RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-wave Radar

Akash Deep Singh*
University of California, Los Angeles
akashdeepsingh@g.ucla.edu

Sandeep Singh Sandha*
University of California, Los Angeles
ssandha@ucla.edu

Luis Garcia
University of California, Los Angeles
garcialuis@ucla.edu

Mani Srivastava
University of California, Los Angeles
mbs@ucla.edu

ABSTRACT

Accurate human activity recognition (HAR) is the key to enable emerging context-aware applications that require an understanding and identification of human behavior, e.g., monitoring disabled or elderly people who live alone. Traditionally, HAR has been implemented either through ambient sensors, e.g., cameras, or through wearable devices, e.g., a smartwatch, with an inertial measurement unit (IMU). The ambient sensing approach is typically more generalizable for different environments as this does not require every user to have a wearable device. However, utilizing a camera in privacy-sensitive areas such as a home may capture superfluous ambient information that a user may not feel comfortable sharing. Radars have been proposed as an alternative modality for coarse-grained activity recognition that captures a minimal subset of the ambient information using micro-Doppler spectrograms. However, training fine-grained, accurate activity classifiers is a challenge as low-cost millimeter-wave (mmWave) radar systems produce sparse and non-uniform point clouds. In this paper, we propose RadHAR, a framework that performs accurate HAR using sparse and non-uniform point clouds. RadHAR utilizes a sliding time window to accumulate point clouds from a mmWave radar and generate a voxelized representation that acts as input to our classifiers. We evaluate RadHAR using a low-cost, commercial, off-the-shelf radar to get sparse point clouds which are less visually compromising. We evaluate and demonstrate our system on a collected human activity dataset with 5 different activities. We compare the accuracy of various classifiers on the dataset and find that the best performing deep learning classifier achieves an accuracy of 90.47%. Our evaluation shows the efficacy of using mmWave radar for accurate HAR detection and we enumerate future research directions in this space.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Hardware** → *Sensors and actuators*; *Sensor devices and platforms.*

KEYWORDS

millimeter-wave, mmWave, radar, human activity recognition, machine learning, neural networks, point-clouds, voxelization, RF

ACM Reference Format:

Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-wave Radar. In *3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems (mmNets'19), October 25, 2019, Los Cabos, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3349624.3356768

*Both authors contributed equally to the paper

mmNets'19, October 25, 2019, Los Cabos, Mexico

ACM ISBN 978-1-4503-6932-9/19/10…$15.00
https://doi.org/10.1145/3349624.3356768

1 INTRODUCTION

Although the recognition and monitoring of human activities can enable safety critical applications–e.g., monitoring disabled and or elderly people living-alone that may need medical attention [1, 10]–the emergence of low cost sensing capabilities have enabled ubiquitous possibilities of human activity recognition for everyday applications. Several context aware applications have recently emerged such as workout tracking and efficacy [12], and factory floor monitoring [6]. Traditionally, human activity has been inferred either through ambient sensors (e.g., cameras) and/or wearable sensors (e.g., smartwatches with IMUs). Although wearables have proven to be effective approaches for human activity recognition, it is not practical to assume that all of the subjects in a space will use wearables that are compatible with the inference model. Ambient sensor approaches are robust to heterogeneous environments since they do not rely on users having a particular device. However, the sensor data from cameras carry a significant amount of ambient information that may be of concern for privacy-sensitive applications. For instance, there have been cases where cameras in the maternity ward of a hospital were used to spy upon female patients [2]. In this context, cameras can be replaced by sensors whose data can provide a sufficient amount of ambient information to realize the same utility, e.g., if we only care about how many patients are in a room or whether there are a certain number of nurses in the ward.

Prior works have shown that less information-rich ambient sensors can effectively infer human activities while not exposing subjects to privacy risks using radio frequency signals. For instance, WiFall [14] showed how WiFi routers can be used to detect whether a human has fallen or not. However, the work is not robust beyond binary classification of two classes that are significantly different from each other. Generally, WiFi has a narrow band (when compared to the high bandwidth of a mmWave radar) and does not have sufficient range resolution to perform robust classification. Radars have begun to emerge as a popular modality for activity recognition[3] as they provide the advantage of operating in any lighting condition and work through a multitude of environmental conditions, e.g., fog and rain. Further, the emergence of millimeter-wave technology has enabled cost-effective distributed sensing applications.

