Vomme: A Multimodal Sensing Platform for Video, Audio, mmWave and Skeleton Data Capturing

Xijia Wei
xijia.wei.21@ucl.ac.uk
University College London
London, UK

Yuan Fang
yuan.fang.20@ucl.ac.uk
University College London
London, UK

Kevin Chetty
k.chetty@ucl.ac.uk
University College London
London, UK

Youngjun Cho
youngjun.cho@ucl.ac.uk
University College London
London, UK

Nadia Bianchi-Berthouze
nadia.berthouze@ucl.ac.uk
University College London
London, UK

Abstract

mmWave sensing has offered a non-intrusive opportunity for human behaviour recognition. However, current mmWave sensing platform is limited for raw signal acquisition together with time-aligned multimedia data recording functions. In addition, there is a lack of open-sourced solution multi-mmWave sensor capturing toolbox. In this paper, we introduce Vomme, a multimodal sensing platform for video, audio, mmWave, and RBG-extracted skeleton data capturing. Vomme supports a series of sensor combination setup for data capturing, demonstrating potentials to be deployed under various application scenarios. Vomme synchronizes multimodal signals via the host computer's timestamp. Hardware-level synchronization is also supported by integrating a micro controller for precise sampling frequency control and avoiding the inter-sensor interference when using multiple mmWave sensors. Vomme is fully publicly open-sourced.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Ubiquitous computing**; • **Applied computing** → **Health care information systems**.

Keywords

mmWave Sensing, Human Activity Recognition, Multimodal Motion Capture

ACM Reference Format:
Xijia Wei, Yuan Fang, Kevin Chetty, Youngjun Cho, and Nadia Bianchi-Berthouze. 2025. Vomme: A Multimodal Sensing Platform for Video, Audio, mmWave and Skeleton Data Capturing. In *ACM Workshop on Access Networks with Artificial Intelligence (ANAI '25), November 4–8, 2025, Hong Kong, China.* ACM, New York, NY, USA, Article 111, 5 pages. https://doi.org/10.1145/3737904.3768536

Authors' Contact Information: Xijia Wei, xijia.wei.21@ucl.ac.uk, University College London, London, UK; Yuan Fang, yuan.fang.20@ucl.ac.uk, University College London, London, UK; Kevin Chetty, k.chetty@ucl.ac.uk, University College London, London, UK; Youngjun Cho, youngjun.cho@ucl.ac.uk, University College London, London, UK; Nadia Bianchi-Berthouze, nadia.berthouze@ucl.ac.uk, University College London, London, UK.

ANAI '25, November 4-8, 2025, Hong Kong, China

ACM ISBN 979-8-4007-1981-3/25/11
https://doi.org/10.1145/3737904.3768536

1 Introduction

With the rapid development of radio-frequency (RF) sensing technology, RF sensing has become a research hotspot for various sensing purposes. Millimetre-wave (mmWave) radar, a type of RF sensor, has emerged as a promising candidate due to its unique advantages: robustness under poor lighting and harsh environmental conditions [2], capability to penetrate certain obstacles such as fog or thin walls [7], and the ability to directly estimate the Doppler velocity of moving targets [5]. Its non-intrusive, ubiquitous nature and penetrate ability make it ideal for diverse applications. For instance, RF sensing and mmWave sensing have been adopted in the field of autonomous driving [11], healthcare monitoring [12], localization and tracking [9, 10, 13], navigation [14–16]. However, there is a lack of open-source toolboxes that offer an easy-to-use sensing approach for synchronously capturing mmWave, video, audio recordings, as well as skeleton keypoints when the presence of a human is detected. To fill this gap, we present the Vomme, a multimodal sensing platform for video, audio, mmWave and skeleton Data capturing.

We list our contributions as follows:

- We present Vomme, a multimodal sensing platform for video, audio, mmWave and skeleton data capturing.

