Capturing Human Pose Using mmWave Radar

Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, Jin Zhang
Southern University of Science and Technology
11510569@mail.sustech.edu.cn, 11749198@mail.sustech.edu.cn, 11611712@mail.sustech.edu.cn,
11611801@mail.sustech.edu.cn, 11612512@mail.sustech.edu.cn, zhangj4@sustech.edu.cn.

***Abstract*—Human pose estimation is an important task. Traditional human pose capturing systems are based on images or videos, which may suffer from bad light and raise the concerns of privacy. In this paper, we proposed an accurate human pose estimation system using the 77GHz millimeter wave radar. It is the first time that people use off-the-shelf millimeter wave radar to complete such a task. Our system requires no camera or specific sensors on the body to estimate the human skeleton. The system first uses two radar data to generate heatmaps and then adopts CNN to transform two-dimensional heatmaps into human pose. We use coordinated heatmaps from radar and visual inputs extracting from camera together to train the designed network. Based on our dataset and system, the proposed method achieves an average OKS value of 0.705 and 0.877 in AP 50.**

***Index Terms*—sensing, mmwave radar, skeleton, neural network**

I. Introduction

Capturing human pose is a long-standing problem and plays an important role in human-computer interaction. It mainly concerns about locating and recognizing different parts of the human body, such as ankles, shoulders, wrists and so on. With the location of each part, system can generate dynamic skeletons of human bodies. This technique can be widely used in intelligent surveillance, gaming, activities analysing, smart home, etc.

Recently, we have witnessed many great improvements in human pose capturing technique using traditional computer vision methods [1], [2]. However, those methods require cameras, which may raise people's concern on privacy and are limited to light condition. To overcome the above problems, researchers start to investigate wireless sensing system by using Wi-Fi or audio signal. Compared to camera-based system, the wireless sensing system can protect the privacy issue and is robust to light conditions. However, the wireless signal like Wi-Fi is not born to achieve such a goal. Some may require people to carry certain equipment, like cell phones. Most of time, the wireless signal can only achieve a rough estimation for position. They can not capture an accurate human pose like we intend to do.

Recently, researchers have developed an RF-based human pose capturing system [3]. They use self-designed RF generating device to do human localization and human skeleton capture. The device is quite large and hard to deploy. In the past few years, the improvement of millimeter wave radar makes them cheaper and smaller. We adopt those advantages and propose our human pose capture system using mmWave radar.

This paper presents a human pose capturing system using millimeter wave radar. We aim to achieve human pose capturing by using off-the-shelf commercial radar devices. The system is free of privacy concern, independent to light condition and easy to deploy. Our millimeter wave radar works on a much higher frequency (77GHz) and the device is quite small due to the smaller wave length. We intend to leverage the information we can extract from the wireless signal to design the human pose capturing system by using neural network. If well trained, our system can achieve human skeleton capturing in both indoor and outdoor scenarios.

Challenges: Firstly, the features extracted from radar devices are not as intuitive as images. Generally, people are not able to read information directly from radar signal while our system should only take radar information as input and estimate human skeleton. Secondly, using the off-the-shelf radars means there is no place for customizing. To achieve such goal, only one radar is not enough because the mmWave radar devices sold in the market mostly contain one radar array only. They can only extract information from two dimensions while we need at least three to rebuild human skeleton.

To overcome these challenges, we design our system in the following ways. First, we use two radar devices simultaneously in both horizontal and vertical direction and combine two signal together in features extraction. In order to realize two-radar capturing, we solve the coordination problems. Then, in order to leverage the radar signal, we design a CNN network to build the relationship between radar extracting features and human pose. An additional camera-based human pose capturing method is used to supervise the learning. If well trained, our system can only take in radar generated information and produce the human skeleton. The paper has the following contributions:

- We develop a human pose capturing system using mmWave radar, the first system to obtain human skeleton by using off-the-shelf 77GHz radar devices.
- We leverage the signal processing algorithm and deep learning technique to ensure our proposed system is able to learn to capture the human pose.
- We conduct the whole experiment and evaluation by establishing our own dataset and prototype.

