V2R: FMCW Radar Data Synthesis from Videos for Long Range Gesture Recognition

Koushik A. Manjunatha, Morris Hsu, Rohit Kumar, Sai Prashanth Chinnapalli
Amazon Lab126, Sunnyvale, CA, USA, 94089

***Abstract*—The increasing popularity of wireless sensing applications has led to a growing demand for large datasets of realistic wireless data. However, collecting such data is often time-consuming and expensive. To address this challenge, we present a novel Video-to-Radar (V2R) framework that generates synthetic mmWave radar data for human gestures using unstructured videos. The V2R framework combines a mesh fitting model to extract 3D spatial features of human subjects from videos, with an FMCW radar physics model to simulate realistic radar signals. This approach enables the generation of diverse synthetic data, which can be used to train and validate gesture recognition models. By incorporating the V2R-generated data, we demonstrate a significant improvement in the classification accuracy of our LSTM-based gesture recognition model, with the overall accuracy improvement of 12%.**

***Index Terms*—Video, Mesh, mmWave Radar, Wireless, Artificial Intelligence**

I. Introduction

Radio frequency (RF) sensing is a promising approach that has seen a lot of development in past decades. These sensors offer signal richness comparable to that of microphones and cameras but without any privacy constraints. The mmWave radar is one of those example that has seen a lot of development over the past years. Hand gesture recognition has become viable through the use of miniaturized, low-power radar sensors, as demonstrated by the Soli project [1]. Subsequent research has focused on improving these results, requiring the collection of significant training datasets for gesture recognition models [2]. These studies typically captured gestures performed close to the device or longer distance (upto 1m) in a controlled setup.

However, the development of AI/ML models for RF sensing as compared to development of signal processing algorithms has always lagged behind due to lack of data. Unlike vision-based systems, the development of mmWave-based sensing faces significant challenges due to the scarcity of training datasets [3]. Collecting such datasets is labor-intensive and time-consuming. Moreover, in case that the sensor hardware (e.g., antenna configuration) or the modulation is modified, the data collection have to be repeated.

Several recent works, such as [3]–[5], have explored synthetic data generation from videos. These studies have leveraged machine learning models to denoise the synthetic signal or have created simulation models of the Frequency-Modulated Continuous Wave (FMCW) radar system, integrating them with animation software (e.g., Blender) to generate synthetic data. However, these approaches have primarily focused on using Doppler features to predict activities. In contrast, [6] considered not only Doppler information but also azimuth and elevation angle data, generating gestures from Blender animations. While animation-based approaches can reliably generate synthetic data, they may struggle to capture the full versatility and real patterns of human activities, such as walking, running, and hand gestures.

In this work, we propose video to radar (V2R) framework, a novel mmWave sensing data synthesis method from unstructured videos to generate realistic data for sensing based applications. The V2R framework builds on two high-level design principles: (i) An AI model to extract 3D spatial features of the human subject from videos. (ii) Integrating a white-box physics-based mmWave simulator with the AI model pipeline to extract realistic RF signal propagation and interaction with the human subject. The pipeline also enables data augmentation where subject could be placed in 3D space at different field of view (FoV) and orientation. The the V2R significantly curtails the cost of RF sensing data collection, and enables several mmWave sensing applications.

The remainder of this paper is structured as follows: Section II introduces the V2R framework, which generates the synthetic dataset. This includes details on the mesh fitting process, viewpoint configuration, and the FMCW radar physics model used to simulate the radar output. Finally, Section III presents the results, discussing the experimental setup, feature extraction, model architecture, and the gesture recognition performance when using the real data as well as the combined real and synthesized data.

II. Video to Radar Framework

Fig. 1 depicts the V2R framework and the components comprising the V2R pipeline. More details on each block is explained in following subsections.

A. Mesh Fitting

To generate mmWave radar signal, we require equivalent 3D data of a user's body against a static background. As the first step in our V2R framework, we compute the position of all vertices of the human body by fitting a mesh to the video data. For this, we utilize the SMPLer-X [7], which estimates the human mesh through the AI model. Given an input video, the SMPLer-X outputs a human pose mesh for each frame.