Millimeter-wave (mmWave) technology operates in the frequency range of 30GHz and 300GHz. Since, antenna size is inversely proportional to frequency, the higher you go up in the frequency spectrum, the lower the size of antenna becomes. As a result, mmWave radars are compact in size. Also, we can pack a large number of antennas into a very small space which enables highly directional beam-forming ($\approx 1^\circ$ angular accuracy). Since these radars have a large bandwidth, they have a superior range resolution. Further, new low cost, off-the-shelf radars have led to an increase in the popularity of mmWave based sensing solutions. However, these devices are resource-constrained and, instead of providing raw data, their output is in the form of point clouds[1]. The number of points in each frame captured by the mmWave radar varies, increasing the complexity of constructing a neural network architecture that can process this data as is. Hence, several feature extraction and data pre-processing techniques have been proposed in previous works which convert this data into a format which is constant in size and can be given as input to a neural network [17, 18].

In this paper, we propose RadHAR, a framework for human activity recognition that utilizes point clouds generated from a mmWave radar. To account for the sparsity of the mmWave radar point clouds, RadHAR leverages the notion that human activities typically last over a few seconds and accumulates point clouds over a sliding time window. Each point cloud is voxelized to overcome the non-uniformity of the data and is then fed into a set of classifiers. We collected a new HAR dataset consisting of point clouds using mmWave radar for 5 different classes of activities. We evaluated RadHAR and compared the accuracy of various classifiers on the collected dataset. In our evaluation, the best performing deep learning classifier composed of a set of convolutional layers and long-short term memory layers can achieve an average test accuracy of 90.47%.

Contributions. Our contributions are summarized as follows.

- We propose RadHAR, a framework that performs human activity recognition using a pre-processing pipeline for point clouds generated by mmWave radar.
- We evaluate different machine learning approaches for human activity detection using point cloud.
- We generate a new point cloud dataset for human activity detection and make it available open-source along with the data processing, classifier training and evaluation code, and pre-trained classifiers.

The rest of this paper is arranged as follows. In Section 2, we present the preliminary information required to understand the RadHAR framework. We provide an overview of RadHAR and its implementation in Section 3, and evaluate the classification approaches in Section 4. The related work is presented in Section 5, and we discuss and conclude in Section 6 and Section 7, respectively.

The source code and datasets of RadHAR are available online at https://github.com/nesl/RadHAR.

2 BACKGROUND

We provide the preliminary information necessary to understand the RadHAR framework. We discuss the basics of mmWave radar physics.

2.1 Millimeter-wave Radar

Over the last several years, there has been a growth in low cost single chip radars that work in the mmWave range. One family of such popular devices are Texas Instruments' mmWave radar. These sensors output the point clouds that contain information like x,y,z positions of each point among other data.

[1]To get raw ADC data from these devices, you need to connect them with expensive hardware

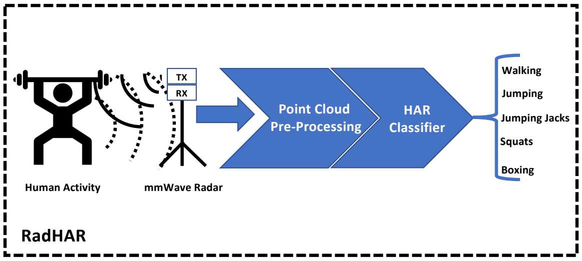

Figure 1: RADHAR framework overview.

Bandwidth and range resolution. Range resolution of a radar is its ability to distinguish between 2 targets present very close to each other. The range resolution and bandwidth are related as,

$$d_{res} = \frac{c}{2B} \tag{1}$$

where d_{res} is the range resolution in m, c is the speed of light in m/s and B is the bandwidth in Hz swept by the chirp of the radar. Hence, if we want a better range resolution, the bandwidth should be high. The maximum continuous bandwidth for the radar that we used is 4 GHz which corresponds to a range resolution of about 4 cm.

3 RADHAR OVERVIEW

The full pipeline of the RADHAR framework is depicted in Figure 1. The framework first collects data from a mmWave radar that is monitoring a human. The point cloud data is pre-processed before being fed into a HAR classifier. We present an overview of each component in detail.