Professor Jin Zhang from Southern University of Science and Technology

II. Related Work

Recent years we have witnessed many great improvements in estimating human pose and people tracking technique using both and computer vision methods and wireless sensing.

(a) Computer Vision: Recognizing human pose is a fundamental task in computer vision. They often take ordinary RGB images or RGB-Depth photos [4] as input. There are two main approaches to solve this problem: top-down [2], [5], [6] and bottom-up [7], [8] methods. The top-down methods first use a human detector and then perform pose estimation. The other is the bottom-up method, which means the system detects every keypoint in the image and later connect them in a certain way.

(b)Wireless Sensing: The past decades have seen many processes made in tracking human and detecting behaviors using wireless systems. Some device-based system requires to carry some certain wireless devices, like their mobile phones [9], which make their system limited by many situations. Some existing works use combined WiFi and cell phone to do multi-person localization [10]. Their accuracy limitations make that system unable to do tasks like skeleton capturing. The other device-free system can achieve such a goal by only analyzing the radio signal reflected from the human body. Recently, there is a new approach to estimate human pose using RF signal [3], [11]. They can also detect some specific behavior, like falling [12]. In such implementation, it uses great more antennas and larger equipment by customizing RF transmitting device.

(c) mmWave Radar: The millimeter wave radar is often used in autopilot, material recognition [13] or hand gesture recognition [14]. Recently, several papers have demonstrated its application in detecting human behaviors, like sleeping condition, or activities like falling, walking and so on [15]. Those works are detecting human behaviors at a high level, which is fundamentally different from ours.

III. The Methodology of mmWave Radar

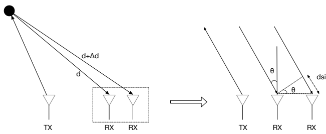

Fig. 1. Angel of Arrival Estimation

A. Radar Primer

A common radar equipment can transmit and receive radio signal in one single board. Due to the time delay during the reflection, information like distance, velocity and angle can be detected. We choose FMCW (Frequency Modulated Continuous Wave) radar in our system, because in order to capture the human skeleton, we need the accurate spacial information of each part of the human body, that is, the distance and angle. In order to coordinate the two radars, the velocity of the object is required. This task can be completed by sending multiple chirps in one frame by FMCW radar.

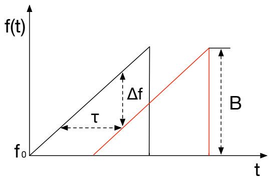

Fig. 2. FMCW wave

Those algorithms will be explained one by one in the following section.

B. Distance Estimation

The FMCW radar transmits signal called chirp. Each chirp is a sinusoidal wave whose frequency changes linearly with time, as the black line shown in Figure 2. The sweeping range of frequency is known as bandwidth B.After transmitting a chirp by TX antenna, the RX antenna will obtain a chirp reflected from an object, as the red line shows in Figure 2. The reflected wave is a delayed signal of the original one. The delay τ is proportional to the frequency difference Δf. Thus we can estimate the distance d between the radar and detected the object by the equation:

$$d = \frac{\tau c}{2}. \tag{1}$$

C. Angle Estimation

In FMCW radar, we use multiple antennas to estimate the angle of the detected object, see Fig.1. The differential distance caused by the location on two RX antennas, see Δd in Fig.1, results in a phase change ω between two receiving chirp:

$$\omega = \frac{2\pi \Delta d}{\lambda}, \tag{2}$$

where λ is the wave length. The phase difference ω can be calculated by 2D-FFT, see Figure 3 and Figure 4. We can use calculated phase difference ω to estimate the angle of the detected object by