- We offers single/multi mmWave sensor raw data capturing supports, where multi-mmWave synchronization is achieved via integrated micro controller, ensuring the precise control across multiple sensors and avoiding inter-sensor interferences.
- We release the Vomme sensing platform with the recommended hardware choices and configuration details fully publicly open-sourced.[1]

2 Background

2.1 mmWave Sensing Methodology

Millimetre-wave (mmWave) radar senses its environment by transmitting a sequence of chirp signals through the transmit antennas (TX). These chirps, generated by a frequency synthesizer, propagate through the medium and partially reflect off objects. The reflected signals return to the radar and are captured by receive antennas (RX). The received signal is then mixed with the transmitted chirp to produce an intermediate frequency (IF) signal, which contains information about the object's range and velocity.

Chirp Signal Model: The transmitted chirp is a linearly frequency-modulated signal. Its instantaneous frequency is:

$$f(t) = f_0 + St, \tag{1}$$

where f_0 is the start frequency and $S = \frac{B}{T_c}$ is the frequency slope, determined by the bandwidth B and chirp duration T_c.

The instantaneous phase $\phi(t)$ of the chirp is the integral of its frequency:

$$\phi(t) = 2\pi \int (f_0 + St)dt = 2\pi f_0 t + \pi S t^2. \tag{2}$$

Thus, the transmitted signal can be expressed as:

$$s_T(t) = \cos\left(2\pi f_0 t + \pi S t^2\right). \tag{3}$$

Received Signal: The reflected signal experiences a round-trip delay $\tau = \frac{2d}{c}$, where d is the target distance and c is the speed of light. Incorporating attenuation α, the received signal is:

$$s_R(t) = \alpha \cos\left(2\pi f_0 (t-\tau) + \pi S (t-\tau)^2\right). \tag{4}$$

IF Signal and Beat Frequency: After mixing $s_T(t)$ and $s_R(t)$ and applying low-pass filtering, the resulting intermediate frequency (IF) signal can be characterized by the following equations:

$$\begin{aligned} f(t) &= f_0 + St \\ \phi(t) &= 2\pi \int (f_0 + St)dt = 2\pi f_0 t + \pi S t^2 \\ s_T(t) &= \cos\left(2\pi f_0 t + \pi S t^2\right) \end{aligned} \tag{5}$$

where f_0 is the start frequency and $S = \frac{B}{T_c}$ is the frequency slope, determined by the bandwidth B and chirp duration T_c. The beat frequency f_{IF} is proportional to the target distance d, forming the basis for range estimation. Phase variations across multiple chirps capture Doppler shifts, which enable velocity estimation.

2.2 FFT-Based Signal Processing

The raw IF signal resides in the time domain. To extract spatial and motion information, it is processed using Fast Fourier Transforms (FFT) along different dimensions:

- **Range FFT:** Applied across ADC samples within a single chirp to estimate target distances. Each range bin corresponds to a specific frequency component related to object range. Magnitude reflects reflection strength.
- **Doppler FFT:** Applied across chirps for each range bin to estimate radial velocity. Motion induces phase shifts between chirps, which appear as peaks in Doppler bins.
- **Azimuth FFT:** Applied across antennas to estimate the angle of arrival (AoA). Phase differences among antennas reveal the signal's spatial origin.

These three steps—Range FFT, Doppler FFT, and Azimuth FFT enable precise estimation of range, velocity, and angle, forming the foundation of mmWave radar perception.

3 Heatmap Generation

After performing the 3D FFT, the radar data is organized into a Range–Azimuth–Doppler (RAD) cube. Its three axes represent the range index (distance), the Doppler index (radial velocity), and the azimuth index (angle of arrival). Since FFT outputs are complex, the magnitude is typically taken:

$$\text{Magnitude} = |X| = \sqrt{\text{Re}(X)^2 + \text{Im}(X)^2}, \tag{6}$$

where X is the complex FFT output. This magnitude represents reflection intensity, which depends on factors such as distance, material, and radar cross-section.

