Radar-Based Human Skeleton Estimation with CNN-LSTM Network Trained with Limited Data

M. Mahbubur Rahman
Dept. of Electrical & Computer Eng.
The University of Alabama
Tuscaloosa, USA
mrahman17@crimson.ua.edu

Dario Martelli
Dept. of Mechanical Eng.
The University of Alabama
Tuscaloosa, USA
dmartelli@ua.edu

Sevgi Z. Gurbuz
Dept. of Electrical & Computer Eng.
The University of Alabama
Tuscaloosa, USA
szgurbuz@ua.edu

***Abstract*—This paper presents a novel framework for human pose estimation using millimeter-wave (mmWave) radar technology, focusing on personalized healthcare applications. The proposed framework utilizes range-azimuth, range-elevation, and range-Doppler maps as inputs to a convolutional neural network (CNN) with a long short-term memory (LSTM) architecture to capture temporal dependencies and achieve improved skeleton estimation. Furthermore, this paper addresses the limitations of current radar-based skeleton estimation techniques, such as inconsistent kinematics and reliance on sparse radar point clouds. Skeleton estimation accuracy attained using diversified simulations is compared with that achieved real RF data, as validated using gold standard Vicon motion capture (MoCap) measurements as ground truth. The results highlight the potential of mmWave radar-based human skeleton estimation for advancing personalized healthcare and improving gait analysis and fall risk assessment.**

***Index Terms*—mmwave Radar, RF-Pose, skeleton estimation, Vicon, health care**

I. Introduction

Radar-based human activity recognition [1] has opened up new opportunities in the design of cyber-physical systems (CPS) for health and safety by providing an ambient, non-contact, non-intrusive way to monitor human movement at any time of the day (24/7). This is important because it can enable the development of RF-based techniques for the early diagnosis and post-treatment monitoring of ailments resulting in symptoms impacting gait, as well as in improving ageing-in-place and quality life by providing gait-based assessments of fall risk - all in a home environment, where the person monitored would be moving in a natural fashion while doing daily activities. As such, it can provide a more realistic assessment of human mobility and gait, as quantitative gait analysis (QGA) methods are often unavailable and a person often walks differently when cognizant of being observed. Moreover, RF technologies have have the potential to improve the accessibility of care while also reducing healthcare costs.

Most radar-based approaches to gait analysis rely on extraction of the micro-Doppler signature - essentially a 2D image that represents the superposition of the time-varying velocity profile of each point on the body. This representation provides a limited characterization of human gait because it does not show the association of the measured velocities cannot with the movement of specific points on the body, i.e. the skeletal representation of the human body.

Radar-based human skeleton estimation was first considered RF-Capture, proposed by Adib, *et. al* [2] in 2015. RF-Capture outputs the positions of coarse human body parts using 5.4 to 7.2 GHz FMCW signals that transmit through an antenna array and then stitches together the identified parts to reconstruct a human figure. In 2018, Khatabi *et.al.*, proposed RF-Pose3D [3], which utilizes a 12 elements T shape antenna array to transmit and receive FMCW signals at 6.3 GHz center frequency and 1.8 GHz bandwidth. This method utilizes Range-azimuth and Range-elevation heatmaps as inputs to a Resnet-based encoder neural network. It uses 12 camera nodes to record RGB-based video and obtain label key points from OpenPose. This data is used to train a region proposal network (RPN) that zooms in on the RF data of the individual person and a CNN with ResNet architecture to extract the 3D skeleton from the region of interest. For 14 key point localization, average errors in the x,y, z axes were reported as 4.2, 4.0, and 4.9cm, based on the key points estimated with OpenPose. Although signficiant at the time for showing that RF skeleton estimation was possible, this approach relied upon of over 17 million data samples acquired over 16 hours of recordings. Not only was the method extremely data greedy, but it required an extremely complex Deep Neural Network (DNN) with high computation costs, hindering the practicality of the approach.