K. A. Manjunatha, M. Hsu, and R. Kumar, S. Chinnapalli, are with Amazon Lab126, Sunnyvale, CA, USA, 94089 (email:{koushiam,rrohk,saic}@amazon.com, mhsu@lab126.com).

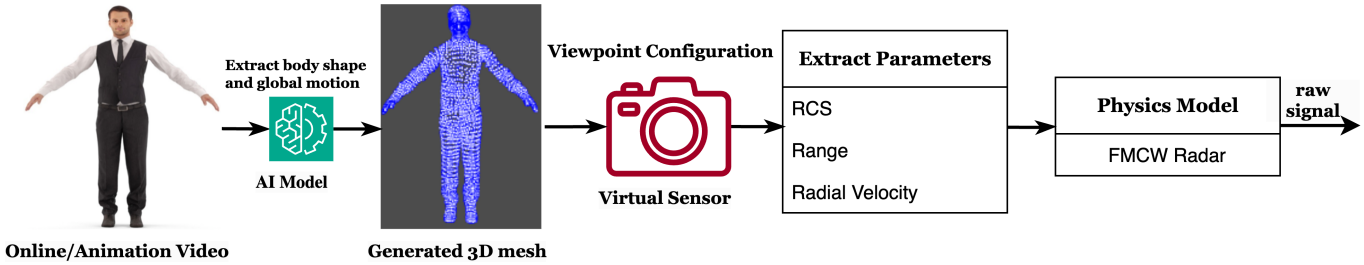

Fig. 1: V2R pipeline with blackbox mesh generation model and FMCW radar physics model.

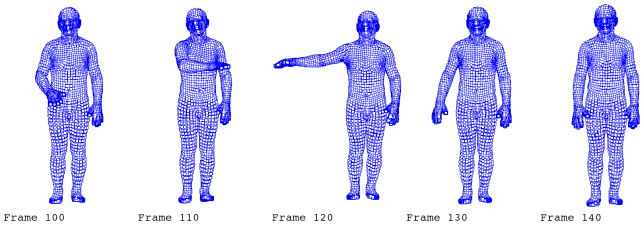

Fig. 2: Mesh generated at different video frames for right-swipe gesture. The frames shown are 100, 110, 120, 130, and 140 and the blue dots in the mesh represents vertices.

An example of a generated mesh from the SMPLer-X model of a human body performing a right swipe gesture is shown in Fig. 2.

B. Viewpoint Configuration

In the Skinned Multi-Person Linear human mesh model, the origin $(x,y,z) = (0,0,0)$ is typically located at the pelvis of the model. Considering the human mesh as the center, to generate data from various viewpoints, the virtual sensor will be placed at different locations $(x_c, y_c, z_c) = (i, j, k)$, where i, j, and k are in real-world units, for the human mesh in each frame.

1) RCS and Radial velocity: The RCS (σ) is set to 1 if the mesh point is not occluded for the set viewpoint else RCS is set to 0. Then, the radial velocity is calculated by the displacement of the each point between frames over set video frame rate.

2) Range Synthesis: The range is calculated using a euclidean distance between each mesh point and the virtual sensor location $(c1, c2, c3)$. As the main motivation of this work is to generate more data for AI models, the ability to place the camera at various locations and also with different view angles enables generating large set of augmented data from a video.

To simulate $N \times M$ multi-input multi-output (MIMO) FMCW radar, the spatial distances from a reference transmit antenna to every other transmit and receive antennas is considered. Typically, the spatial distance will be $\lambda/2$, where λ is wavelength of the mmWave signal. In this work, 1×3 FMCW radar is considered with Tx and Rx location shown in Fig 3. For the considered Tx-Rx layout in Fig 3 with $Tx = (c1, c2, c3)$, the receiver locations are derived as $Rx_1 = (c1, c2 + dy, c3)$, $Rx_2 = (c1 - dx, c2, c3)$, and, $Rx_3 = (c1 - dx, c2 + dy, c3)$. The euclidean distance between the Rx locations and each vertex of the mesh is used to extract range as below

$$\vec{r}_{ij} = \|Rx_j - V_i\| \tag{1}$$

where $i = 1, 2, ..., n$, $j = 1, 2, 3$ are vertex number and receiver, respectively.