By projecting the RAD cube into any two dimensions, different heatmaps can be generated. The **Range–Angle (RA) heatmap** shows intensity versus range and angle, and can be used to detect targets at different distances and directions. The **Range–Doppler (RD) heatmap** displays range versus radial velocity and contains Doppler patterns that can be further exploited for classifying specific motions. The **Doppler–Angle (DA) heatmap** represents velocity versus angle, but it is less commonly used in practice.

3.1 MediaPipe Skeleton Detection

The skeleton keypoints are extracted from the RGB camera via the MediaPipe framework [8]. MediaPipe extracts the skeleton joint information per frame.

[1] https://weixijia.github.io/Vomme

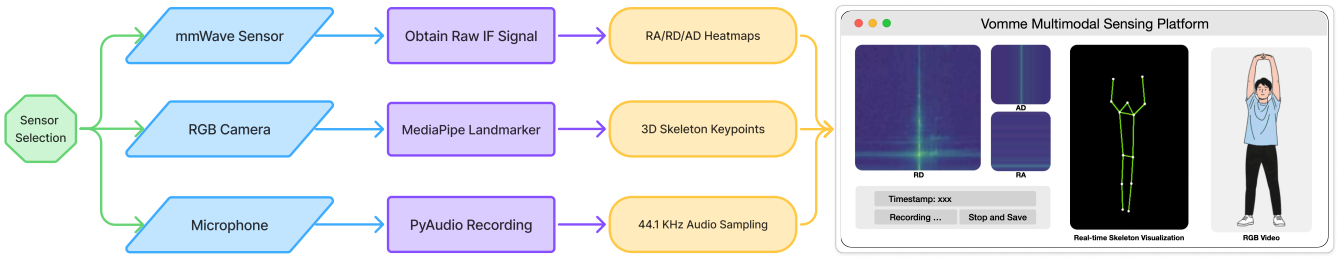


Figure 1: Vomme System Overview

4 Vomme Sensing Platform

Vomme is a modular, multimodal sensing platform for human motion capturing, integrating mmWave radars, RGB cameras, depth or ToF cameras (e.g., Kinect), and microphone sensors, where the overview of the system is shown in Figure 1. It supports both single and multi-radar configurations. Deploying multiple radars with complementary layouts and viewpoints can expand coverage and mitigate elevation limitations. Accordingly, Vomme enables flexible placement and synchronization of one or more single-chip mmWave radars to broaden the field of view and enhance sensing performance.

Figure 2a illustrates Vomme's system connection. It consists of one or multiple hardware-triggered mmWave radars synchronized with the host computer and multimedia sensors (e.g., cameras, microphone) for multimodal data capturing. Figure 2b shows an example of Vomme's sensor connection, consisting of two orthogonally placed mmWave radars, an RGB camera for video recording (and for MediaPipe-based skeleton extraction), with a microphone for audio capture. Time synchronization is achieved by the micro controller (attached on a bread board placed on the left).

4.1 Multi-Radar Placement Layouts

Vomme supports three general layouts for multi-radar deployment: (i) an orthogonal (90° rotated) pair, (ii) a line-up (stacked / collinear) arrangement, and (iii) a distributed (surround-view) setup. This design allows users to adjust the sensing coverage, resolution, and hardware according to application needs.

Orthogonal (90° rotated) pair. Two identical single-chip mmWave radars can be placed close to each other and rotate one unit by 90°, as demonstrated in [3, 6]. The rotation maps the azimuth array of the rotated unit onto the elevation axis. Consequently, one radar provides high-resolution azimuth while the rotated unit provides the same high-resolution in elevation, yielding near-balanced 2D angular resolution. The design is simple and cost-effective. It substantially improves elevation resolution and can even outperform cascaded systems limited to four vertical virtual antennas (VAs), while avoiding their added cost and complexity.

