2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3315-1484-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/VRW66409.2025.00299

Full-Body Interaction in Mixed Reality using 3D Pose and Shape Estimation

Hong Son Nguyen[*]
Korea University

Andrew Chalmers[†]
Victoria University of Wellington

DaEun Cheong[‡]
Korea University

Myoung Gon Kim[§]
Korea University

Taehyun Rhee[¶]
The University of Melbourne

JungHyun Han[‖]
Korea University

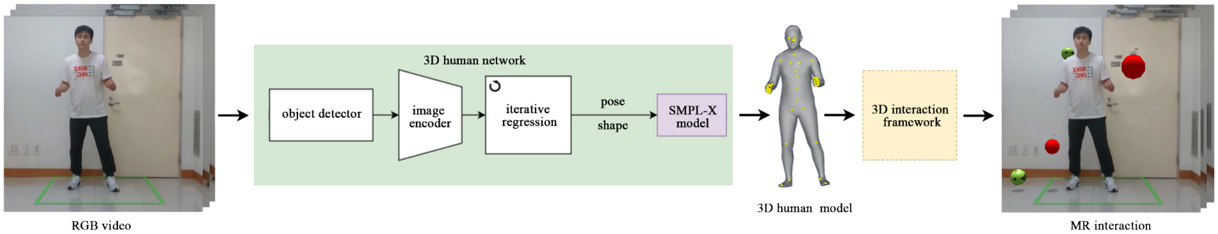

Figure 1: The proposed pipeline is composed of 3D human network and 3D interaction framework.

Index Terms: Mixed reality, Full-body interaction, Pose and shape estimation.

Abstract

This paper presents a pipeline that estimates a user's 3D pose and shape to facilitate a user's full-body interaction with virtual objects in a mixed reality environment. The usability and effectiveness of the pipeline are demonstrated through a user study.

1 Introduction

Estimating 3D human pose has been studied extensively. For example, Desai et al. [1, 2] utilized skeletal data in mixed reality (MR) games. However, they constrained the collider representation to simple geometric shapes, such as boxes or capsules. Escalona et al. [3] proposed a framework for rehabilitation exercises using 3D pose and shape estimation. However, it was not targeted at MR interactions. In this paper, we present a comprehensive pipeline that takes an RGB video as input, estimates a user's 3D pose and shape, and facilitates the user's full-body interaction with 3D virtual objects in an MR environment.

2 Methodology

The first component of our pipeline shown in Fig. 1 is named *3D human network*. Located at its front end is the object detector, and the detected human is passed to the image encoder, ConvNeXt-T [4], which is followed by an iterative regression module proposed by Kanazawa *et al.* [5]. It returns the 3D pose and shape of the parametric human model, SMPL-X [6]. Using them, the human's 3D triangular mesh is reconstructed. Then, the joints' 3D positions are derived from the mesh. The 3D human model given in Fig. 1 consists of the mesh and joints.

The reconstructed human model is taken as a *proxy* that interacts with virtual objects. It is used to detect collisions with the virtual objects. The second component of our pipeline, *3D interaction framework*, is built upon Unity engine.

Our MR environment is sketched in Fig. 2-(a). A user stands on the rectangular workspace and the RGB camera captures the user's full body. The large TV screen behind the camera displays the captured environment, which is augmented with virtual objects.

In our experiment, each user played a game, where virtual *bombs* were repeatedly spawned one at a time and the user was asked to avoid the bombs while staying within the workspace. See Fig. 2-(b). In addition to bombs, *balls* were intermittently spawned, and the user was asked to touch them. See Fig. 2-(c). The user lost scores if hit by the bombs but earned scores by touching the balls.

Our user study is designed to compare our human model with the baseline models that lack detailed shape information. We use three different colliders for the 3D human proxy, as shown in Fig. 3:

- For collision detection, we use the proxy's mesh extracted from the SMPL-X model. This collider is denoted as **mesh**.
- Each *bone*, which connects a pair of adjacent 3D joints of the human model, is associated with either a *capsule* or a *box*. For thin and elongated bones such as arms and legs, we use capsules. In contrast, boxes are used for the other bones such as head and torso. This collider is denoted as **capsule&box**.
- For each 3D joint of the human model, we assign a sphere so that the collider is represented as a connected set of spheres. This collider is denoted as **sphere**.