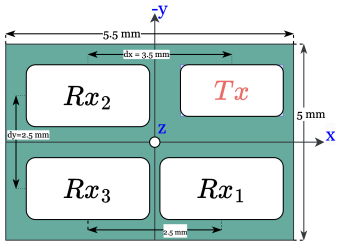

Fig. 3: Infineon MIMO radar Rx-Tx layout in 3D space.

C. FMCW Radar Physics Model

The FMCW radar operates by transmitting a linear frequency chirp signal and mixing it with the received echoes to obtain the beat signal. The transmitted FMCW chirp signal can be mathematically expressed [8] as:

$$t = m \times T + t_m, \text{ and } f(t) = f_0 + B \times \frac{t_m}{T} = f_0 + \gamma t_m \tag{2}$$

where t_m is the time within the m-th chirp, $f(t)$ is the instantaneous frequency of the chirp signal,and $\gamma = \frac{B}{T}$ is the frequency slope. Following this, we can represent the transmitted mmWave signal $u^{Tx}(t)$ as:

$$u^{Tx}(t) = A \times e^{j\phi} \tag{3}$$

where A is the amplitude and ϕ is the initial phase of the signal. The integral $\int_0^{mT+t_m} f(x), dx$ represents the total number of signal periods that have occurred since the start of the frame.

Since the phase ϕ of the signal advances by 2π with each full period, we can relate the phase ϕ to the number of periods as:

$$\phi = 2\pi \left(\int_0^{mT+t_m} f(x) \cdot dx \right) + \phi_0 \tag{4}$$

Substituting this expression for ϕ into $u^{Tx}(t)$ [8]

$$u^{Tx}(t) = Ae^{j\left(2\pi\left(f_0 t + \gamma\left(t_m^2 + mT^2\right)\right) + \phi_0\right)} \tag{5}$$

where f_0 is the start frequency of the chirp and γ is the chirp rate or slope.

The transmitted signal $u^{Tx}(t)$ will be received back from a target at range R and with a radial velocity v with respect to the receiving antenna. The received signal $u^{RX}(t)$ can be viewed as a time-delayed (and also attenuated) version of the transmitted signal, with a latency $\tau = \frac{2R}{c}$, where c is the speed of light. Thus, the received signal $u^{RX}(t)$ can be expressed as:

$$u^{RX}(t) = A' e^{j\left(2\pi\left(f_0(t-\tau)+\gamma\left((t_m-\tau)^2+mT^2\right)\right)+\phi_0\right)} \tag{6}$$

where A' is the attenuated amplitude, obtained according to the radar communication principle.

By mixing the transmitted and received signals, the base-band beat signal can be obtained as:

$$u^{IF}(t) = A' e^{2\pi j\left(f_0\tau - \frac{B\tau^2}{2T} + \frac{Bt_m\tau}{T}\right)} = A' e^{2\pi j\left(\frac{2\gamma R}{c}t + \frac{2f_0 v}{c}t\right)} \tag{7}$$

For a MIMO FMCW radar scenario with N transmit antenna and M receive antenna, the radar-detected signal between the i^{th} transmit antenna and k^{th} receive antenna for the entire human mesh is constructed by applying the superposition principle as:

$$S_{ik}(t) = \sum_n \sum_m A'_{n,m} e^{2\pi j\left(\frac{2\gamma R_{n,m}}{c}t + \frac{2f_0 v_{n,m}}{c}t\right)} \tag{8}$$

where $R_{n,m}$ is the range of the n-th human mesh point at the m-th chirp, $v_{n,m}$ is the velocity of the n-th human mesh point at the m-th chirp for the i^{th} transmit antenna and k^{th} receive antenna pair.

III. Experimental Results

A. Experimental Setup

For our experimental setup, we configured the Infineon's 60 GHz radar sensor which has a sampling rate of 750 KHz and a total bandwidth of 460 MHz, spanning from $f_{\text{low}} = 61.02$ GHz to $f_{\text{high}} = 61.48$ GHz. The IF gain is set to 40 dB. The radar system performs 128 chirps per frame, with each chirp lasting 0.23 milliseconds, and the frame repetition time is 30 milliseconds. The system collects 32 samples per chirp and has a frequency slope of 16.0156 GHz/s, enabling high-resolution distance measurements with a range resolution of 30 cm. Along the side of real mmWave radar data collection, video camera at *30fps* was used to record videos.

