Robust Indoor Human Detection and Tracking via Fusion of mmWave Radar and Vision Sensor

Jae Myung Shin[†], Jae Yoon Jung[‡], and Kae Won Choi[*]
[†*]Department of Electrical and Computer Engineering, Sungkyunkwan University, Republic of Korea
[‡]Department of AI System Engineering, Sungkyunkwan University, Republic of Korea
Emails: [†]sjm2442@g.skku.edu, [‡]jungjy1018@g.skku.edu, [*]kaewonchoi@skku.edu

***Abstract*—Human detection and tracking in indoor environments are essential for applications such as smart homes, security monitoring, and elder care. Vision-based approaches provide detailed imagery, but are often constrained by occlusions, illumination changes, and privacy concerns. mmWave radar offers a robust alternative; however, the sparsity of its point clouds poses challenges for accurate skeleton reconstruction. To overcome this limitation, we present a deep learning framework that reconstructs 3D human skeletons from mmWave signals. The proposed system aligns radar point clouds with skeleton annotations from a vision sensor, allowing the network to learn a direct mapping from sparse radar inputs to skeletal structures. Experimental results confirm that our method enables reliable and privacy-preserving indoor human detection and tracking.**

***Index Terms*—mmWave radar, deep learning, human sensing**

I. Introduction

In indoor scenarios, accurate human detection and tracking are critical for applications such as smart homes, security, and elderly care. Conventional methods mainly rely on vision systems, which provide high-resolution image but are vulnerable to issues such as lighting fluctuations, occlusion, and privacy concerns. To overcome these limitations, mmWave radar has attracted attention as a promising alternative. mmWave radar is robust to lighting and clothing variations and can even penetrate nonmetallic obstacles, thereby enabling reliable human sensing. Nevertheless, hardware constraints yield sparse point clouds that hinder precise reconstruction of shapes and poses.

To mitigate these challenges, recent deep learning methods have been proposed to enrich semantic representations. For example, [1] uses motion capture labels to train a network for reconstructing 3D human meshes from sparse radar data, while [2] uses vision-based annotations to fine-tune a detector for robust localization. These works demonstrate that deep learning can significantly improve the applicability of sparse point clouds to human-centered perception.

In this paper, we propose a framework that fuses mmWave radar point clouds with 3D skeleton annotations obtained from a vision sensor. Radar point clouds serve as the input to a deep learning model, while vision-based skeletons provide ground truth during training. This approach enables the reconstruction of accurate human skeletons from sparse radar returns, leading to improved detection and tracking performance in indoor environments. Moreover, since radar data do not contain identifiable visual information, our approach inherently preserves privacy while improving sensing capability.

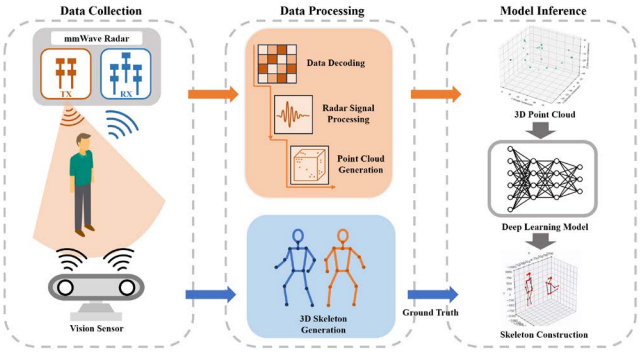

Fig. 1. Overview of the proposed system

II. System Overview

In this section, we describe the overall architecture of the proposed system. The system is designed to reconstruct skeletal representations in real time, enabling robust human detection and tracking. As shown in Fig. 1, the framework is organized into three stages: data collection, data processing, and model inference.

A. Data Collection

To enable real-time processing, a commercial frequency modulated continuous wave (FMCW) radar [3] transmits chirp signals and mixes the received echoes to produce intermediate-frequency (IF) samples. In parallel, a vision sensor [4] captures synchronized frames and extracts 3D skeletons using its SDK. Timestamping both radar and vision data ensures precise temporal alignment, while rigid co-mounting of the two devices maintains consistent geometry. Fig. 2 shows the integrated setup with the subject positioned at a fixed distance and orientation.

B. Data Processing

IF packets are first decoded and passed through a range FFT to obtain range bins, followed by a doppler FFT that produces a 2D range–doppler map. A constant false alarm rate (CFAR) detector is then applied to identify significant targets, and an angle of arrival (AoA) FFT on the virtual array resolves azimuth and elevation. Combining range and angle estimates yields 3D point locations, which are further augmented with radial velocity and signal strength before being paired with the corresponding 3D skeleton frame for dataset construction.

