Demo: Hands-Free Human Activity Recognition Using Millimeter-Wave Sensors

Soo Min Kwon, Song Yang, Jian Liu, Xin Yang,
Wesam Saleh, Shreya Patel, Christine Mathews, Yingying Chen
Rutgers University, New Brunswick, NJ 08901, USA
Email: {smk330, sy540}@scarletmail.rutgers.edu, jianliu@winlab.rutgers.edu,
{xy213, ws394, sp1729}@scarletmail.rutgers.edu, {christine.m, yingying.chen}@rutgers.edu

***Abstract*—In this demo, we introduce a hands-free human activity recognition framework leveraging millimeter-wave (mmWave) sensors. Compared to other existing approaches, our network protects user privacy and can remodel a human skeleton performing the activity. Moreover, we show that our network can be achieved in one architecture, and be further optimized to have higher accuracy than those that can only get singular results (i.e. only get pose estimation or activity recognition). To demonstrate the practicality and robustness of our model, we will demonstrate our model in different settings (i.e. facing different backgrounds) and effectively show the accuracy of our network.**

***Index Terms*—hands-free, millimeter-wave, human activity recognition, pose estimation, machine learning**

I. Introduction

Driven by a wide range of real-world applications, Human Activity Recognition (HAR) networks have been widely explored. Many existing hardware devices (e.g. smart watches, mobile phones) are generally used for HAR, but have limitations due to cost and discomfort. Moreover, other vision-based sensing devices, such as cameras, can potentially lead to privacy issues, if leaked. Recent works that utilize wireless infrastructures (e.g. WiFi signals) have been studied to address these issues, yet, these issues still continue to remain as primitive challenges. For example, Wi-Motion [1] analyzes the amplitude changes in Channel State Information (CSI) data that result from human interference. However, extracting features and information directly from WiFi amplitude changes can result in occlusions, such as noise. Further, the accuracy in these models can decrease by removing certain subcarriers used by the CSI data. Although this process is necessary for dimension reduction, it is possible that the removal of subcarriers can also withdraw information valuable for HAR.

This demo proposes the use of mmWave sensors to combat the burden of hardware devices and the obtrusive designs of other device-free networks. The advantage of utilizing these sensors is that they are less susceptible to noise and can be highly accurate when sensing the range of the object. In addition, rather than a feature selection-based network for HAR, such as [2], we show that we can leverage the recent works of human pose estimation for activity classification. Though advantageous, maneuvering mmWave sensors in this manner is nevertheless a challenging task. Firstly, the sensor data does not hold any relative information on human pose estimation. Thus, recreating a skeleton on mmWave data is difficult. Secondly, studies utilizing mmWaves have not been widely explored. This results in time constraints due to having to collect a plethora of data to train our network. To further address these challenges, we employ existing work on pose estimations and show a network that can learn on smaller amounts of data. We summarize the major contributions of our work as follows:

- We propose a hands-free system using mmWave sensors that can achieve HAR and create a pose estimated skeleton performing the classified activity. The proposed system is robust and can remain accurate during environmental changes (i.e. change in scenario).
- We explore the fiducial features in our mmWave signals and propose a Convolutional Neural Network (CNN) as the base of our model.
- We explore different methodologies to address the scarcity in our collected data. We build a prototype that can train on static data (i.e. no body movement), but can still test and recognize dynamic data.
- We develop a teacher-student framework to guide the processed mmWave data to learn human pose estimations. We leverage this to make HAR.

II. System Overview

In this section, we elaborate on the architecture of our proposed framework as shown in Figure 1. Our network consists of four major elements: data collection & processing, teacher-student network, training process and testing process.

A. Data Collection & Processing

Firstly, the person performs an activity in front of the camera and sensor setup. The mmWave sensor captures reflected signals, while the camera synchronously takes pictures (i.e. person is the focal point of both camera and sensor). Secondly, the images taken by the camera are processed through OpenPose [3], which provides human body key points to be used as labels. OpenPose provides a total of 18 key points formatted as a 2-D matrix. However, we detach some of the other given features (e.g. confidence) and work directly with the key points. These points are in the form of a standard coordinate system with two variables, x and y. Thirdly, the