$$\theta = \sin^{-1}(\frac{\lambda\omega}{2\pi d}). \tag{3}$$

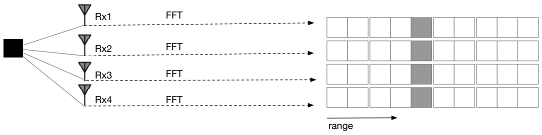

Fig. 3. 2D-FFT in Range-Angle

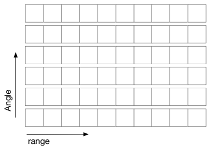

Fig. 4. 2D-FFT in Range-Angle

D. Velocity Estimation

The velocity of the detected object can be estimated by transmitting two separate chirps consecutively. The measured phase difference ω calculated by FFT is caused by a moving distance vT_c of the object. So the velocity of the object can be estimated by the following equation:

$$v = \frac{\lambda\omega}{4\pi T_c}. \tag{4}$$

IV. System Design

A. System Overview

This section presents an overview of the system, see Figure 5. The overall target of our system is to take radar information as input and use them only to estimate human pose. The system consists of:

- The radar front end—— two fixed indentical radar. The radar takes in reflected FMCW chirp and stores them for processing in the next component. A camera is needed when collecting training data.
- A signal processing and heatmap generation component, which processes raw data stored from the radar front end and generates heatmaps for the input of CNN network.
- A coordination component to coordinate two radars. It also works for the camera in the training stage.
- A CNN network that takes range-angle heatmap as input and produces confidence map for each component of the human body. If in the training stage, openpose is used to supervise the learning.
- A skeleton generation component to locate keypoint position from confidence map and link them to produce the human skeleton.

B. Radar Front End

The radar front end is a two-radar system which generates FMCW chirps and stores reflected signal data. Due to the arrangement of antenna array, the radar is only sensitive to one specific dimension. However, the necessary information needed to rebuild human pose should contain three-dimensional information, while one radar only gives two. To solve the problem, we use two identical radars in a fixed coordinate to obtain information from both horizontal and vertical. This component will store the original data sampled from FMCW chirp. In the training stage, this component will add one more fixed camera to collect photos for labelling.

C. Signal Processing and Heatmap Generation

The signal processing component takes stored signal data as input and produces distance-angle heatmaps as output. Firstly, we perform static cluster removal on original raw data before 2D-FFT in order to remove noise and achieve a better result in moving targets. In order to get spatial information of reflected human parts, we first perform FFT in range dimension and followed by another FFT in antenna dimension. This algorithm is also known as 2D-FFT, see section III. The heatmap represents the reflected signal strength in both range and angle. For example, a dominant peak in a specific point means a main object. We use two radar separately to obtain range and angle information in horizontal and vertical views (x, y and y, z respectively), see Figure 6. The result is stored in a two-dimensional matrix in order to store complex values.

D. Coordination

One of the main challenges of implementing such a system is coordination because we need to use two identical radars at the same time. If collecting training data, we should use one more camera. Bad coordinated data may cause the network unable to learn from the radio signal.To deal with the coordination problem within the three devices, we use a quick hand waving movement as a landmark at the beginning of every data collection. The direction of hand waving is towards the radar to achieve a vivid max speed, see Figure 7. Another camera is placed on the right side of the subject to record the hand movement and we can find the same max speech from video. The side camera can be coordinate with the main front camera using time stamp.

E. CNN Network Design

The designed network will take range-angle heatmap generated from signal processing components as input. Our network will encode useful information from the two-dimensional heatmaps and later decodes them into the view of the camera.

The CNN network needs supervision while training. Labelling all the data by hand is a heavy and almost impossible job. Inspired by the work in [3], we can use CV model to produce labels for our dataset. We use a pre-trained COCO model in [16] to supervise the training of our network. Our network is designed to predict 14 keypoints of the human body, including nose, shoulders, elbows, wrists, hips, knees and ankles.