Line-up (stacked/collinear). Multiple mmWave radars can also be placed along a line (often at different heights) so that each covers a distinct elevation band or horizontal slice, and their fields of view jointly expand the vertical coverage, where the similar layout setting is used in [1]. This configuration is practical when most activity is in front of the sensors but spans a big height range (from low to high), e.g., for human skeleton extraction.

Distributed (surround-view). mmWave radars can be arranged around the subject like a motion-capture rig to provide complementary viewpoints for robust 3D reconstruction and skeleton tracking. This surround layout reduces self-occlusion and view dependence, improving completeness and tracking stability.

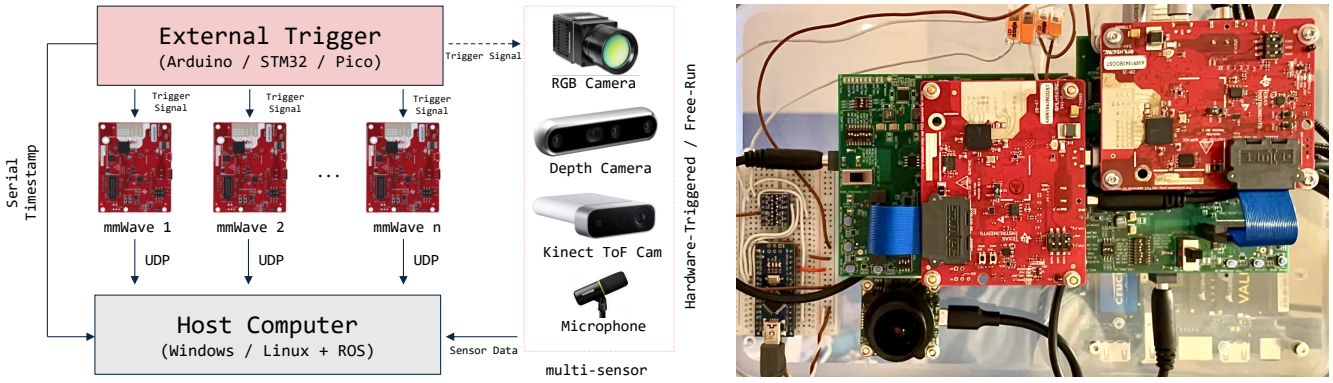
4.2 Hardware Trigger and Synchronization

Inter-radar interference is the main challenge when deploying multiple mmWave sensors, as sensors sharing the same band and point into the same direction or facing to each other. One of the receivers possibly pick up the other's echo, leading to ghost targets and biased estimates. A straightforward and effective solution is time-division that only trigger one radar at a time and stagger their acquisitions so that chirp frames could avoid overlapping.

Given number of N mmWave radars sharing the same band, let $T_{\text{frame}} = 40.8\,\text{ms}$ be the per-radar frame time and $T_{\text{cycle}} = 100\,\text{ms}$ the acquisition interval (10 Hz). To avoid overlap within each cycle, choose a start delay t_{delay} (the idle time between the end of one frame and the start of the next) that meets

$$N\,(T_{\text{frame}} + t_{\text{delay}}) \le T_{\text{cycle}} \tag{7}$$

In a dual-radar configuration where $N = 2$, choosing $t_{\text{delay}} = 9.2\,\text{ms}$ yields a simple schedule: Radar 1 runs from $t = 0\,\text{ms}$ to 40.8 ms; wait 9.2 ms; Radar 2 runs from 50.0 ms to 90.8 ms; the remaining 9.2 ms in the 100 ms cycle stays idle and this prevents any overlap between the dual radar. For



(a) System Connections **(b) Sensor Layout Example**

Figure 2: Vomme hardware connection diagram with an example

$N > 2$, schedule additional radars sequentially within the 100 ms interval. If needed, shorten T_{frame} (e.g., fewer chirps or a higher sample rate) so that Eq. (7) remains satisfied.

Triggering multiple radars requires an external controller to generate trigger pulses. Common choices include Arduino[2], STM32[3], and Raspberry Pi Pico[4]. The same trigger source can also be logged for precise timestamping and shared with other sensors (e.g., cameras) for hardware synchronization. Notably, when hardware synchronization is unavailable, these sensors can free-run and be aligned to the radar using timestamps easily.