In 2020, Sengupta *et. al* proposed mmPose [4], which predicts over 15 joints. To capture high-quality azimuth and elevation information, they used two IWR1443 radars, where one radar is flipped by 90 degrees with respect to the other radar. The point clouds from both radars are fed into a forked CNN, which is later combined to predict key points with 3.2, 2.7, and 7.5 cm localization errors in x,y, z axes, respectively, as compared with skeleton key points estimated by a Kinect RGB-D camera, not a gold standard system. Moreover, it should be noted that this approach did not exploit the Doppler and reflected signal intensity information provided by the radar system at all. This approach suffered from jitter in the animations of the estimated skeletons. In 2022, the authors proposed the use of additional filters [5] to mitigate jitter and provide a more temporally stable skeleton. However,

This work was funded by National Science Foundation Award #2238653.

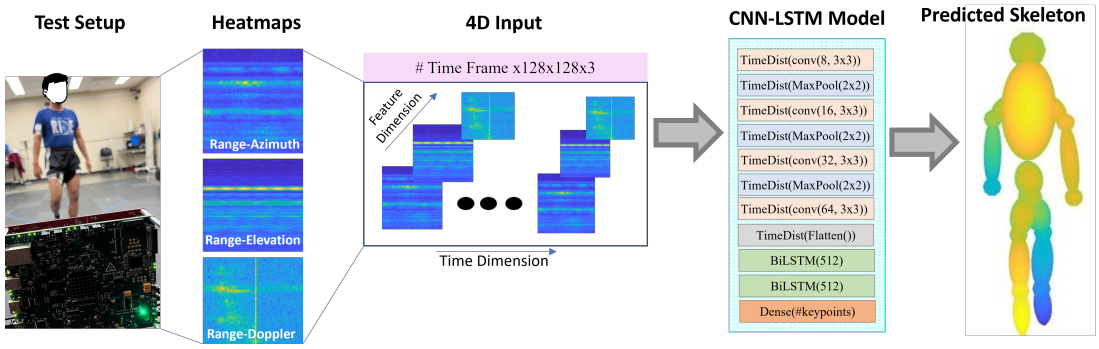

Fig. 1. Proposed skeleton estimation framework.

this approach merely patches over the issues with the initial skeleton estimates, and do not provide for fundamentally more accurate skeleton estimation.

In 2021, Sizhe *et. al* proposed Mars [6], which utilizes IWR1443 radar and off-the-shelf software provided point cloud information (x,y,z coordinate) as well as Doppler and intensity information for each point. The accuracy for 19-point predictions in the x, y, and z, directions respectively were reported as 6.99, 4.07, and 6.65 cm, respectively, again based on comparisons with skeletons estimated by a Kinect camera.

Current works on RF skeleton estimation are thus limited by several key issues. First, most current methods rely on radar point clouds, provided by the software that comes with commercially available RF sensing systems, as input to DNNs to learn the skeletal key points. Thus, the Doppler measurements of the radar and the temporal constraints incurred during movement that are reflected in the micro-Doppler signature are ignored. In contrast, physical temporal correlations are essential aspects of human biomechanics, which also constrain the temporal relationships between the skeletal key points. Thus, considering temporal correlations is a critical aspect of skeleton estimation that has been ignored in the estimation process of currently proposed techniques.

Second, the use of Kinect as ground truth is inappropriate, because Kinect has significant estimation error during skeleton tracking. It has been shown in [7] the Kinect estimates a simplistic form of the actual skeleton, and Kinect's estimation is erred compared to the marker-based skeleton estimation.

Third, current techniques rely on complex models that require large amounts of data; thus, exploring lighter-weight models that can give improved estimates is important for the achievement of practical skeletal estimation, especially using mobile computing platforms.

This work aims to address these challenges by proposing a model trained with just 17 thousand samples that utilizes not just a Convolutional Neural Network (CNN), but also an Long Short-Term Memory (LSTM) recurrent neural network to capture the temporal correlations that are critical to human kinematics. Moreover, the proposed approach utilizes not just spatial data from the radar, but the entire 4D radar tensor, from which range-Azimuth, range-Elevation, and range-Doppler maps can be extracted and provided as input to a DNN model. The model predicts the x, y, and z coordinates of different joints of the body part by considering Vicon MoCap measurements as the ground truth. More details of the proposed skeleton estimation technique are given in Section II, while Section III presents results for the estimation accuracy using synthetic data and with real data, computing estimation error in comparison with that of simultaneously acquired Vicon measurements. The paper concludes in Section IV.

