EgoLiveHAND: Real-Time Egocentric 3D Hand Pose Estimation with Stereo Fisheye Glasses

Hyejin Hur
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
hyejin.hur@etri.re.kr

Seongmin Baek
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
baeksm@etri.re.kr

Hyunjun Lim
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
tp02134@etri.re.kr

YounHee Gil
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
yhgil@etri.re.kr

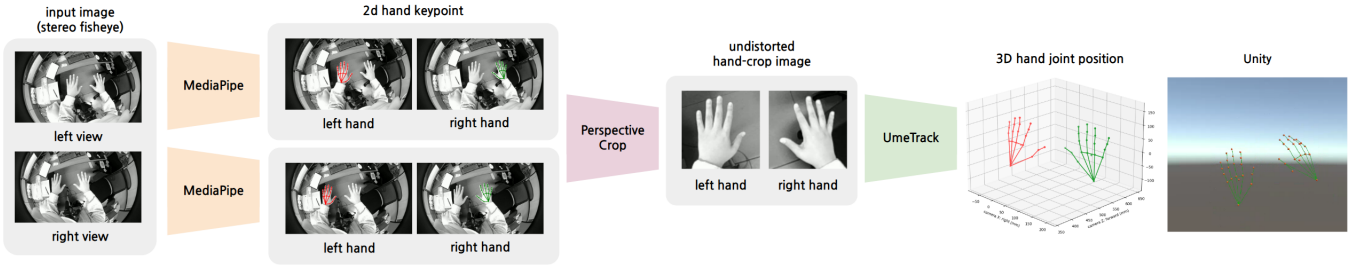


Figure 1: Overview of the EgoLiveHAND pipeline.

Abstract

In this paper, we propose EgoLiveHAND, a real-time egocentric 3D hand pose estimation system. It runs on a lightweight, glasses-type wearable device equipped with stereo fisheye cameras. The wide field of view of the fisheye cameras keeps the hands within the frame even during a wide range of motion, and the system runs at 41.5 fps with real-time Unity visualization. The estimated hand poses are visualized in Unity, enabling real-time interaction in virtual environments.

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → **Virtual reality**.

Keywords

3D Hand Pose Estimation, Egocentric Vision, Virtual Reality

ACM Reference Format:
Hyejin Hur, Seongmin Baek, Hyunjun Lim, and YounHee Gil. 2026. EgoLiveHAND: Real-Time Egocentric 3D Hand Pose Estimation with Stereo Fisheye Glasses. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3822517.3849925

1 Introduction

Estimating 3D hand pose from an egocentric viewpoint has become increasingly important with the advancement of HMDs and AI glasses. However, capturing a wide range of hand motion requires a wide field of view (FoV) camera, such as a fisheye camera, which in turn introduces lens distortion.

To address this, we develop a real-time hand pose estimation system built on a glasses-type wearable device equipped with fisheye cameras to capture a wide range of hand motion. The distortion is resolved via perspective cropping: MediaPipe [2] is used to infer the 2D hand joint positions required for perspective cropping, and the resulting undistorted hand crop images are then fed into UmeTrack [1], which infers the 3D hand joint positions.



Figure 2: Our stereo fisheye glasses prototype for egocentric hand pose estimation.

2 Method

Stereo Fisheye Glasses. The wearable prototype, shown in Fig. 2, integrates two synchronized grayscale fisheye cameras into a compact glasses-type form factor. The device weighs approximately 67 g, including the glasses frame and camera sensors but excluding the USB Type-C cable. The stereo camera module captures 2560 × 720 grayscale stereo images at up to 60 fps, where the left and right fisheye views each occupy 1280 × 720 pixels. The split stereo views are then processed by the real-time hand pose estimation pipeline. The cameras use global shutter sensors to reduce motion artifacts during rapid hand movements.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

VRST '26, Sendai, Japan

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3849925

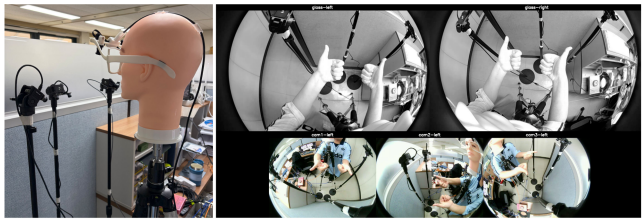


Figure 3: Multi-camera setup for 3D hand joint triangulation.