Among the colliders, **mesh** is most faithful to the user's actual shape whereas **capsule&box** mimics the shape and **sphere** is an overly simplified approximation. Our hypothesis is that the orders of *usability* and *effectiveness* are **mesh**, **capsule&box** and **sphere**.

For our experiment, an additional collider is defined by combining all of **mesh**, **capsule&box** and **sphere**. It is named **combined**. When playing the game with **combined**, collisions are tested "simultaneously" with all of its three components, and whether a collision occurs is recorded for each component.

The game was played as follows:

- The game playing time with a collider is three minutes, i.e., 180 seconds, and a break of two minutes is provided between the games with different colliders.
- For each collider, the virtual objects are spawned every 1.2 seconds, making the total number of objects 150 (= 180/1.2).
- The ratio of bombs and balls is 4 : 1, i.e., we have 120 bombs and 30 balls for each collider.

For the user study, we recruited 37 participants, 11 males and 26 females between the ages of 20 and 35 ($M = 26.86$ and $SD =$

[*] e-mail: nguyenhongson303@gmail.com
[†] e-mail: andrew.chalmers@vuw.ac.nz
[‡] e-mail: wjdekdms001@korea.ac.kr
[§] e-mail: m_gon_kim@korea.ac.kr
[¶] e-mail: taehyun.rhee@unimelb.edu.au
[‖] e-mail: jhan@korea.ac.kr

DOI 10.1109/VRW66409.2025.00299

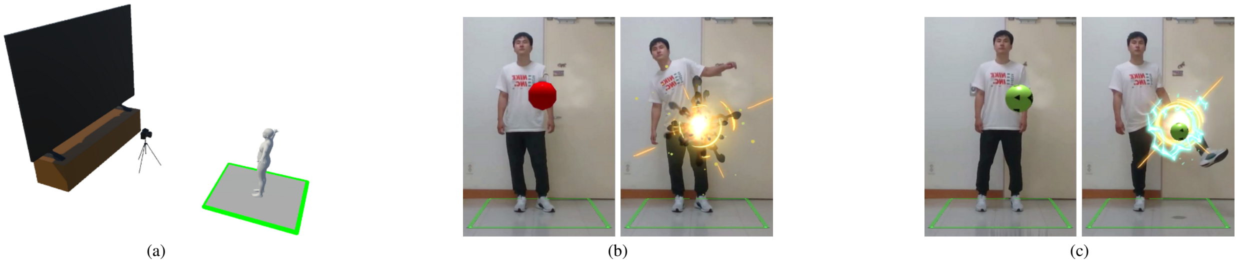

Figure 2: Experiment setup: (a) Our MR environment [7]. (b) Standing on the rectangular workspace, the user fails to avoid the red bomb, which then explodes due to collision. The user loses scores. (c) The user successfully touches a green ball with his leg and earns scores.

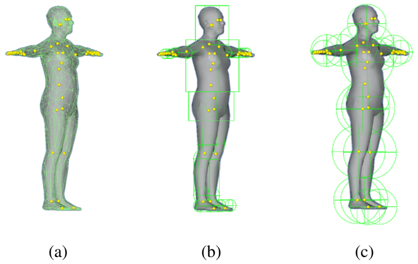

Figure 3: Three representations of a 3D human proxy: (a) **mesh**. (b) **capsule&box**. (c) **sphere**.

		mesh	capsule&box	sphere
SUS ↑	*M*	**76.14**	73.98	64.39
	SD	15.14	13.74	21.01
	p		< 0.01 **	
NASA-TLX ↓	*M*	**4.68**	4.83	4.87
	SD	0.94	0.92	0.86
	p		> 0.05	

Table 1: ANOVA test results: SUS and NASA-TLX.