3.1 Data Collection & Pre-processing

We have used TI's IWR1443BOOST [8] radar to collect the new point cloud dataset called *MMActvity* (millimeter-wave activity) dataset. It is a FMCW (Frequency Modulated Continuous Wave) radar which uses a chirp signal. This radar works in the 76-GHz to 81-GHz frequency range. The radar includes four receiver and three transmitter antennas, which enable tracking multiple objects with their distance and angle information. This antenna design enables estimation of both azimuth and elevation angles, which enables object detection in a 3-D plane [7].

3.2 MMActvity Dataset

For data collection, the radar is mounted on a tripod stand at a height of 1.3m. The data from the radar is sent to the laptop via ROS (Robot Operating System) messages over USB. The ROS node running on the laptop is developed and described in [17]. To record and store the data, we use rosbag which another ROS package. Finally, we convert these rosbags into .txt files. These files are then used to create voxelized representation of the point clouds. The data collection and pre-processing pipeline is describe in Figure 3.

Figure 2: Data collection setup.

We have collected the data from two users[2]. The users perform 5 different activities in front of the radar as shown in Figure 2. These activities are: Walking, Jumping, Jumping Jacks, Squats and Boxing. The data is collected in a continuous periods of about 20 seconds for a subject performing the same activity. Some of the data files are longer than 20 seconds. In total, we have collected 93 minutes of data. The description of the dataset can be found in Table 1.

The captured point clouds contains spatial coordinates (x,y,z in meters) along with velocity in meters/second, range (distance of the point the from radar) in meters, intensity (dB) and bearing angle (degrees). The sampling rate of the radar is 30 frames per second.

3.3 Data Pre-processing

We divided collected data files into separate train and test files with 71.6 minutes data in train and 21.4 minutes data in test. To overcome the non-uniformity in number of points in each frame, we converted the point clouds into voxels of dimensions 10x32x32 (depth=10) which makes the input of constant size irrespective points of the number of points in the frame. We decided these dimensions empirically by testing their performance. In our voxel representation, the value of each voxel is the number of data points present within its boundaries. While having large number of voxels may represent underlying information well, it increases the data size by several orders of magnitudes.

Since activities are performed over a period of time, the time window from activities are generated in-order to capture the temporal dependencies. We create windows of 2 seconds (60 frames) having a sliding factor of 0.33 seconds (10 frames). The 2-second window was chosen based on the previous works in human activity recognition from multimodal timeseries datasets [15] and human identification using point clouds [18]. Finally, we get 12097 samples in training and 3538 samples in testing. We use 20% of the

[2]The data is collected from the authors and thus does not require approval from IRB.

Activity	# of data files	Total duration (seconds)
Boxing	39	1115
Jumping Jacks	38	1062
Jumping	37	1045
Squats	39	1090
Walking	47	1269

Table 1: Details of the MMActivity dataset.

training samples for validation. In the *time window voxelized representation*, each sample has a shape of $60 * 10 * 32 * 32$.

3.4 Classifiers

We evaluate different classifiers on the MMActivity dataset. We train Support Vector Machine (SVM), multi-layer perceptron (MLP), Long Short-term Memory (LSTM) and convolution neural network (CNN) combined with LSTM. We compare the inference capability of these classifiers on the same train and test split of MMActivity dataset. These deep learning classifiers are generally adapted in a wide range of applications. LSTM and CNN combined with LSTM architectures are inspired from [18]. Next, we explain the details of classifiers (data inputs, architectures and training details).

3.4.1 SVM Classifier. The input to the Support Vector Machine (SVM) classifier is generated by flattening the time window voxelized representation ($60 * 10 * 32 * 32$) and then applying the Principal Component Analysis (PCA) for dimensionality reduction. We used PCA to reduce the dimensions of data from 614400 ($60*10*32*32$) to 6000 which explained 80% of variance in data. SVM with RBF kernel was used.

3.4.2 MLP Classifier. It is composed of fully-connected layers and an output layer. We flatten the time window voxel representation ($60 * 10 * 32 * 32$) of the sample to create input size of 614400 dimensions for the MLP classifier. The MLP classifier has 4 fully connected layers followed by the output layer. We use dropout layers to avoid overfitting. It has 39.35 million trainable parameters.