II. Proposed Approach

To make the most of the information that the radar can sense, an RF-based skeleton estimation framework is proposed that utilizes all of the spatial and velocity measurements provided by a radar, as illustrated in Figure 1. The proposed framework uses range-azimuth, range-elevation, and range-Doppler maps as inputs to the neural network. Through a CNN-LSTM network, the temporal features from the time dimension are extracted. The 4D inputs are fed into a CNN-LSTM model, which consists of 4 convolutional layers and two BiLSTM layers. The convolutional layers are associated with the Time-Distributed wrapper, which allows the application of a convolution to every temporal slice of the input. This helps to extract the temporal features from the various 2D maps. The time-distributed convolutional layers are followed by the maxpooling layers to extract the spatio-temporal features corresponding the reflection from the human body. These spatio-temporal features contain information regarding the range, azimuth angle, elevation angle, Doppler, and the intensity of the moving human targets in front of the radar. The BiLSTM layers learn the dependencies of the extracted spatio-temporal features between the consecutive time-frames. Finally, the key points (x,y,z coordinates of the joints) are predicted through a dense layer with linear activation.

The loss function of the model is computed as the mean squared differences between the Vicon ground truth and the output of the predicted dense layer. For prediction of K keypoints, the loss function is given as follows:

$$Kpoints_{loss} = \frac{\sum_{i=1}^{k}(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2}{3K} \quad (1)$$

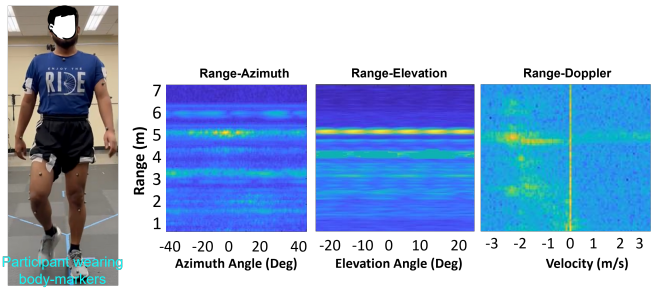

Fig. 2. Bio-markers were attached to the different body joints of the participants during the experiment (left). The raw adc radar data has been processed to generate 3 different feature heatmaps (right).

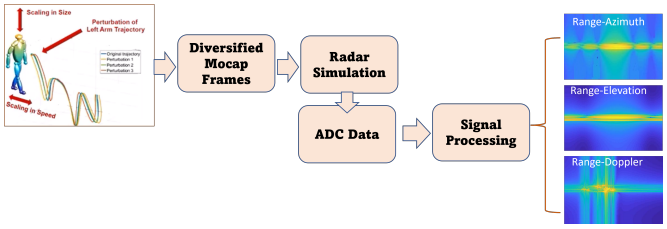

Fig. 3. Simulation of raw radar data utilizing the diversified motion capture data.

As the proposed framework utilizes the raw Range-Azimuth, Range-Elevation and Range-Doppler maps as the 3 feature inputs along with the time being the fourth dimension, the learning model has access to all the acquired information by the mmwave radar sensors. This includes the noise and clutter of the surrounding environments as well. Therefore, the feature extractor backbone network is robust to noise and other environmental factors.

III. Experimental Results

A. Data Acquisition and Pre-Processing

Both RF and MoCap data were collected simultaneously for a free-moving human walking. The data were collected in a laboratory setting, where a Vicon MoCap System with 8 cameras records the coordinates of bio-markers attached to the joints of the body. The RF data were collected with TI IWR2243 cascade radar operated on 77-81 GHz band with 12 Tx antenna and 16 Rx antenna, 1 meters above the ground.

Four healthy participants of different ages, heights, and weights participated in an IRB-approved study in which each participant walked back and forth in front of the radar for 10 minutes. A total of 40 minutes of RF data was acquired. The radar data were collected at 10 frames per second (FPS), whereas the Vicon captured the ground truth skeleton coordinates at 100 FPS. Therefore, the Vicon data was down-sampled to match the radar's frame rate. The Vicon and radar frames were synchronized using the frame where the participants made the first turn. There were 20k synchronized frames.

