A Scalable and Cost-Effective Multi-View Capture System for Photorealistic Dynamic Human Reconstruction

Steffen-Sascha Stein[1*] [†] Dennis Amuser[1*] [‡] Kai Altwicker[1*] [§] David Mertens[1] Matthias Bullert[1]
Alisa Rüge[1] David Martin Karg[1] Kristoffer Waldow[1,2] [¶] Arnulph Fuhrmann[1] [‖]

[1]TH Köln, Computer Graphics Group
[2]Technical University of Munich (TUM), Human-Centered-Computing and Extended Reality Lab

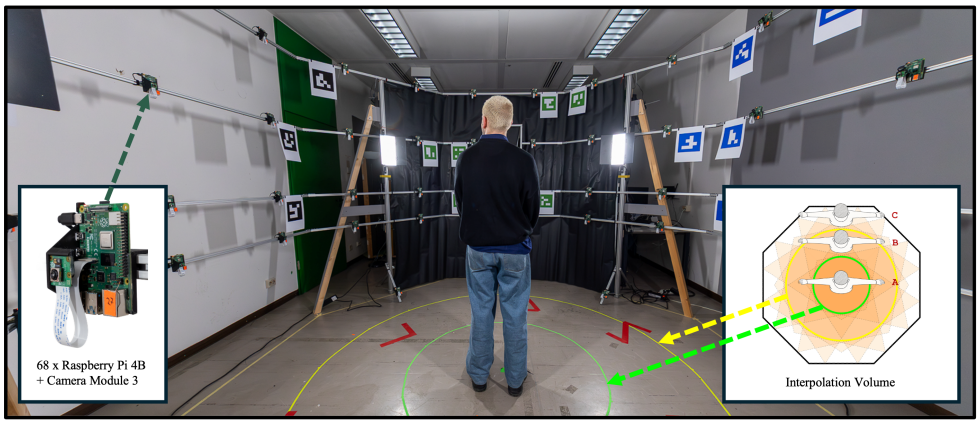

Figure 1: Our multi-view caputre system consists of 68 capture units (**left**). The horizontal interpolation volume (**right**) describes three different areas: (**A**) Optimal capture zone (≈1.34 m ∅) where all camera views overlap. (**B**) Usable volume limit (≈2.6 m ∅) where camera coverage is reduced. (**C**) Areas outside the interpolation volume resulting in significant reconstruction artifacts.

Abstract

Photorealistic human representation is a key factor for embodiment and presence in Virtual and Social Virtual Reality. While recent approaches such as 3D Gaussian Splatting enable high-quality reconstructions, dynamic human capture remains challenging due to the need for synchronized multi-view data. Existing capture systems are often costly and inaccessible, limiting reproducibility and dataset availability. We present a low-cost, reproducible multi-view capture rig consisting of 68 synchronized Raspberry Pi cameras arranged for full 360° coverage. The system enables reliable acquisition of dynamic human data suitable for 3D Gaussian Splatting and VR applications.

Index Terms: Avatars, Capture System, Virtual Humans, Radiance Fields

1 Introduction

The creation of immersive virtual environments is a central goal of Virtual Reality (VR) and, in particular, social VR applications [4]. A key factor contributing to a greater sense of presence in such environments is embodiment, which is defined as the user's perception of owning and controlling a virtual body [3]. Embodiment is typically achieved through mesh-based avatars, which serve as the user's visual and interactive representation in the virtual world. Although stylized avatars are commonly used, photorealistic self-representations have been shown to further enhance presence, identification, and social interaction [6, 5].

However, creating such avatars through manual modeling by artists is time-consuming and costly. As an alternative, reconstruction-based approaches, such as multi-view stereo (MVS), can be used to generate 3D meshes with textures [1]. While effective for static objects, these methods often produce non-manifold meshes with self-intersecting triangles, which is particularly problematic for dynamic elements such as clothing. Occlusion further limits reconstruction quality, as hidden areas are rendered incompletely or in poor quality, especially for dynamic scenes. Moreover, mesh-based reconstructions struggle to capture fine-scale details such as hair, leading to noticeable visual artifacts. 3D Gaussian Splatting (3DGS) [2] allows some of these problems to be mitigated, as this type of radiance field representation works with independent geometric primitives in the form of anisotropic Gaussians.

Although training times and visual quality are more advanced than those of MVS, the quality of the output from both methods depends heavily on the quality of the underlying images. This necessitates multi-view camera systems, as monocular capture setups are fundamentally insufficient for accurate dynamic reconstruction. Existing multi-view capture systems are typically expensive, complex, and difficult to access. Furthermore, publicly available datasets for dynamic, multi-view human reconstruction are scarce and offer limited variability.

To address these challenges, we present a reproducible capture rig comprising 68 Raspberry Pi cameras arranged in an octagonal configuration, providing full 360° coverage of a human subject (see Fig. 1). Our system enables synchronized multi-view video capture suitable for dynamic human reconstruction using 3D Gaussian

[†]e-mail: stein.steffen@hotmail.de
[‡]e-mail: mail@dennis-amuser.de
[§]e-mail: altwicker@tallaldiproduction.de
[¶]e-mail: kristoffer.waldow@th-koeln.de
[‖]e-mail: arnulph.fuhrmann@th-koeln.de

*These authors contributed equally to this work.