The radar device was equipped with three receive antennas arranged in an L-shape configuration, which allowed for the estimation of the angle of the scattering target in two planes. The received signal from each chirp was converted to an intermediate frequency, passed through an anti-aliasing filter, and digitized with 32 samples at a rate of 2 MHz.

B. Feature Extraction

In this work, we examine gestures performed by subjects close to the radar device. The participants carried out the gestures within a field of view of $-20°$, $0°$, and $20°$ at distances of 1 m, 1.5 m, and 2 m from the sensor with a varying mounting height of the radar (0.95m, 1.35m). The set of gestures included four directional swipes *(left, right, up , down)* towards the sensor, as shown in Fig. 4.

The gestures can be fully described by a time series of RF scattering characteristics of the moving hand, including Range, Doppler, Azimuth angle, Elevation angle, and Power. By capturing these key features, we are able to synthesize the millimeter-wave signals that would be observed by the radar sensor during the execution of these gestures.

The received radar signal is first segmented using short-term averaging (STA) and long-term averaging method (LTA). Once the waveform is segmented, for each Rx antenna, FFT along samples is performed (range FFT) and then after static removal, FFT along chirps or Doppler FFT is generated. This creates a complex valued Range-Doppler (RD) map.

From the RD map, the first step is to estimate the human body around zero range-Doppler bin. Once the human body is estimated, we create a region of interest (RoI) with ranges from the body position to 3 range bins (1 m as each range bin has a resolution of 30 cm) from it and (2 bin or 60 cm) beyond it. This RoI is then used to extract features over time (search the high Doppler values that relates to Hand, and track the peak Doppler over frames to extract range and estimate angles).

The dataset collected from our experimental setup was first processed through the signal processing pipeline. The same pipeline was then used to process the synthetic raw samples generated from the V2R pipeline using video recordings.

C. Gesture Model Architecture

A 2 layer LSTM model is used and training in a streaming model pattern. The model structure represents a single-layer LSTM model, where the LSTM layers are responsible for capturing the sequential patterns in the input data, and the single dense layer with softmax activation for the classification having total of 3765 trainable parameters.

Total sequence length (negative data plus gestures) of 100 frames were considered as a sample. The data was normalized to have values ranging between 1 and +1. The data is highly imbalanced towards negative/background. To compensate for that the *classweight* for each class of gesture is determined and used during the training as a sample weight for every training sample to avoid bias in the model update.

D. Gesture Recognition Performance

The neural network was trained on a dataset with an 80:20 split between training and validation sets. The validation set contained gesture samples from real devices belonging to 5 different subjects, while the training data had samples from 15 subjects, including some synthesized V2R data.

The model was trained using sequences of 100 time steps, with the categorical cross-entropy loss optimized using the Adam algorithm with a learning rate of 1e-3 and a batch size of 8. Training was run for 150 epochs, and the model with the best validation accuracy was selected.

