are used to reconstruct a 4D scene model from a set of videos to render novel views from unseen viewpoints. However, the quality of the rendered objects is directly dependent on the density of coverage by reference views. To ensure maximum QoE, we propose a dual rig of cameras, a central rig that captures the details of the gesture zone of the training scenarios and a peripheral rig that captures the environment of the room and the interactions occurring around the gesture zone. The central rig provides dense coverage of the central content, facilitating high-quality reconstruction on novel views of the captured gestures. Recordings include audio interactions of multiple actors, captured by ambisonic microphones spatially distributed around the scene. The captured scenes are real-world educational content for medical courses, so this dataset provides a rare opportunity to assess the QoE of volumetric video techniques on realistic content, and to compare their pedagogical capabilities with standard multi-view video content.**

***Index Terms*—database, multi-view video, spatial audio, free navigation, volumetric video, QoE**

I. Introduction

Over the past few decades, video consumption and video devices have become widespread globally. In 2014, the emergence of "consumer-grade" virtual reality headsets was followed by a growing number of 360° videos. While this marked the democratization of immersive video, users were initially confined to a fixed position within the chosen scene. By 2022, second-generation devices like augmented reality glasses, new VR headsets, LIDAR-equipped tablets, and smartphones employing advanced SLAM techniques became commercially available. These innovations enabled real-time user position estimation, unlocking three additional degrees of freedom and promising new 6-degree-of-freedom (6DoF) immersive video experiences. Nevertheless, various challenges still hinder the production, distribution, and playback of high-quality volumetric videos.

Synthetic environments represent the majority of available 6DoF navigable content. However, shaping realistic synthetic environments is difficult and costly. In 2020, the groundbreaking Neural Radiance Field (NeRF) paper [1] introduced a new way to generate high-quality free viewpoint renderings of real scenes from only sparsely captured views. Follow-up research has led to faster and more flexible methods, such as the widely used 3D Gaussian Splattings [2]. These methods can render photorealistic views in the right conditions, but are prone to very visible artifacts if the input views are too few and sparse. Dense enough array of cameras are effectively used in studio environments [4] covering all elements of the scene with many view angles. In-situ recording, on the othe hand, is more challenging as it requires restricting the number of camera used and adapting to the shape of the environment.

Many multi-view video datasets have been made available for volumetric video research, but there is a lack of datasets that cover real uses-cases of immersive displays. Especially, we denote a lack of audio-video datasets of real-world trainings for 6DoF immersive experience.

We propose an in-situ recording rig that has been validated on three different medical training scenarios. Key extracts of the trainings are publicly available for scientific research on https://volumetric-repository.labs.b-com.com/#/immersive_learning, and the full sequences will be shared on demand. In this paper, we detail our dual rig configuration: a central rig captures the scene area with precise medical gestures performed to ensure high fidelity and rendering quality. A second peripheral rig captures the training environment for immersion and to provide sufficient coverage of actions and interactions outside the gesture zone. Spatial audio is captured in two of the three scenes by 4 ambisonic