4.3 Real-Time Processing

Vomme acquires raw data from one or multiple Texas Instruments AWR1843 Boost single-chip mmWave radars with DCA1000 EVM capture card. Other TI single-chip devices that support DCA1000 LVDS streaming are also supported. Radar configuration parameters are set in the official mmWave Studio software, but once streaming starts, our custom pipeline takes over for continuous acquisition, decoding, post processing, heatmap generation, and online visualization; unlike the official workflow, which supports only offline processing. To achieve real-time heatmap generation, we use CUDA [4] to achieve millisecond-level real-time visualization. The toolchain is also cross-platform, with optional GPU acceleration for heatmap generation, and provides ROS/ROS2[5] publishers for easy integration with multi-sensor stacks. In practice, the system enables plug-and-play, real-time mmWave data capture and post-processing.

[2]Arduino is an open-source electronics platform.
[3]STM32 is a 32-bit microcontroller and microprocessor integrated circuits by STMicroelectronics.
[4]Raspberry Pi Pico is the microcontroller chip designed by Raspberry Pi built using RP2040
[5]The Robot Operating System (ROS) is a set of software libraries and tools for building robot applications.

4.4 Video Recording Function

Video recording is achieved by the integrated RGB camera. Each radar frame is aligned with the corresponding camera image using timestamps when the camera operates in free-run mode, or through hardware triggering to ensure precise synchronization. Vomme also supports the integration of the Kinect camera. During video recording is activated, Vomme leverages a parallel threading to ensure the video recording would not be impacted by other sensing IO.

4.5 Skeleton Extraction

Vomme supports two modes of skeleton extraction. When using an RGB camera for video recording, Vomme supports to extract the skeleton information via the MediaPipe at a frame level. For MediaPipe extracted skeleton, the saved CSV file includes the three-dimensional joint coordinates, visibility scores, and corresponding timestamps. When recording video using the Kinect sensor, Vomme leveraged the Kinect pre-built pose landmaker models for directly extracting outputs skeleton keypoints.

4.6 Audio Recording Function

Audio recording is done via the plugged-in microphone. Again, audio recording function is achieved by creating a parallel thread by using the PyAudio[6] package. Audio is sampled at a frequency of 44.1KHz.

5 Data Validity

Figure 3 presents examples of four distinct human actions, each illustrated by 3 frames, alongside their corresponding Range-Doppler (RD) representations. The RD frames reveal unique micro-Doppler patterns associated with different body gestures. Notably, actions involving larger movements

[6]PyAudio is a Python-based library for audio recording

and higher speeds produce broader regions of Doppler shift in the RD images, reflecting the increased velocity and spatial extent of the motion.

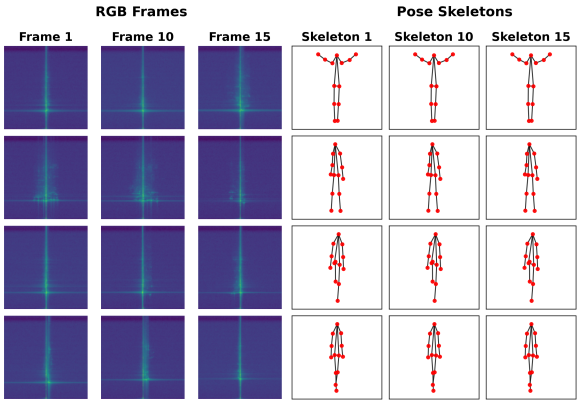


Figure 3: Visualization of samples captured via Vomme platform. We report the RD spectrogram at frame 1, 10, 15 with the corresponding skeleton figures, where skeletons figures are visualized based on 13 keypoints from the raw 33 extracted keypoints for visualization simplification.

