Implementation of an Occupancy Detection and Tracking System Integrating Radar and Vision Sensor

Jae Myung Shin, Jae Yoon Jung, Kae Won Choi*

Sungkyunkwan Univ.

sjm2442@g.skku.edu, jungjy1018@g.skku.edu, *kaewonchoi@skku.edu

레이더와 비전 센서를 융합한 재실자 감지 및 추적 시스템 구현

신재명, 정재윤, 최계원*

성균관대학교

Abstract

This paper presents a real-time pipeline that synchronizes mmWave radar intermediate-frequency (IF) streams with 3D skeletons from a stereo camera, using cascaded range, doppler, and angle of arrival (AoA) FFTs with constant false alarm rate (CFAR) filtering to reconstruct denoised 3D point clouds. Streaming over UDP ensures precise alignment of radar and vision data. The resulting multimodal dataset of paired frames enables end-to-end neural models for robust indoor occupancy detection and trajectory tracking.

Ⅰ. Introduction

In indoor environments, human detection and tracking are essential for applications such as smart homes, security surveillance, and elder care. Conventional object detection approaches primarily depend on visual sensors like cameras. While camera-based systems offer high-resolution visual information, they remain vulnerable to lighting variations, occlusions, and privacy concerns. To address these shortcomings, mmWave radar has gained traction due to its insensitivity to lighting and clothing variations and its ability to penetrate non-metallic barriers, thereby providing reliable detection. However, radar performance is limited by hardware specifications, and the resulting point clouds are extremely sparse, rendering fine-grained shape and pose estimation challenging. Recognizing these challenges, recent deep learning-based approaches have sought to extract richer semantic information from sparse radar returns. The mmMesh framework [1] employs VICON motion-capture data as ground truth to train an SMPL-based model capable of reconstructing dynamic 3D human meshes from only a few dozen radar points per frame. In parallel, mmYodar [2] converts radar point clouds into pseudo-images and uses Azure Kinect-derived bounding-box labels to optimize a YOLOv3 network for accurate object detection. Building on these advances, we propose a framework that (1) fuses constant false alarm rate (CFAR)-filtered mmWave point clouds with 3D skeleton data from a ZED 2i stereo camera, and (2) integrates these modalities within a deep learning model to substantially enhance human detection and tracking performance in indoor settings.

Ⅱ. Method

This paper focuses on the design and implementation of synchronized data collection and processing pipeline for indoor occupancy detection and tracking. As illustrated in Fig 1, our pipeline comprises three sequential stages—data collection, data processing, and dataset generation—that work in concert to produce time-aligned mmWave point clouds and 3D skeletons. In the first stage, raw sensor outputs are streamed into the PC; the second stage applies signal-processing routines to extract 3D points; and the final stage assembles point-and-skeleton data pairs for subsequent deep-learning training and evaluation.

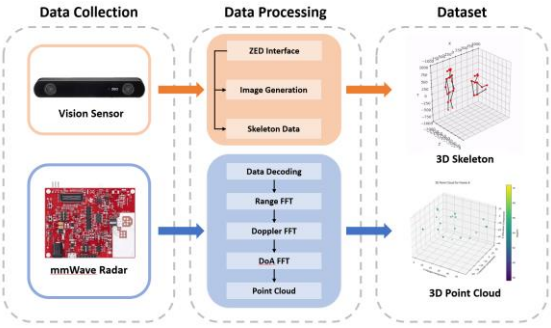

Fig 1. Overview of Data Pipeline

A. Data Collection

For real-time operation, a commercial frequency modulated continuous wave (FMCW) radar continuously emits chirped waveforms and mixes received signal

with the transmitted signal to yield intermediate-frequency (IF) samples. These IF data are packetized and streamed over a custom UDP protocol, ensuring minimal latency and packet loss. In parallel, a vision sensor captures synchronized video frames from which we extract 3D skeletons via its onboard SDK. By timestamping both the IF packets and skeleton outputs at acquisition, we guarantee precise temporal alignment between the radar returns and ground-truth poses. To ensure consistent sensor geometry throughout our experiments, both devices are rigidly co-mounted on a single frame. To illustrate this setup, Fig 2. shows the testbed of the integrated radar and vision sensor, with the subject positioned at a fixed distance and orientation relative to both sensors.