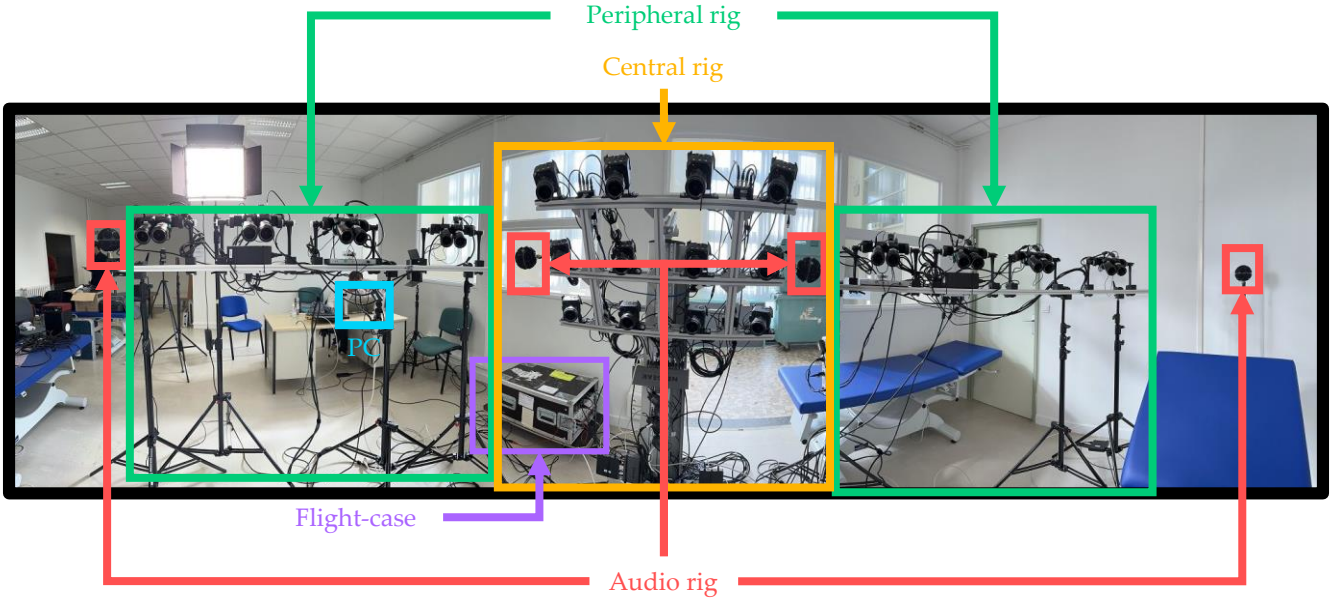


Fig. 1: Our full in-situ capture setup consists in a central camera array of 12 Panasonic BGH1 cameras, a linear peripheral camera rig. of 16 Blackmagic micro cinema cameras and an audio rig of 4 Zylia ZM-1 microphones.

microphones distributed throughout the training room.

In section II we review other datasets for volumetric video. We describe our rig design choices in section III, and the captured content in section IV. The dataset characteristics are listed in section V.

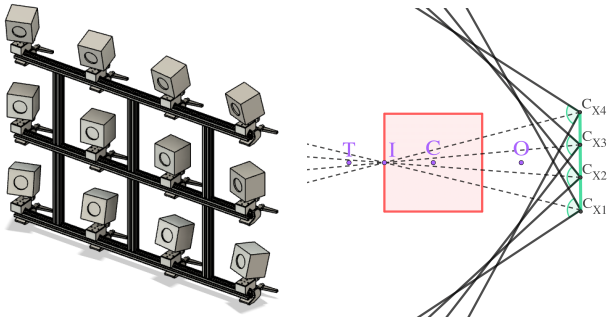
II. Related work

As mature 4D scene reconstruction methods have emerged, many multi-view datasets have been proposed. We focus on multi-view video content for free viewpoint rendering, leaving aside existing content designed for concurrent tasks such as avatar reconstruction [5], scene graph [6]

Datasets captured in studio feature long sequences of diverse and complex human interaction and manipulation with high camera density, such as the Panoptic Studio dataset [4], the NeuralDome dataset [10]. The Replay dataset [3] contains audio-video recordings of various human social activities. Contrarily to these sequences, we propose in-situ recordings of real training scenarios to preserve the natural environment that is a key element in most demonstrations.

Other multi-view datasets for 4D scene reconstruction capture human activities in-situ with a portable dense array of cameras, such as the Light Field Video dataset [9], the Immersive light Field dataset [7] and the Plenoptic Video dataset [8]. These datasets are limited to short sequences and feature compact camera rigs limiting view interpolation to a 3DoF+ context with only a few steps of movement possible.

With the exception of [3], none of the presented dataset feature audio capture, where audio is captured with binaural microphones, while we capture sound with ambisonic spatial audio microphones, which are more suitable for free viewpoint sound restitution. Our dataset is the only one that proposes real training situations captured in-situ. Inspired by in studio rigs and 3DoF+ in-situ rigs, we propose a close array rig that captures the precise gesture of the demonstrator and a sparser rig that captures the environment.



(a) Frontal view (b) Horizontal cross-section

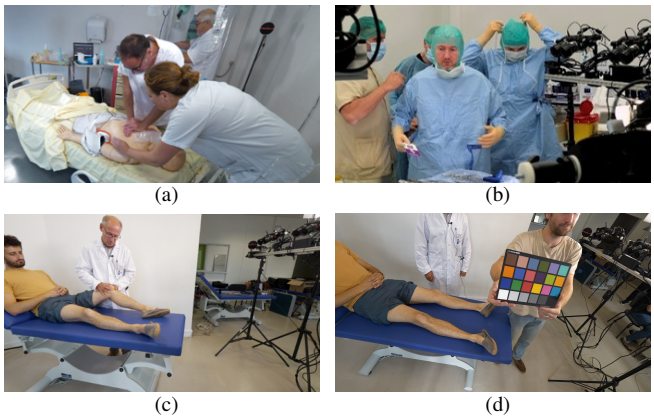
Fig. 2: Detailed views of our central camera rig. Fig. (b) shows a horizontal cross section of the central rig. The rig, in green, captures the view of an observer O, so that all cameras point to the intersection I at the edge of the gesture zone, in red. The trainer is positioned at T and his gesture is centered at C.

III. Rig design

Our multi-view camera setup, see fig. 1, consists of two rigs. A central rig captures the central navigation area from the student's ideal position. A peripheral rig captures the entire scene and forms a navigation path from which the student can observe.

A. Central Camera Rig

Our central camera rig consists of 12 Panasonic BGH1 cameras arranged in a 4x3 horizontal rectangle, see fig. 2. The



(a) (b) (c) (d)

Fig. 3: The three recorded sequences are shown in (a)-(c). For each sequence, a color checker is passed in front of all cameras in a calibration rush, shown in (d).

central rig captures the gesture zone of the training scenario: the trainer's gesture. The central rig is compact enough to fit on a wheeled support.