6 Conclusion

In this paper, we present Vomme, an open-source multimodal sensing platform that enables time-aligned acquisition of video, audio, mmWave data (including both raw signals and extracted RD/RA/AD heatmaps), and RGB-extracted skeleton data. Additionally, Vomme supports multi-mmWave sensing setups, addressing inter-mmWave signal interference by leveraging an additional microcontroller for precise sampling control at the hardware level. Vomme is fully publicly open-sourced, with hardware recommendations and configuration details. We hope that the release of Vomme can facilitate easier collection of mmWave signals alongside multimedia signals, reduce the cost and complexity of setting up data collection platforms through an out-of-the-box solution, and accelerate the adoption of mmWave sensing in various real-world scenarios.

References

[1] Han Cui and Naim Dahnoun. 2021. Real-time short-range human posture estimation using mmWave radars and neural networks. *IEEE Sensors Journal* 22, 1 (2021), 535–543.

[2] Fangqiang Ding, Xiangyu Wen, Yunzhou Zhu, Yiming Li, and Chris Xiaoxuan Lu. 2024. Radarocc: Robust 3d occupancy prediction with 4d imaging radar. *Advances in Neural Information Processing Systems* 37 (2024), 101589–101617.

[3] Yuan Fang, Fangzhan Shi, Xijia Wei, Qingchao Chen, Kevin Chetty, and Simon Julier. 2025. CubeDN: Real-Time Drone Detection in 3D Space from Dual mmWave Radar Cubes. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 3977–3983. doi:10.1109/ICRA55743.2025.11127766

[4] Pawan Harish and Petter J Narayanan. 2007. Accelerating large graph algorithms on the GPU using CUDA. In *International conference on high-performance computing*. Springer, 197–208.

[5] Tianshu Huang, John Miller, Akarsh Prabhakara, Tao Jin, Tarana Laroia, Zico Kolter, and Anthony Rowe. 2024. Dart: Implicit doppler tomography for radar novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24118–24129.

[6] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5715–5724.

[7] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 14–27.

[8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[9] Fangzhan Shi, Wenda Li, Chong Tang, Yuan Fang, Paul V. Brennan, and Kevin Chetty. 2024. Decimeter-Level Indoor Localization Using WiFi Round-Trip Phase and Factor Graph Optimization. *IEEE Journal on Selected Areas in Communications* 42, 1 (2024), 177–191. doi:10.1109/JSAC.2023.3322812

[10] Fangzhan Shi, Wenda Li, Chong Tang, Yuan Fang, Paul V. Brennan, and Kevin Chetty. 2025. ML-Track: Passive Human Tracking Using WiFi Multi-Link Round-Trip CSI and Particle Filter. *IEEE Transactions on Mobile Computing* 24, 6 (2025), 5155–5172. doi:10.1109/TMC.2025.3529897

[11] Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. 2021. RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization. *IEEE Journal of Selected Topics in Signal Processing* 15, 4 (2021), 954–967. doi:10.1109/JSTSP.2021.3058895

[12] Xijia Wei, Temitayo Olugbade, Fangzhan Shi, Shuang Wu, Amanda William, Nicolas Gold, Youngjun Cho, Kevin Chetty, and Nadia Bianchi-Berthouze. 2023. Leveraging WiFi Sensing toward Automatic Recognition of Pain Behaviors. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8.

[13] Xijia Wei and Valentin Radu. 2019. Calibrating recurrent neural networks on smartphone inertial sensors for location tracking. In *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[14] Xijia Wei and Valentin Radu. 2021. MM-Loc: Cross-sensor indoor smartphone location tracking using multimodal deep neural networks. In *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[15] Xijia Wei and Valentin Radu. 2022. Leveraging transfer learning for robust multimodal positioning systems using smartphone multi-sensor data. In *2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[16] Xijia Wei, Zhiqiang Wei, and Valentin Radu. 2021. Sensor-fusion for smartphone location tracking using hybrid multimodal deep neural networks. *Sensors* 21, 22 (2021), 7488.