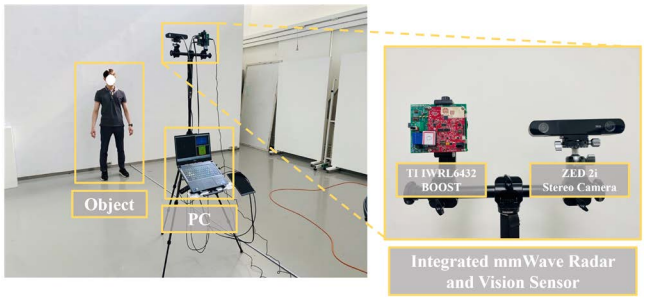

Fig. 2. Testbed of the integrated radar and vision sensor system

C. Model Inference

As illustrated in Fig. 3, the model employs a query-based transformer architecture to directly regress 3D human skeletons from mmWave radar point clouds. The input to the network is a set of 5D radar points per frame, where each point comprises x, y, z coordinates, radial velocity, and signal power. Each frame's point cloud is first normalized and padded to a fixed length for uniform processing. These point features are then tokenized via a linear projection into d-dimensional embeddings, yielding a sequence of input tokens for the transformer. A set of learnable object queries then attends to these tokens through multi-head attention, following a DETR-inspired design [5]. This mechanism allows each query to focus on relevant regions of the point cloud and aggregate cues for a potential human target. The transformer's output embeddings corresponding to each query are fed into three parallel prediction heads: one for the 3D bounding box parameters of the detected person, another for the 3D coordinates of the person's skeletal keypoints, and a third for an objectness confidence score. During training, Hungarian matching is applied to enforce a one-to-one assignment between predicted queries and ground-truth instances. A composite loss is then computed only over these matched pairs, combining contributions from bounding box regression, keypoint localization, and objectness. This end-to-end framework enables the model to reliably detect and reconstruct human skeletons in 3D from sparse radar inputs.

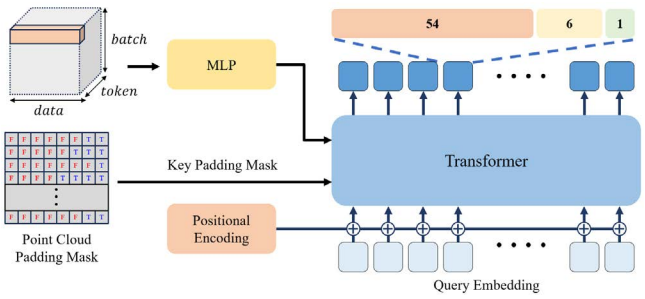

Fig. 3. Model architecture for skeleton reconstruction

III. Experiments

Our experiments were conducted using the setup in Fig. 2, where the mmWave radar and vision sensor were co-mounted to collect synchronized data streams. The subject performed various activities such as standing, walking, and sitting, and the resulting radar point clouds were paired with vision-based skeletons to form a multimodal training dataset. Fig. 4 presents representative results: the red skeleton corresponds to the ground truth from the vision sensor, while the yellow skeleton shows the reconstructed output of our model, demonstrating accurate alignment between input radar data and predicted skeletal structures.

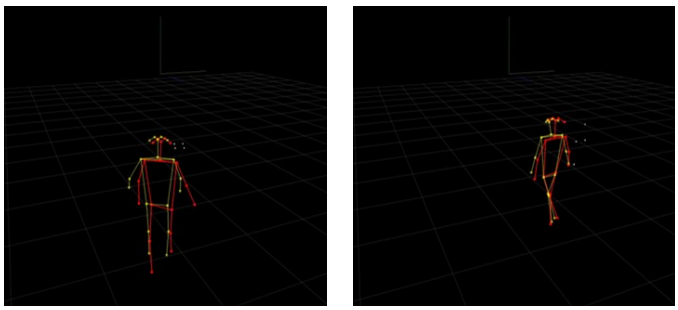

Fig. 4. Visualization of skeleton reconstruction

IV. Conclusion

In this paper, we presented a deep learning framework for indoor human detection and tracking that integrates mmWave radar with vision-based supervision. The system collects radar signals, processes them through cascaded FFTs to form point clouds, and aligns them with 3D skeleton annotations. Using this multimodal dataset, our model reconstructs accurate human skeletons from sparse radar returns, achieving reliable performance while inherently preserving privacy.

Acknowledgment

This work was partly supported by the BK21 FOUR Project and the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience Program(IITP-2023-2020-0-01821) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)"

References

[1] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 269–282.

[2] C. Yuance, H. Ding, D. Han, T. Zhang, G. Wang, C. Zhao, F. Wang, W. Xi, and J. Zhao, "mmYodar: Lightweight and robust object detection using mmwave signals," in *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2023, pp. 151–159.

[3] Texas Instruments, "IWRL6432 BoosterPack™ evaluation module for single-chip 60GHz mmWave low-power sensor," 2022. [Online]. Available: https://www.ti.com/tool/IWRL6432BOOST

[4] Stereolabs Inc., "ZED 2i Stereo Camera." [Online]. Available: https://www.stereolabs.com/en-pt/store/products/zed-2i

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020. [Online]. Available: https://arxiv.org/abs/2005.12872