B. Peripheral camera Rig

Our peripheral rig, shown in fig. 1, consists of two units of 8 cameras each. Each linear unit has two parts, 6 linear cameras, and two linear cameras connected by a hinge, to allow smoother angles between the units. The role of the peripheral rig is to capture both the gesture zone and the training environment, it is placed on the side of the central rig as far away from the scene as possible. We grouped the cameras into stereo pairs to allow comparison of novel volumetric reconstruction methods with more classical stereo-based rendering techniques.

C. Audio capture equipment

Our audio capture rig is designed for spatialized audio capture and signal processing for 6DoF audio rendering. We use four Zylia ZM-1 19 capsules ambisonic microphones for spatial audio capture. The microphones are placed around the audio scene as shown in fig. 1. Two microphones are attached to the central rig to maximize QoE around the gesture zone, and the other two are placed at other extremities of the scene.

IV. Sequence description

We recorded three sequences of approximately one hour each, from three different medical training scenarios. Each sequence was recorded in-situ at the respective hospital. For color calibration of each sequence, a first take is made with a color checker running in front of all cameras, as shown in fig. 3d. Another take for spatial calibration of the microphones is done by moving around a speaker that emits high frequency audio sweeps. The sequences IV-A and IV-C will be publicly available to the research community with the consent of the participants and the sequence IV-B will have restricted access for privacy and ethical reasons.

A. Training in first aid gestures

The training in first aid gestures and care took place in a room at the Center and Laboratory for Simulation Learning in Health, located within the Health Personnel Training Center of the Rennes University Hospital. For this training, three trainers from the PFPS played the roles of health professionals, as shown in fig. 3a. This team of actor-caregivers simulated first aid gestures on an animated connected mannequin, representing a patient in anaphylactic shock following the administration of an antibiotic. The first aid gestures demonstrated included chest compressions, intubation, ventilation, and defibrillation.

B. Training on the insertion of a trochanteric nail

The capture of the training on surgical procedures, in particular the insertion of a trochanteric nail, took place in the PLaTIMed platform room of the Brest University Hospital, as shown in fig. 3b. During this training, two surgeons performed a trochanteric nail insertion using surgical tools from a specific accessory designed for this type of procedure. To reduce the length of the content, five takes were made to capture the key stages of the surgery.

C. Training for the Lower Limb Musculoskeletal Examination

The recording of the training on the examination of the musculoskeletal system of the lower extremities took place in an examination room of the Morvan Hospital of the Brest University Hospital, as shown in fig. 3c. During this training, a professor demonstrated the various gestures of the lower limb examination on a patient (see fig. 6). A single 24-minute take was sufficient to capture the entire course.

V. Applications

We planned and shot three medical scenarios with our rig. We demonstrated the effectiveness of our rig for in-situ indoor situations with limited camera rig space, with the central camera rig placed less than two meters from the trainer. Because the camera rig consists of three separate modules, only two people are needed to transport and assemble the rig. The modules also provide flexibility in the overall camera configuration, which we demonstrate by shooting in three rooms of different shapes. The use of pre-assembled units significantly reduces installation time, which can be completed by two people in 30 minutes.

We captured real medical training situations and proved that our rig is well suited to existing immersive training needs. Since training requires audio playback we proposed an audio rig that can be easily added to any multi-camera recording. We are releasing recordings IV-A and IV-C to the research community. Our sequences can be used for multimodal immersive rendering of 6DoF and high quality 3DoF volumetric video. Since we cover real-world medical training scenarios, the sequences can be used to assess the QoE of training with such immersive technologies as well as to compare their training potential with classical multi-view video for medical learning.

REFERENCES

[1] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.

[2] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 1-14.

[3] SHAPOVALOV, Roman, KLEIMAN, Yanir, ROCCO, Ignacio, et al. Replay: Multi-modal Multi-view Acted Videos for Casual Holography. In : Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 20338-20348.

[4] JOO, Hanbyul, LIU, Hao, TAN, Lei, et al. Panoptic studio: A massively multiview system for social motion capture. In : Proceedings of the IEEE international conference on computer vision. 2015. p. 3334-3342.

[5] IONESCU, Catalin, PAPAVA, Dragos, OLARU, Vlad, et al. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 2013, vol. 36, no 7, p. 1325-1339.

[6] ÖZSOY, Ege, ÖRNEK, Evin Pınar, ECK, Ulrich, et al. 4d-or: Semantic scene graphs for or domain modeling. In : International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham : Springer Nature Switzerland, 2022. p. 475-485.

[7] BROXTON, Michael, FLYNN, John, OVERBECK, Ryan, et al. Immersive light field video with a layered mesh representation. ACM Transactions on Graphics (TOG), 2020, vol. 39, no 4, p. 86: 1-86: 15.

[8] LI, Tianye, SLAVCHEVA, Mira, ZOLLHOEFER, Michael, et al. Neural 3d video synthesis from multi-view video. In : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 5521-5531.

[9] SABATER, Neus, BOISSON, Guillaume, VANDAME, Benoit, et al. Dataset and pipeline for multi-view light-field video. In : Proceedings of the IEEE conference on computer vision and pattern recognition Workshops. 2017. p. 30-40.

[10] ZHANG, Juze, LUO, Haimin, YANG, Hongdi, et al. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 8834-8845.