3.4.3 Bi-directional LSTM Classifier. A bi-directional LSTM layer consists of two LSTM layers operating in parallel. The input to the first layer is provided as-is whereas the input is reverse copy of the data for the the second layer. As a result, a bi-directional LSTM layer preserves the information from both the future and the past. This network consists of the Bi-Directional LSTM layer followed by the 2 fully connected layers and an output layer. The input (60*10240) to the network is created by preserving the time dimensions (60) and flattening the spatial dimensions in the samples (10*32*32). We used Bi-Directional LSTM with size of 64 and 64 hidden units. The Bi-directional LSTM classifier has 5.29 million trainable parameters.

3.4.4 Time-distributed CNN + Bi-directional LSTM Classifier. Time-distributed CNN applies CNN layers to every temporal slice of the input data. The architecture of Time-distributed CNN + Bi-directional LSTM classifier consists of 3 time distributed convolutional modules (convolution layer + convolution layer + maxpooling layer) followed by the bi-directional LSTM layer and an output layer. Overall the network has 291k trainable parameters. This classifier is directly trained on the the input sample with its time and spatial dimensions.

Training and implementation. The classifiers were implemented using Sklearn and Keras. We use GridSearchCV function from sklearn to optimize the hyperparameters (C and gamma) of the SVM. Adam optimizer with a learning rate of 0.001 was used to train deep learning classifiers. The models with minimum loss on the validation data was saved after training for 30 epochs.

S.No	Classifier	Accuracy
1	SVM	63.74
2	MLP	80.34
3	Bi-directional LSTM	88.42
4	Time-distributed CNN+ Bi-directional LSTM	90.47

Table 2: Test accuracy of different activity recognition classifiers trained on the MMActivity Dataset.

We now evaluate the aforementioned approaches.

4 EVALUATION

As shown in Table 2, the classifiers trained for the human activity recognition have different performance. The table reports average results of 5 different training sessions. The SVM classifier has poor performance with test accuracy of 63.74%. One reason might be the input to the SVM is not using the domain specific feature extraction approaches as used by Kim et al.[9] where they use a Doppler radar (2.4 GHz) and then convert the output into micro-Doppler signatures. All the three deep learning classifier are working directly on the time window voxel data. MLP classifier consists of the fully connected layers which doesn't assume anything about the input data and has test accuracy of 80.34%. Bi-directional LSTM classifier tries to learn the sequence of input data. The input data to the LSTM preserve the time component. Since human activities are performed over a duration and due to preserving time sequence for input to LSTM, the LSTM performance is significantly better then the MLP with test accuracy of 88.42%. The best performing classifier is Time-distributed CNN + Bi-directional LSTM which has test accuracy of 90.47%. Time-distributed CNN layers learn the spatial features from the data, as the point clouds are spatially distributed and the Bi-directional LSTM layers learn the time dependency for the activity windows.

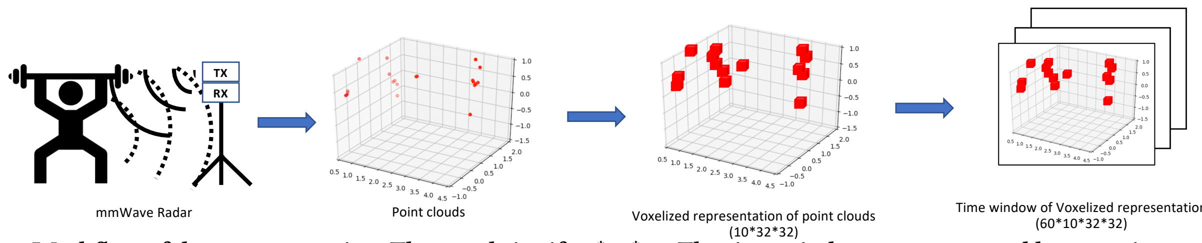

Figure 3: Workflow of data preprocessing. The voxel size if 10*32*32. The time windows are generated by grouping 60 frames (2 second) together.