Figure 2: Static 3DGS [2] reconstruction, trained and rendered in Jawset Posthot (v.0.5.115), based on datasets generated with our capture system. From left to right: ground truth; full reconstruction; and cropped reconstruction. High quality cropping was easily achieved by limiting the rendering volume.

Splatting. Our rig is modular, cost-effective (ca. €7,500), and designed to facilitate reliable dataset generation.

2 Method

Before construction, we determined the number of cameras to maintain a high reconstruction quality for 3DGS and keep the cost factor of our system low. Therefore, a synthetic dataset based on a static 3D scene centring a high-quality virtual human was generated in Blender, trained, and evaluated. We identified a requirement of 60 to 70 cameras for 360° reconstruction without the appearance of strong artifacts. Based on our findings, the final rig was equipped with 68 synchronized capture units. Each unit consists of a Raspberry Pi 4B (1GB) and a Camera Module V3. This setting enables us to generate dynamic datasets in Full HD at 30 frames per second, as well as static captures in UHD.

The physical structure is an octagonal rig with a 4 m diameter and 2.6 m height, providing an internal interpolation volume with full camera overlap of approximately 1.34 m in diameter (see Fig. 1). To ensure stability and scalability, the frame utilizes modular industrial aluminum profiles, while cost-effective and reproducible 3D-printed polylactide mounts secure the cameras. To overcome tracking failures in subsequent processing steps caused by the rig's high symmetry and featureless room walls, 24 randomized ArUco markers were integrated into the setup to provide unique optical anchor points for the camera localization process, using Structure-from-Motion. We perform manual focus calibration for every camera using Siemens stars, as production variances and gravity impact the internal voice-coil actuators.

To achieve frame synchronization without network jitter, we employ a local Network Time Protocol (NTP) server hosted on a dedicated Raspberry Pi. All capture agents synchronize their system clocks with this server, after which a master node broadcasts an absolute future timestamp for the capture start, enabling simultaneous recording across cameras. However, due to hardware and I/O limitations, individual Raspberry Pi units may occasionally drop frames, causing temporal misalignment in the recorded streams. To address this, we monitor sensor timestamps during capture and log deviations from the expected frame interval as dropped frames, which are provided as metadata for subsequent processing.

3 Discussion

Our evaluation of the needed number of cameras based on a synthetic dataset enables low costs while maintaining photorealistic quality. Using as few cameras as possible also has a positive effect on the reconstruction process, as a smaller dataset leads to lower VRAM consumption. In dynamic settings in particular, this helps to keep the memory footprint lower than the available GPU memory allowing efficient training.

In terms of reconstruction quality, we evaluate our capture rig using 3DGS human reconstruction and the Structural Similarity Index Measure (SSIM). The reconstructed results shown in Figure 2 achieve SSIM scores of 0.931 and 0.949, indicating a high degree of structural consistency compared to the reference images.

Our multi-view capture system is an alternative to the low-cost volumetric capture approaches currently available, such as the method proposed by Boensch et al. [1]. As their system restricts subjects to a volume of approximately 0.8 m^3, our rig offers an optimal capture volume (see Figure 1) of approximately 3.67 m^3, which is around 4.6 times larger.

4 Conclusion and Future Work

This work presents a low-cost, multi-view capture rig designed to provide 360° coverage for both static and dynamic scenes. Alongside frame synchronisation and frame drop detection, it ensures sufficient data integrity and quality to guarantee photorealistic reconstructions of dynamic human movements. We use 3DGS to investigate this, as it is ideal for playing back recorded animations in VR applications. Overall, our work establishes an accessible foundation for capturing dynamic human datasets. This supports research on embodiment and presence in VR by lowering the barrier to photorealistic virtual human reconstruction.

For future work, other camera modules could be investigated, such as those that enable global shutter or the attachment of different lenses to increase temporal and visual resolution. However, these modules could incur higher costs. We also plan to develop automatic calibration and focusing of all camera units to avoid the time-consuming manual process and achieve higher visual quality. Further plans include generating and publishing datasets, and making this work open source to ensure reproducibility.

Acknowledgments

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) within the StartUpLab@TH Köln project under grant number 13FH015SU8.

References

[1] A. Bönsch, A. W. Feng, P. Patel, and A. Shapiro. Volumetric video capture using unsynchronized, low-cost cameras. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2019, Volume 1: GRAPP, 2019*, pages 255–261. SciTePress, 2019. 1, 2

[2] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), July 2023. 1, 2

[3] K. Kilteni, R. Groten, and M. Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012. 1

[4] D. Roth, J.-L. Lugrin, D. Galakhov, A. Hofmann, G. Bente, M. E. Latoschik, and A. Fuhrmann. Avatar realism and social interaction quality in virtual reality. In *2016 IEEE virtual reality (VR)*, pages 277–278. IEEE, 2016. 1

[5] K. Waldow, A. Fuhrmann, and S. M. Grünvogel. Investigating the effect of embodied visualization in remote collaborative augmented reality. In *International Conference on Virtual Reality and Augmented Reality*, pages 246–262. Springer, 2019. 1

[6] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652, 2018. 1