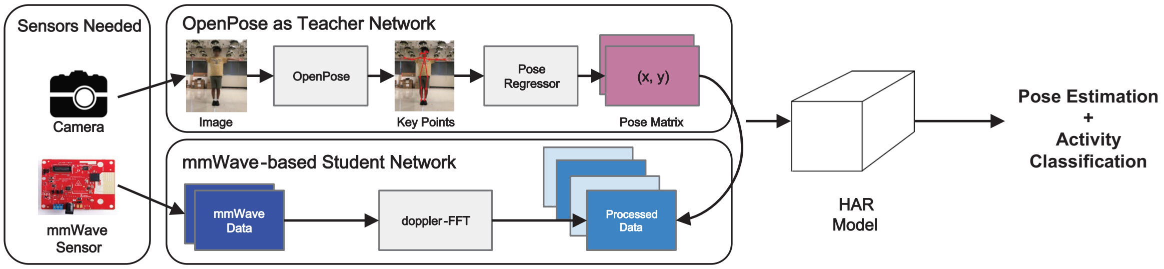

Fig. 1. Proposed mmWave sensor-based framework.

mmWave sensor data is processed by performing the doppler-FFT. The doppler-FFT is, in simpler terms, a 2-D Fast Fourier Transform, along the rows and columns respectively. The key point labels given from OpenPose and the processed sensor data are aggregated and served as the training data for our model.

B. Teacher-Student Network

The aggregation mentioned in the previous section transpires from our teacher-student network, similar to that of [4]. In order for our mmWave data to estimate human poses, we must append the key point labels to its respective sensor data, similar to a supervised machine learning task. This data is forwarded to our HAR model to learn and make predictions.

Our current HAR model is built using a CNN with 4 convolution layers, each followed by a batch normalization, ReLU activation, and a dropout layer. The final dense layer has a total of 36 neurons, reshaped from the 18 by 2 matrix given from the key point labels. The model was trained with a total of 1050 samples for 150 epochs and an Adam optimizer. The model can classify amongst three different activities: stretching, raising dumbbells and sitting down. The breakdown of the samples was 450, 300, and 300 samples respectively.

C. Training Process

The efficiency in our proposed network lies in the training process. Many related networks need an abundance of either dynamic data or a series of static data. However, in order to classify activities in our model, we only need two to three samples of static data to train on. For example, if we wanted to train our network to recognize a person curling dumbbells, we would need 2 static images for training, as shown in Figure 2. In detail, each time our sensor captures data, we trigger a total of 150 frames. Considering that static data does not change per frame, we can use each frame as a data sample to train our network.

D. Testing Process

Though our network was trained using static data, we further test our model with dynamic data. The dynamic data also contains 150 frames, each frame showing slight movement of the complete activity. Our model makes a prediction on the key points in a single frame, and then consecutively plots to recreate the skeleton performing the activity. The activity classification is done similar to a clustering algorithm, where the true labels act as the centroids of the clusters. The model maps the predictions to the closest centroid and forecasts a result.

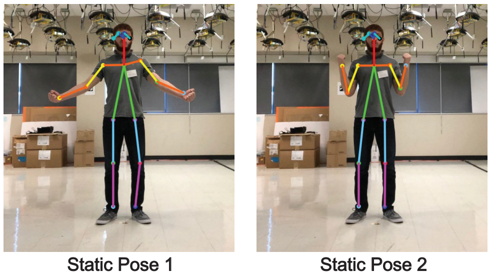

Fig. 2. Two static postures required for dumbbell raising classification.

III. Demonstration Setup

Our demo will showcase our system under two different scenarios (i.e. facing two different backgrounds). This will cover the practicality of our model, in that a change in scenery does not affect our model as long as the distance between the sensor and the human is relatively fixed.

The facilities required by our demo include: (1) a table to deploy the sensor, (2) a chair to classify the ”sitting down” activity, (3) power outlets for the mmWave sensor and laptop. The estimated setup time required for our demo is about 10 minutes.

Acknowledgment

This work was supported in part by the National Science Foundation Grant CNS-1826647 and ARO Grant W911NF-18-1-0221.

References

[1] H. Li, X. Chen, H. Du, X. He, J. Qian, P.-J. Wan, and P. Yang, “Wi-Motion: A Robust Human Activity Recognition Using WiFi Signals,” *arXiv e-prints*, p. arXiv:1810.11705, Oct 2018.

[2] M. Zhang and A. A. Sawchuk, “A feature selection-based framework for human activity recognition using wearable multimodal sensors,” in *BODYNETS*, 2011.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *arXiv e-prints*, p. arXiv:1812.08008, Dec 2018.

[4] F. Wang, S. Panev, Z. Dai, J. Han, and D. Huang, “Can WiFi Estimate Person Pose?” *arXiv e-prints*, p. arXiv:1904.00277, Mar 2019.