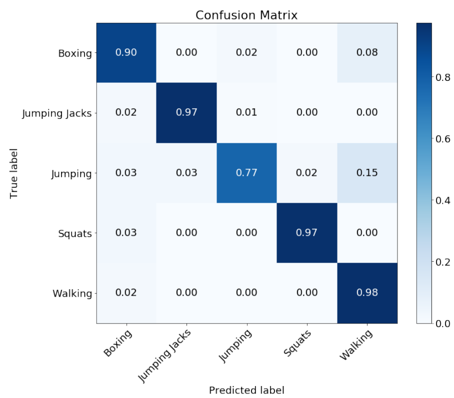

Figure 4: Confusion matrix of time-distributed CNN + bi-directional LSTM classifier.

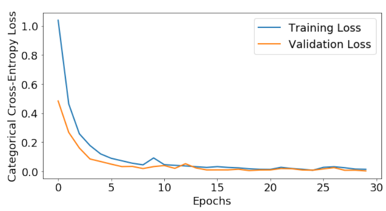

Figure 5: Variation of training and validation loss of Time-distributed CNN + Bi-Directional LSTM

Our evaluation shows specialized spatial and temporal layers in deep learning architectures can result in boost in accuracy. The confusion matrix for one of the trained Time-distributed CNN+ Bi-directional LSTM classifier is shown in Figure 4. As seen from the figure, the activity of jumping and boxing is confused with the walking. The reason might be the similarities in the data for these activities. The variation of the training loss and the validation loss for Time-distributed CNN + Bi-directional LSTM classifier with the training epochs is shown in Figure 5.

Voxelized representation with velocities. All the evaluation presented above used the voxel representation, where the value of each voxel is the number of data points present within its boundaries. We also evaluated with the voxelized representation where the value of each voxel is the sum of the velocity of all the points present within its boundaries. The Time-distributed CNN + Bi-directional LSTM classifier trained on velocity voxel representation also had the similar performance as shown in Table 2. We now discuss the works directly related to the RadHAR framework.

5 RELATED WORK

Human activity recognition is widely explored using various sensing modalities. Researchers have used sensors like cameras and inertial measuring units [4, 13, 15], sound [15, 16] and WiFi [14]. However, optical sensors like cameras capture a significantly large amount of information and sensors like inertial measuring units need to be present on the user body.

Micro-Doppler spectrograms using radars for human activity recognition have been studied in detail over the last decade [3, 5, 9]. In [9], the authors use a Doppler radar to collect data of 12 subjects performing 7 different activities. They create micro-Doppler spectrograms from this data and extract 6 features from it to train an SVM classifier. Unlike our work, the radar used here is in the S-band (2-4 GHz).

Recently, researchers have exploited low-cost single-chip mmWave radar systems for person identification and tracking [18] and human activity recognition [17]. In [17], the authors convert the point cloud data into micro-Doppler spectrograms before using a CNN to classify it. In [18], the authors use voxelized representation of point clouds for human identification using a LSTM and a CNN + LSTM classifier. Our deep learning classifier architectures are inspired from [18], however, we are targeting a different problem of human activity recognition.

In this work, we show that the time window voxel representation of the sparse points clouds can be used for human activity recognition. We evaluate multiple classifiers and achieve test accuracies as high as 90% percent for the deep learning classifier based on the convolutional layers and long-short term memory layer. Our evaluation shows that deep learning approaches can achieve comparable performance to the previous domain specific feature extraction approaches like used by Kim et al.[9].

6 DISCUSSION AND FUTURE RESEARCH

We now discuss the limitations of our approach and enumerate future research directions.

Spatial and temporal dependencies in point clouds. As shown in Table 2, MLP classifier has poor performance. The reason might be due to fact the fully connected layers in MLP classifier makes no spatial and temporal assumption about the data. On the other hand, Time-distributed CNN + Bi-directional LSTM classifier assumes spatial and temporal dependency in the data and hence performs better.

Limitations of voxelized representation. Voxels result in significant increase in the required memory and computation. This can be seen in dimensionality of each input sample (60*10*32*32 = 614400), which has to be processed by the deep learning classifier. This begs the need for neural networks which are trainable directly on point clouds. One such network is the PointNet [11] which can be used for applications like object classification, part segmentation and scene semantic parsing.