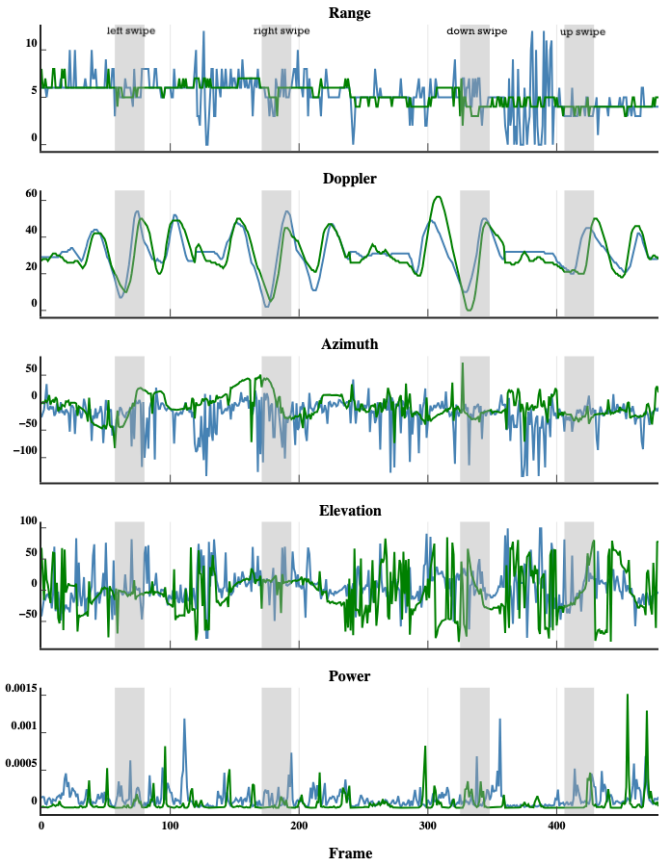

Fig. 4: Range, Doppler, Azimuth, Elevation, and Power features generated from V2R (Green), and real device (Blue).

TABLE I: Confusion matrix for model trained on i. real data, and ii. real plus V2R data.

Train Data	Predicted / Actual	left	right	up	down	negative
real data	left	12.88	46.85	28.95	8.97	2.331
	right	2.25	66.796	8.886	21.456	0.606
	up	14.859	26.122	44.14	5.232	9.646
	down	1.434	9.688	9.986	69.032	9.856
	negative	4.056	2.959	4.403	7.596	80.984
real + V2R data	left	60.989	26.015	5.577	4.612	2.805
	right	8.886	75.422	5.353	8.755	1.582
	up	6.303	6.655	78.880	3.38	4.78
	down	2.934	4.499	3.353	83.966	5.245
	negative	7.389	6.048	6.165	8.267	72.129

The presented confusion matrix shows the performance on two datasets - real-world radar data, and a combination of real data and synthesized V2R data. The real data baseline showed moderate classification accuracy, with significant misclassification between certain gestures.

However, incorporating the synthesized V2R data during training led to significant improvements in classification accuracy across all gesture classes. For example, the left gesture accuracy increased from 12.88% to 60.98%. This demonstrates the benefits of using synthesized data to enhance the model's ability to generalize and correctly identify the target gestures.

The one exception was a dip in prediction accuracy for the negative/background data, potentially due to label leakage from the non-gesture samples. Overall, the results highlight the value of leveraging synthesized data to improve gesture recognition performance.

IV. Conclusion

We present a novel V2R framework that generates synthetic millimeter-wave radar data for human gestures. This framework combines a mesh fitting model with an FMCW radar physics model, enabling the generation of diverse synthetic data to train and validate gesture recognition models. The V2R-generated data substantially improves the performance of the classification model, demonstrating the effectiveness of the signal synthesis approach. Future work will expand the V2R capabilities and explore techniques to enhance the fidelity of the synthesized data.

References

[1] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proceedings of the 29th annual symposium on user interface software and technology*, 2016, pp. 851–860.

[2] M. Strobel, S. Schoenfeldt, and J. Daugalas, "Gesture recognition for fmcw radar on the edge," in *2024 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNeT)*. IEEE, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1109/WiSNeT59910.2024.10438579

[3] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition," ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445138

[4] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 1, mar 2023. [Online]. Available: https://doi.org/10.1145/3580872

[5] C. Williams and C. Li, "Prediction and simulation of fmcw radar hand gesture detection based on captured 3d motion data," in *2023 IEEE Radio and Wireless Symposium (RWS)*. IEEE, 2023, pp. 55–57.

[6] A. Ninos, J. Hasch, M. E. P. Alvarez, and T. Zwick, "Synthetic radar dataset generator for macro-gesture recognition," *IEEE Access*, vol. 9, pp. 76 576–76 584, 2021.

[7] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, Y. Wang, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu, "Smpler-x: Scaling up expressive human pose and shape estimation," 2024. [Online]. Available: https://arxiv.org/abs/2309.17448

[8] X. Zhang, Z. Li, and J. Zhang, "Synthesized millimeter-waves for human motion sensing," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 377–390. [Online]. Available: https://doi.org/10.1145/3560905.3568542