

# Poster: Efficient Convolutional Neural Network for FMCW Radar Based Hand Gesture Recognition

**Xiaodong Cai**  
xiaodong.cai@intel.com  
Intel Corporation  
Shanghai, China

**Jingyi Ma**  
jingyi.ma@intel.com  
Intel Labs China  
Beijing, China

**Wei Liu**  
wei.liu@intel.com  
Intel Labs China  
Beijing, China

**Hemin Han**  
hemin.hani@intel.com  
Intel Corporation  
Shanghai, China

**Lili Ma**  
michael.ma@intel.com  
Intel Corporation  
Shanghai, China

## ABSTRACT

FMCW radar could detect object's range, speed and Angle-of-Arrival, advantages are robust to bad weather, good range resolution, and good speed resolution. In this paper, we consider the FMCW radar as a novel interacting interface on laptop. We merge sequences of object's range, speed, azimuth information into single input, then feed to a convolution neural network to learn spatial and temporal patterns. Our model achieved 96% accuracy on test set and real-time test.

## CCS CONCEPTS

• Human-centered computing; • Computing methodologies;

## KEYWORDS

FMCW radar, hand gesture recognition, signal processing, Convolutional Neural Network

## ACM Reference Format:

Xiaodong Cai, Jingyi Ma, Wei Liu, Hemin Han, and Lili Ma. 2019. Poster: Efficient Convolutional Neural Network for FMCW Radar Based Hand Gesture Recognition. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*, September 9–13, 2019, London, United Kingdom. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3341162.3343768>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*UbiComp/ISWC '19 Adjunct*, September 9–13, 2019, London, United Kingdom

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6869-8/19/09.

<https://doi.org/10.1145/3341162.3343768>





Sensor	Field of View	Range resolution	Information captured	Anonymity	Bad weather	Price	Computation
 radar	★★★★★ 90°	★★★★★ 2mm	★★★★★ Distance, velocity, angle	★★★★★ Yes	★★★★★ Good in fog, dust, smoke	★★★★★ \$5	★★★★★ Medium
 IR TOF Camera	★★★★★ 60°	★★★★★ 2cm	★★★★★ Texture, RGB, IR	★★★★★ No	★★★★★ Sensitive to light	★★★★★ \$5	★★★★★ High
 PIR	★★★★★ 100°	★★★★★ N/A	★★★★★ Human in motion	★★★★★ Yes	★★★★★ Sensitive to light, heat	★★★★★ \$1	★★★★★ Low
 ultrasonic	★★★★★ 100°	★★★★★ 3mm	★★★★★ distance	★★★★★ Yes	★★★★★ Bad in smoke, dust	★★★★★ \$1	★★★★★ Low

Figure 1: Sensors comparison

## 1 INTRODUCTION

For camera based gesture recognition, there are many commercial solutions, such as Kinect, leap motion, RealSense. These solutions suffer from privacy issues, while FMCW radar has no such limits. As FMCW radar can only estimate object's range, speed and angle information, so it capture human actions, but the information is not enough to identify the user. Detailed comparison among different sensors are shown in Figure 1. It is also friendly for industrial design that radar doesn't need hole-punch, while microphone and camera does.

## FMCW Radar Basics

FMCW radar is short for Frequency Modulated Continuous Wave Radar. Radar transmits a continuous carrier modulated by a periodic function such as a sinusoid wave to provide range data. At each period  $T_c$  (also called a chirp), radar transmitter emits a sinusoidal wave, with frequency modulated from  $f_{min}$  to  $f_{max}$ . Bandwidth  $B = f_{max} - f_{min}$ , which is proportional to radar spatial resolution. We set modulation slope  $\alpha = B/T_c$ . In this paper, we use a 2Tx, 4Rx FMCW Radar, sweeping frequency 57-64GHz,

Consider an object, e.g. palm of hand, with an initial range  $R_0$  at  $t_0 = 0$ , and radial velocity  $v$ , where  $v \ll c$ , so the signal travel time  $\tau = \frac{2}{c}(R_0 + v \cdot t)$ . FMCW radar transmitter