The TI IWR2243 cascade radar was operated in TDM-MIMO settings with 12 Tx and 16 RX, forming a total of 192 virtual channels. Of these, 86 channels correspond to the azimuth virtual antenna and only 4 channels with an aperture size of 7 $\frac{\lambda}{2}$ correspond to the elevation virtual antenna. Consequently, the radar's azimuth resolution is significantly greater than the elevation resolution. There are 256 ADC samples per chirp, and 128 chirp-loops per frame were transmitted. Thus, the raw ADC data was decomposed into range-azimuth (256x86), range-elevation(256x7), and range-Doppler(256x128) maps. FFT-based match filtering is applied in each plane's direction to reveal target features. For example, he range-azimuth heat map is found by computing an FFT along the range dimension, followed by an FFT (128 points) across azimuth angle. As the target moved within a 0.2 to 6-meter range, the range dimension is lowered by truncating the unused range bins. In this way, a 128x128 range-azimuth map is generated. The same procedure is followed to generate 128x128 range-elevation and range-Doppler maps.

B. Simulation of Radar Data from Diversified MoCap

When training deep learning models, it is generally recommended to use a large amount of well-curated, diverse and representative training dataset so that the model can learn generalizable features. This is because deep learning models, particularly neural networks, have many parameters that need to be optimized, and a larger dataset provides more information for the model to learn from and can lead to better performance on unseen data. Training on a large dataset is not an issue for computer vision tasks, but it is a big struggle to accumulate a huge amount of radar data. Therefore, to initially train the model, this work proposes to generate simulated radar data from MoCap data. Diversification of human skeleton motion capture (Mocap) data can be used to generate raw radar data using multiple-input and multiple-output (MIMO) radar simulation with 3 transmit (Tx) antennas (one of them is elevation antenna) and 4 receive (Rx) antennas imitating the antenna configuration of TI2243Boost radar. The Mocap data can be used to create a virtual model of the human body, and this model can be used to simulate how radar signals would reflect off of the various parts of the body. The simulation can be run for different scenarios, such as different body types, sizes, and speeds of movement, and for different positions and orientations of the body relative to the radar. This will generate raw radar data that can be analyzed to extract information about the body's motion, such as walking speed, arm and leg movement, and other details. In prior work [8], diversification was shown to yield a synthetic micro-Doppler signature database that was effective for model training.

In this work, from 20 Kinect MoCap samples of 10 sec long each, 40k radar raw data frames have been synthesized. The synthetic data frames are processed to generate range-azimuth, range-elevation, and range-Doppler heatmaps. The Radar ADC data simulation workflow has been shown in Figure 3.

C. Training and Evaluation

First, the proposed framework is trained on simulated radar data, with Kinect V2 being used as ground truth. A total

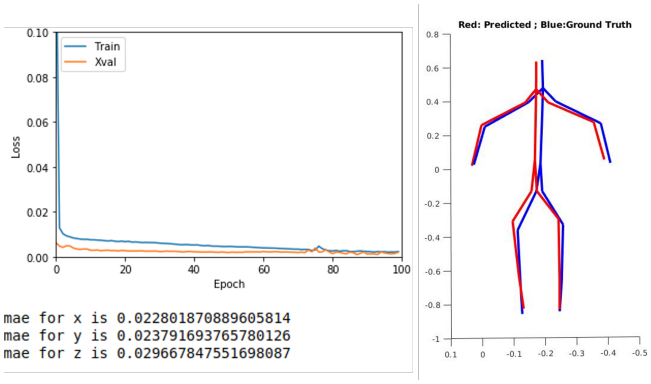

Fig. 4. Human pose estimation for 21 key points prediction with Mocap simulated Radar data in the training while using Kinect as the ground truth. Training loss (left) and predicted pose (right) are also shown.

of 30k frames were utilized in training the framework, and the remaining 10k frames were used in the inference stage. The model was trained for 100 epochs for a 21 key points prediction task. The mean absolute error in x, y, and z were found to be 2.2cm, 2.3 cm, and 2.9 cm, respectively. Figure 5 shows the training and validation loss curve, as well as the predicted skeleton, in comparison with the ground truth skeleton. Note that this result surpasses the accuracy attained by other techniques in the literature, which also compare on Kinect data, but by using fewer training samples.