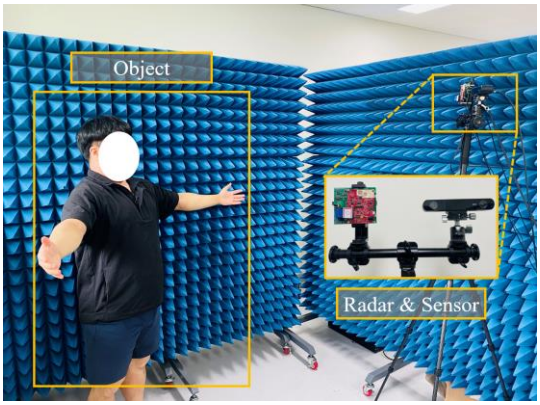

Fig 2. Testbed of Integrated Radar-Vision System

B. Data Processing

Upon arrival, IF packets are digitally decoded and fed into a cascaded FFT pipeline. First, a range-FFT converts time-domain samples into range bins, followed by a doppler-FFT that produces a 2D range-doppler map. We then apply a CFAR detector—but instead of discarding zero-doppler cells as clutter, we retain their amplitude information for ghost removal and background modeling. Only peaks exceeding a minimum velocity threshold are promoted to 3D points, while static returns remain available for filtering but are not emitted as point-cloud coordinates. Next, an angle of arrival (AoA)-FFT on the virtual array resolves both azimuth and elevation. By combining each moving target's range and angle estimates, we reconstruct its denoised 3D location. Finally, we augment every point with its radial velocity and signal-strength feature, then pair the resulting point cloud with the corresponding 3D skeleton frame for dataset creation.

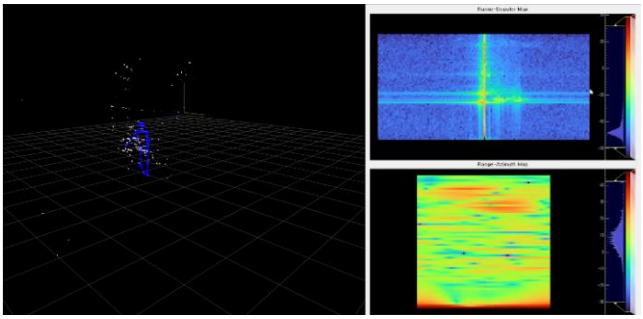

Fig 3. Results of Processing Pipeline

Fig 3. presents representative outputs of our preprocessing pipeline in a single composite layout. On the left, the 3D point cloud depicts the reconstructed spatial geometry of the moving subject. In the top-right, the raw range-doppler map—prior to CFAR filtering—visualizes signal intensity across distance and radial velocity, illustrating the full spectrum of motion and background returns. The bottom-right shows a range-azimuth heatmap constructed exclusively from zero-doppler (static) returns, revealing the positions of fixed reflectors—such as walls and furniture—to provide essential spatial context and support robust ghost removal. Together, these views demonstrate how the pipeline progresses from raw measurements to filtered, structured data for deep-learning: initial FFT outputs, background characterization, and final point-cloud reconstruction.

Ⅲ. Conclusion

This work presents data collection and processing pipeline for indoor human detection and tracking, integrating mmWave radar streams with 3D skeletons from a stereo camera and applying cascaded range, doppler, and AoA FFTs to generate point clouds precisely aligned with pose annotations. The resulting multimodal dataset—composed of paired point-cloud and skeleton frames—lays the foundation for subsequent model development.

In the next phase, these processed data will be fed into our neural network architecture to perform end-to-end occupancy detection and trajectory tracking. Future work will focus on quantitative evaluation of detection accuracy and tracking stability under varying lighting and occlusion conditions, exploration of multi-person scenarios, and the incorporation of additional sensor modalities to further improve robustness and generalization.

ACKNOWLEDGMENT

This work was supported by the BK21 FOUR Project.

REFERENCES

[1] H. Xue et al., "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in Proc. ACM MobiSys, 2021, pp. 269-282.

[2] C. Yuance et al., "mmYodar: Lightweight and Robust Object Detection using mmWave Signals," 2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Madrid, Spain, 2023.