The network structure is mainly illustrated in Figure 8. At the beginning of our model, we firstly perform 3D convolution on both horizontal and vertical heatmaps. The reason to use 3D convolution is that the reflection of the human body may not always occur in one heatmap, so it is necessary to take the spatiotemporal information to predict all the keypoints. The system later uses a 3D fractionally strided convolutions to decode the features into the confidence map generated from the camera. To supervise the decode network, we generate the confidence maps C_p from the detected keypoint location. We apply a Gaussian filter to each point to produce confidence maps. The value at location p for keypoint k in position x_k in the confidence map is defined as

$$C_k(p) = exp(-\frac{\|p - x_k\|_2^2}{\sigma^2}). \tag{5}$$

The surpervise process is minimizing the loss between its prediction and the prediction of openpose network. The loss function $L(G, P)$ is defined as the summation of binary cross entropy of ground truth G and prediction P for each point the confidence maps:

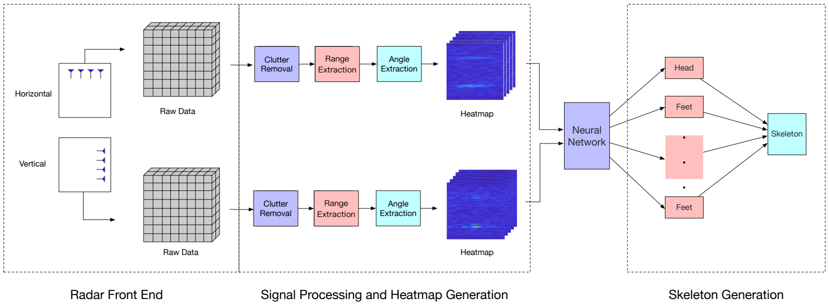

Fig. 5. System Structure

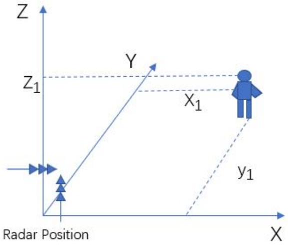

Fig. 6. Radar Deployment

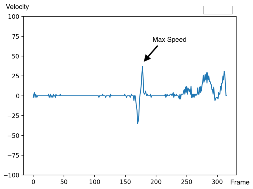

Fig. 7. Speed of waving hands from radar

$$L(G,P) = -\sum_{c}\sum_{i,j} P_{ij}^{c}\log G_{ij}^{c} + (1-P_{ij}^{c})\log(1-G_{ij}^{c})), \quad (6)$$

where G_{ij}^{c} and P_{ij}^{c} are the confidence scores for the pixel value at the position i, j on the confidence map C_k.

CNN Network: The encoding network takes 60 frames (2 seconds) of radar heatmaps from both horizontal and vertical dimensional as input. It uses 5 layers of 9*5*5 3D convolution. We use batch normalization in every layer and no max pooling. As for the activation functions, we use ReLU at the end of every layer.

Fractionally Strided Convolution Network: After the CNN, the rest of network takes in the concatenated features. We use 3D fractionally strided convolution [17] to decode the features into the human pose. The decoding network includes 4 layers of 3*6*6 deconvolution with a stride of 1*2*2. Specially, we use sigmoid activation in the last layer and the loss function is binary cross entropy. The other layers use Parametric ReLU (PReLU) [18].

TABLE I
RADAR CONFIGURATION

Parameter	Value	Unit
Frequecny	77	GHz
Slope	64.019	MHz/us
Chirp	16	per TX in a frame
Samples	512	per chirp
Frame Duration	33.333	ms
Range resolution	0.0458	m
Velocity resolution	0.7417	m/s

F. Human Pose Generation

The output of the designed network produces confidence maps for all keypoints of people. In the post-processing component, we detect the exact location of each keypoint. Inspired by traditional ways in the computer version, we perform non-maximum suppression (NMS) on the confidence maps to get the exact location of keypoint candidates. Since in our implementation, we only consider there will be only one person in the scene, so after we obtain the keypoint candidates, we can simply line them up to get the skeleton.

V. SYSTEM IMPLEMENTATION

The radar we used is produced by Texas Instruments [13]. The IWR1642 radar we used can generate an initial 77GHz FMCW wave with a maximum bandwidth of 4GHz. We use a DCA1000 to connect IWR1642 to perform raw data capture, see Figure 9. The radar is set to 2TX and 4RX modes to generate 8 virtual antennas. Both radars we use in this experiment are initialized with the same parameters. During the whole experiment, we fix two radars and mobile phone in a certain relative position to make sure our network can learn a proper relationship between heatmaps and the human skeleton.