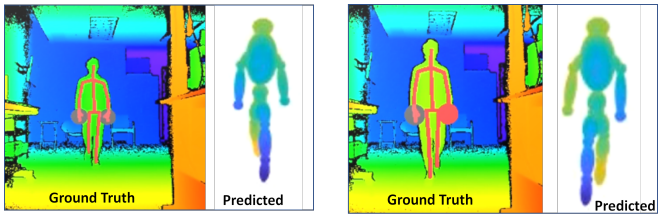

Fig. 5. Visual comparison of ground truth pose vs predicted pose. The ellipsoidal model is applied to the predicted skeleton to make it visually appealing.

Next, we evaluate skeleton estimation accuracy by applying the proposed approach on real data, but with error now computing using Vicon data as ground truth. The proposed framework is trained with real data from 3 participants (17k frames) for 100 epochs with 1e-4 learning rate, Adam optimizer, and 512 GRU in BiLSTM layers with 0.5 dropouts. The Vicon only provided the x,y, and z coordinates for 13 joints. Therefore, in the initial attempts 39 (13x3) neurons were set in the output layer. While tested on 4th subjects data, the mean absolute error across the x, y, and z were 5 cm, 2cm, and 10 cm, respectively, achieved for prediction of 13 key points.

These results show the discrepancy between Kinect-based error evaluation and Vicon-based error evaluation, as well as the impact of the amount of data on the resulting accuracy. Moreover, the results show the benefits of utilizing synthetic data in training, resulting in a lighter weight neural network achieving improved results.

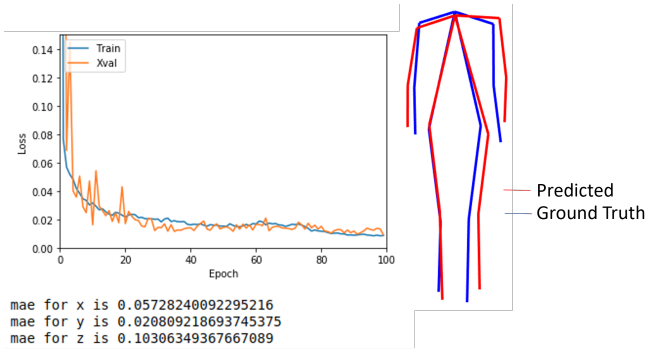

Fig. 6. Human pose estimated while training on real radar data with Vicon Mocap data as ground truth. The training loss and predicted pose are shown side by side.

TABLE I
SKELETON ESTIMATION PERFORMANCE

Data source	Key points	Train/Test Frames	Ground Truth	MAE (cm)		
				x	y	z
Simulated Radar Data	21	30k/10k	Kinect	2.2	2.3	2.9
Real Radar Data	13	17k/3k	Vicon	5.7	2.0	10.3

IV. CONCLUSION

This paper presents a novel framework for human pose estimation using millimeter-wave (mmWave) technology, addressing the limitations of current RF-based pose estimation techniques. The proposed framework utilizes raw mmWave sensor data and a CNN-LSTM architecture to accurately estimate the coordinates of body joints, showcasing its potential for advancing personalized healthcare and improving gait analysis and fall risk assessment. In future work, we plan to expand the study to include more participants and quantify the impact on radar transceiver architecture on accuracy.

REFERENCES

[1] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[2] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–13, 2015.

[3] M. Zhao, D. Katabi, and et.al., "Rf-based 3d skeletons," in *Proc. Conf. ACM Special Interest Group on Data Comm.*, 2018, pp. 267–281.

[4] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[5] S. Hu, A. Sengupta, and S. Cao, "Stabilizing skeletal pose estimation using mmwave radar via dynamic model and filtering," in *IEEE Int. Conf. Biomedical and Health Informatics (BHI)*, 2022, pp. 1–6.

[6] S. An and U. Y. Ogras, "Mars: mmwave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.

[7] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect," in *Int. Conference on Healthcare Informatics*. IEEE, 2015, pp. 380–389.

[8] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Dnn transfer learning from diversified micro-doppler for motion classification," *IEEE Trans. Aerosp. Elec. Sys.*, vol. 55, no. 5, pp. 2164–2180, 2018.