7 CONCLUSION

We presented RadHAR framework for HAR using the time window voxel representation of sparse mmWave radar point clouds. Our evaluation of the classifier shows that deep learning classifiers can be directly trained on the time window voxel representation and can achieve test accuracy greater than 90%. The classical machine learning approaches require domain specific feature extraction and shows poor performance on the voxels. Deep learning classifiers are able to learn the feature extraction transformation by directly training on the voxels. The classifiers which are designed to handle the spatial and temporal dependencies in data perform better the fully connected deep learning classifier.

We also presented a new MMActivity dataset of point clouds along with source code and pre-trained models which are available open-source.

ACKNOWLEDGMENTS

The research reported in this paper was sponsored in part by the National Science Foundation (NSF) under award #CNS-1329755, by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, and by the Army Research Laboratory (ARL) under Cooperative Agreement W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARL, DARPA, NSF, SRC, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. 2015. Physical human activity recognition using wearable sensors. *Sensors* 15, 12 (2015), 31314–31338.

[2] John Bonifield. 2019. Cameras secretly recorded women in California hospital delivery rooms. https://www.cnn.com/2019/04/02/health/hidden-cameras-california-hospital/index.html

[3] Bahri Çağlıyan and Sevgi Zübeyde Gürbüz. 2015. Micro-Doppler-based human activity classification using the mote-scale BumbleBee radar. *IEEE Geoscience and Remote Sensing Letters* 12, 10 (2015), 2135–2139.

[4] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*. IEEE, 168–172.

[5] Dustin P Fairchild and Ram M Narayanan. 2016. Multistatic micro-Doppler radar for determining target orientation and activity classification. *IEEE Trans. Aerospace Electron. Systems* 52, 1 (2016), 512–521.

[6] Jinwen Hu, Frank L Lewis, Oon Peen Gan, Geok Hong Phua, and Leck Leng Aw. 2014. Discrete-event shop-floor monitoring system in RFID-enabled manufacturing. *IEEE Transactions on Industrial Electronics* 61, 12 (2014), 7083–7091.

[7] Texas Instruments. 2018. IWR1443BOOST Evaluation Module User's Guide. *http://www.ti.com/lit/ug/swru518c/swru518c.pdf* (2018). Accessed: 2019-07-05.

[8] Texas Instruments. 2019. IWR1443 single-chip 76-GHz to 81-GHz mmWave sensor evaluation module IWR1443BOOST (ACTIVE). http://www.ti.com/tool/IWR1443BOOST Accessed: 2019-07-05.

[9] Youngwook Kim and Hao Ling. 2009. Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Transactions on Geoscience and Remote Sensing* 47, 5 (2009), 1328–1337.

[10] Bingbing Ni, Gang Wang, and Pierre Moulin. 2011. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 1147–1153.

[11] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 652–660.

[12] Chenguang Shen, Bo-Jhang Ho, and Mani Srivastava. 2017. Milift: Efficient smartwatch-based workout tracking using automatic segmentation. *IEEE Transactions on Mobile Computing* 17, 7 (2017), 1609–1622.

[13] Xiao Sun, Li Qiu, Yibo Wu, and Guohong Cao. 2017. ActDetector: Detecting Daily Activities Using Smartwatches. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 1–9.

[14] Yuxi Wang, Kaishun Wu, and Lionel M Ni. 2016. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing* 16, 2 (2016), 581–594.

[15] Tianwei Xing, Sandeep Singh Sandha, Bharathan Balaji, Supriyo Chakraborty, and Mani Srivastava. 2018. Enabling Edge Devices that Learn from Each Other: Cross Modal Training for Activity Recognition. In *Proceedings of the 1st International Workshop on Edge Systems, Analytics and Networking*. ACM, 37–42.

[16] Yi Zhan and Tadahiro Kuroda. 2014. Wearable sensor-based human activity recognition from environmental background sounds. *Journal of Ambient Intelligence and Humanized Computing* 5, 1 (2014), 77–89.

[17] Renyuan Zhang and Siyang Cao. 2018. Real-Time Human Motion Behavior Detection via CNN Using mmWave Radar. *IEEE Sensors Letters* 3, 2 (2018), 1–4.

[18] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mID: Tracking and Identifying People with Millimeter Wave Radar. In *International Conference on Distributed Computing in Sensor Systems (DCOSS)*.