Table 1: 3D joint accuracy (MPJPE) against the multi-camera triangulation baseline. (Abs: no alignment, RA: root-aligned, PA: Procrustes-aligned)

Hand	Abs-MPJPE↓ (mm)	RA-MPJPE↓ (mm)	PA-MPJPE↓ (mm)	Frames
Left	26.83	23.57	8.24	763/819
Right	18.52	20.12	7.94	723/841
Both	22.79	21.89	8.10	1486/1660

EgoLiveHAND. As shown in Fig. 1, our pipeline takes stereo fisheye grayscale images from the wearable glasses as input and estimates the final 3D hand joint positions. MediaPipe infers 2D hand keypoints from the left and right fisheye views, which are used to generate a perspective crop for each detected hand (left and right hand) in each view, yielding undistorted hand crop images. For each hand, the corresponding crop images from the left and right camera views are then provided as multi-view input to UmeTrack, which estimates the wrist 3D position, wrist rotation, and joint angles; a generic hand model then computes the 3D joint positions via forward kinematics. The resulting 3D hand joint positions are transformed into the left camera coordinate system for real-time 3D visualization in Unity.

3 Experiments

Multi-Camera Triangulation Baseline. To evaluate 3D hand joint estimation accuracy, we placed three external RGB cameras around the workspace, providing left, right, and elevated front views (Fig. 3). Each camera was calibrated using a checkerboard pattern. For each frame, MediaPipe was used to detect 2D hand keypoints in all three views. The corresponding 3D joint positions were then reconstructed via multi-view triangulation and used as a comparison baseline. To ensure a consistent camera pose during evaluation, the stereo fisheye glasses were fixed on a mannequin.

3D Accuracy. For evaluation, a participant performed a sequence of diverse hand gestures with both hands, including showing the front and back of the hands, thumbs-up, V-sign, and peace-sign gestures. We evaluate 3D joint accuracy using Mean Per Joint Position Error (MPJPE). Abs-MPJPE is computed as the mean Euclidean distance between the predicted and baseline 3D joint positions without alignment. RA-MPJPE and PA-MPJPE are computed after root alignment at the wrist joint and Procrustes alignment, respectively. We compute the modified z-score of the per-frame mean joint error using the median and median absolute deviation (MAD). Frames with a score above 3.5 are excluded as outliers. As shown in Table 1, EgoLiveHAND achieves a PA-MPJPE of 8.10 mm across both hands.

Table 2: End-to-end pipeline FPS with and without the Unity interaction client.

Configuration	Mean (fps)	Std (fps)	Max (fps)
With Unity (n=11)	41.52	5.29	50.68
Without Unity (n=13)	48.91	5.56	54.60

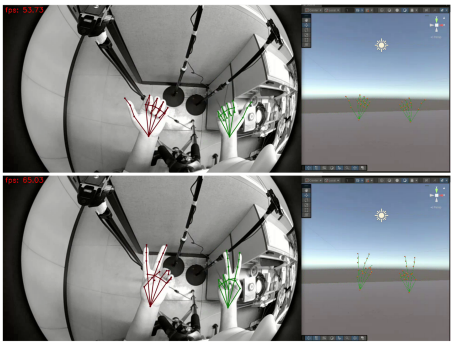


Figure 4: Real-time 3D hand pose estimation and Unity visualization with EgoLiveHAND.

Real-Time Operation. We run the pipeline on an NVIDIA RTX 4090 connected to the glasses via USB-C. As shown in Table 2, EgoLiveHAND runs at 48.91 fps without Unity and 41.52 fps with Unity visualization. Fig. 4 shows the estimated 3D hand joints projected onto the left fisheye camera view of the glasses and simultaneously visualized in Unity for real-time interaction.

4 Conclusion and Future Work

We presented EgoLiveHAND, a real-time egocentric 3D hand pose estimation system using lightweight stereo fisheye glasses. Our system enables real-time 3D hand interaction in virtual environments through perspective cropping and 3D hand pose estimation. Our evaluation is limited to a single participant, a fixed glasses setup, and an estimated triangulation reference. Future work will evaluate more realistic wearable settings and integrate EgoLiveHAND with body pose estimation for real-time full-body pose estimation.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World].

References

[1] Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. 2022. UmeTrack: Unified Multi-View End-to-End Hand Tracking for VR. In *Proceedings of SIGGRAPH Asia 2022 Conference Papers*. ACM, 1–9.

[2] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-Device Real-Time Hand Tracking. *arXiv preprint arXiv:2006.10214* (2020).