Our networks training are implemented in Keras using Tensoflow backend. The batch size is 8 and we use Adam optimization algorithm .

VI. SYSTEM EVALUATION

A. Experimental Setup

We present how we collect our dataset in this section. To collect data from different environments, we place the two connected AWR1643 and DCA1000 EVM and camera in

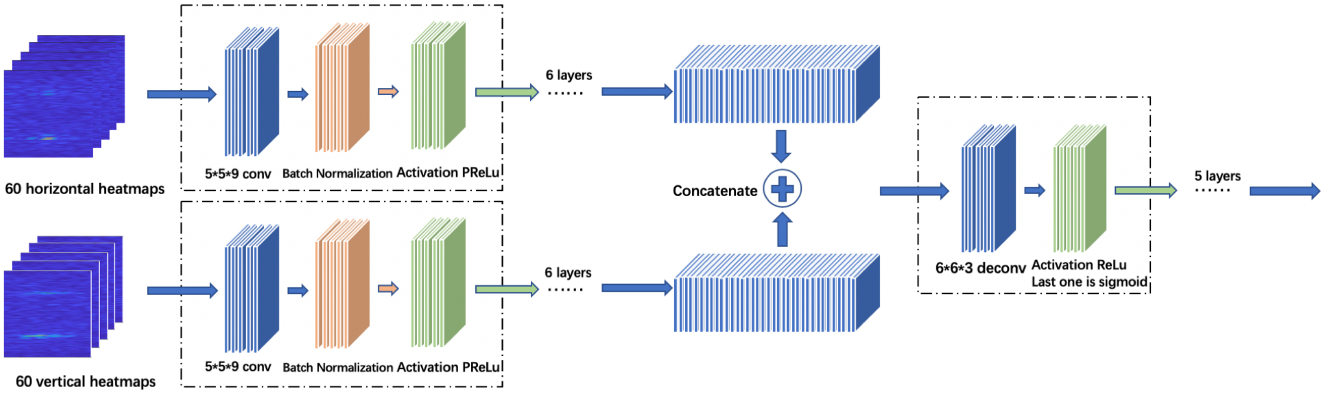

Fig. 8. CNN network

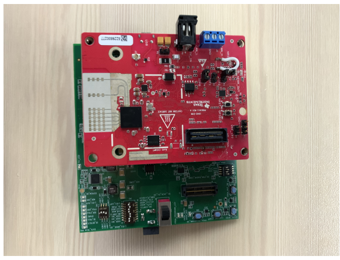

Fig. 9. mmWave Radar Device (AWR1642 (Red) and DCA1000 (Green))

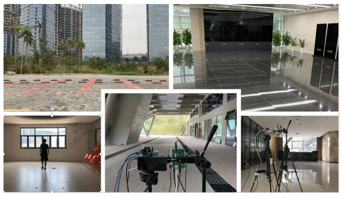

Fig. 10. Experimental Scenario

a relatively fixed position. For the video recording device, we use a XIAOMI 8 to record the video. We collected coordinated radar and vision data in different environments around the campus, which includes hall, gym, and so on, see Figure 10. The dataset contains 2 hours data from 5 different environments. Six single people are walking when collecting. The whole data is split into training and testing sets. We make sure the testing set wouldn't be used while training.

B. Experimental Results

The output of the system can be seen in Figure 11. For evaluation, like many tasks in object detection, we report object keypoints similarity (OKS) and the mean average precision (AP) over different OKS thresholds. The OKS is defined as:

$$OKS_p = \frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)}, \qquad (7)$$

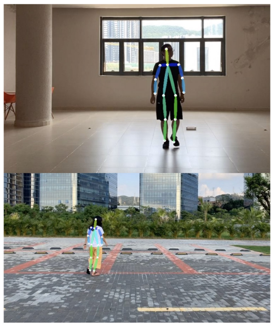

Fig. 11. Experiment results

where p is id for a person in ground truth, i for id of keypoint, d_{pi} represents for the Euclidean distance of keypoints between prediction and ground truth. S_p is the root of area of ground truth and σ_i is the normalization factor of the i keypoint. $\delta(v_{pi} = 1)$ determines whether to use this keypoint to calculate.