3.03). To mitigate learning effects, we employed a counterbalanced testing approach. The participants tested **mesh**, **capsule&box** and **sphere** in a manner where the order was counterbalanced, and then tested **combined**. After completing a test with each collider, the participants filled out a post-test questionnaire on usability, utilizing two established metrics: the System Usability Scale (SUS) and the NASA Task Load Index (NASA-TLX).

3 Results

We performed ANOVA tests for both SUS and NASA-TLX, and the results are presented in Table 1. There were significant differences in the SUS scores ($p < 0.01$) and therefore multiple paired t-tests with Bonferroni correction were conducted as post-hoc tests. The results are shown in Fig. 4a.

During the test made with **combined**, we recorded the number of colliding bombs and that of touched balls for each of the individual colliders (**mesh**, **capsule&box** and **sphere**). We conducted ANOVA test to find that there were significant differences in both bomb and ball counts, as shown in Table 2. Therefore, multiple paired t-tests with Bonferroni correction were conducted as post-hoc tests, and Fig. 4b shows the results.

4 Conclusion

This paper presents an MR pipeline, which facilitates natural full-body interaction between a real human and 3D virtual objects. Taking only a single RGB video feed as input, it supports 3D shape-based natural interactions. Through a user study, we assess the impact of human reconstruction on virtual object interactions. The analysis with two baselines, which lack detailed shape information, revealed enhanced usability and effectiveness brought by our human model.

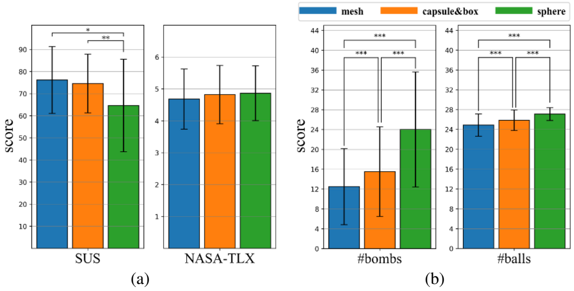

Figure 4: Qualitative analysis with the paired t-tests: (a) Usability. (b) Effectiveness.

		mesh	capsule&box	sphere
#bombs	*M*	12.5	15.5	24.03
	SD	7.8	9.19	11.80
	p		< 0.001 ***	
#balls	*M*	24.89	25.85	27.10
	SD	2.28	2.08	1.31
	p		< 0.001 ***	

Table 2: ANOVA test results: bomb and ball counts.

Acknowledgments

This work was supported by IITP grants funded by the Korea government(MSIT), No.RS-2020-II200861 and ICT Creative Consilience Program (IITP-2025-RS-2020-II201819). The Korea University IRB reviewed and approved the study protocol (IRB-2024-0075) and our user study was conducted under the protocol.

References

[1] K. Desai, K. Bahirat, S. Ramalingam, B. Prabhakaran, T. Annaswamy, and U. Makris. Augmented reality-based exergames for rehabilitation. In *Proceedings of the 7th International Conference on Multimedia Systems*, pages 1–10, 2016. 1

[2] K. Desai, S. Raghuraman, R. Jin, and B. Prabhakaran. Qoe studies on interactive 3d tele-immersion. In *2017 IEEE international symposium on multimedia (ISM)*, pages 130–137. IEEE, 2017. 1

[3] F. Escalona, E. Martin, E. Cruz, M. Cazorla, and F. Gomez-Donoso. Eva: Evaluating at-home rehabilitation exercises using augmented reality and low-cost sensors. *Virtual Reality*, 24:567–581, 2020. 1

[4] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of CVPR*, pages 11976–11986, 2022. 1

[5] A. Kanazawa, M. Black, D. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of CVPR*, pages 7122–7131, 2018. 1

[6] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of CVPR*, pages 10975–10985, 2019. 1

[7] H. Nguyen, A. Chalmers, D. Cheong, M. Kim, T. Rhee, and J. Han. A simple but effective ar framework for human-object interaction. In *EuroXR 2024: Proceedings of the 21st EuroXR International Conference*, pages 67–71, 2024. 2