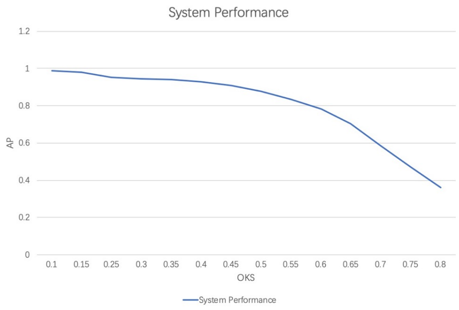

Fig. 12. System Performance

Overall Evaluation: To evaluate the performance of the system and compare our system with the baseline, we manually label around 3000 pictures as the testing set. The result is can be seen in Table II and Figure 12. Particularly, we give

TABLE II
EVALUATION

System	Average OKS	mAP	AP50	AP75
Openpose	93.3	92.0	98.8	93.3
Our System	70.5	50.0	87.7	47.1
Leave One Out	66.6	42.4	83.8	34.8

TABLE III
DIFFERENT PARTS EVALUATION

	Head	Neck	Sho	Elb	Wrist	Hip	Knee	Ank
OKS	42.5	73.2	61.1	55.5	47.1	69.1	70.9	65.5
AP50	47.8	83.9	67.4	60.6	51.7	79.5	80.0	73.5
mAP	29.6	59.3	48.1	41.6	32.3	58.4	58.7	50.2

the AP and OKS result of leaving one person out in training and evaluating system performance on his data only.

Difference Between Body Parts: There is a big difference within OKS values of different body parts, see table III. Mainly, it is because that the radar can not always receive the reflection of every part of human body. The lack of reflection information leads to a worse result of the parts like wrists. Result of parts like hips and shoulders is better due to a larger reflection area.

VII. DISCUSSION

The capture of our system depends on radar signal, which leads to some limitations: In our implementation, we only consider one person situation this time. Secondly, the performance of our system highly depends on the training set. Thirdly, the detecting range of the mmWave radar is limited to its power. This time we focus on a range of 4 to 13 meters. Furthermore, we find the accuracy of the system highly relys on the number of antennas.

VIII. CONCLUSION

In this paper, we proposed a human pose capturing system using mmWave radar, which is a novel application of 77GHz radar device. The devices we use are much smaller than ever and it is off-the-shelf. We use a new coordination method to align two-radars devices and label-used camera, which is the key to use two radars to realize the two-radar system. Then, we implement the system and create our own dataset to complete evaluation. With our neural networks well trained, we can transform radar producing heatmaps into the human skeleton. The system can achieve human pose capture while protecting the privacy and it is still valid in bad light condition. It can be potentially used in healthcare, smart homes or other applications.

REFERENCES

[1] D. L. Ke Sun, Bin Xiao and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[2] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334–2343, 2017.

[3] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7356–7365, 2018.

[4] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[5] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *European Conference on Computer Vision*, pp. 627–642, Springer, 2016.

[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

[7] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, pp. 34–50, Springer, 2016.

[8] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937, 2016.

[9] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system," in *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, pp. 71–84, 2013.

[10] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via {RF} body reflections," in *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pp. 279–292, 2015.

[11] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pp. 267–281, ACM, 2018.

[12] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 137, 2018.

[13] T. Instruments. http://www.ti.com/.

[14] G. Malysa, D. Wang, L. Netsch, and M. Ali, "Hidden markov model-based gesture recognition with fmcw radar," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1017–1021, IEEE, 2016.

[15] F. Jin, R. Zhang, A. Sengupta, S. Cao, S. Hariri, N. K. Agarwal, and S. K. Agarwal, "Multiple patients behavior detection in real-time using mmwave radar and deep cnns,"

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[17] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.