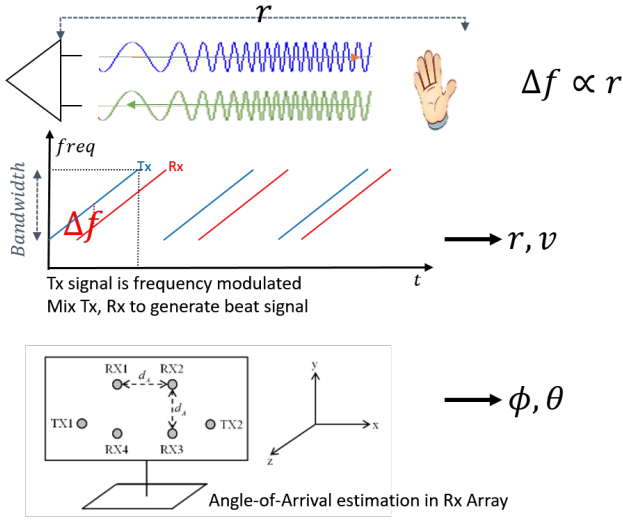


Figure 2: Radar basics

emits sinusoidal signal,  $s(t) = A \cdot \cos(\phi_s \cdot t)$ , the received signal  $\phi_r(t) = s(t - \tau)$ , mixed signal  $\phi_w(t) = \int (\phi_s(t) \phi_r^*(t))$ , where the beat frequency in  $\phi_w(t)$  is the range of the object; and the speed of the object can be estimated by frequency shift  $f_d$ ,  $v = \frac{f_d \lambda}{2}$ . Angle-of-Arrival is calculated via phase difference between 2 receivers. Same range bin and speed bin in Range-Doppler Map (RDM), Receiver 1  $r_1^{(i,j)}$  and Receiver 2  $r_2^{(i,j)}$ , Angle-Of-Arrival  $\theta = \sin(\frac{\Delta \lambda}{d})$ , where  $\Delta \lambda = \frac{2\pi}{\lambda} \Delta \beta$ ,  $\Delta \beta = \text{atan}(\frac{r_1^{(i,j)}.imag}{r_1^{(i,j)}.real}) - \text{atan}(\frac{r_2^{(i,j)}.imag}{r_2^{(i,j)}.real})$ , illustrated in Figure 3. How FMCW radar estimate hand movement is illustrated in Figure 2, and readers can refer to [4] for technical details.

All in all, for a Multiple Input Multiple Output (MIMO) radar system,  $Q$  objects' range, speed, angle can be estimated via 3D FFT, where  $l$  is the antenna index,  $n$  is sampling index,  $p$  is the chirp index,  $Q$  is the number of object,  $\alpha^{(q)}$  is the amplitude factor,  $K$  is modulation slope,  $R^{(q)}$  is range of  $q$  th object,  $f_d^{(q)}$  is the frequency shift of  $q$  th object,  $f_c$  is radar's center frequency,  $d$  is the distance between antenna,  $\theta^{(q)}$  is the azimuth of  $q$  th object,  $c$  is the speed of light.

$$d(l, n, p) \approx \sum_{q=0}^{Q-1} \alpha^{(q)} \cdot e^{j2\pi [\frac{(2K R^{(q)} + f_d^{(q)})n}{f_s} + \frac{f_c l d \sin \theta^{(q)}}{c} + f_d q p T_0 + \frac{2f_c R^{(q)}}{c}]}$$

### Gesture Definitions

we defined four gestures: left wave, right wave, click, and wrist in Table 1. We also plot theoretical analysis of range, speed, azimuth trajectories on the predefined gestures set, shown in Figure 4. From Figure 4, those gesture trajectories

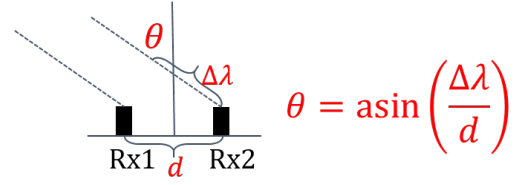


Figure 3: Angle of Arrival estimation

Table 1: Gesture Definitions

Gesture	Hand Movement	Meaning of gesture
LEFT	Move from right to left	Browse previous item
RIGHT	Move from left to right	Browse next item
CLICK	Finger pointing to radar	Select an item
WRIST	Hand making fist	Return to main menu

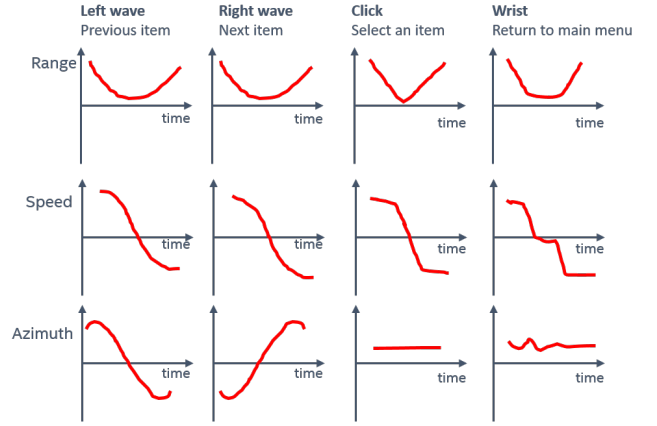


Figure 4: Theoretical analysis of hand movement

are quite distinctive. We build a template-matching algorithm for gesture recognition, the accuracy is 70-80%, and it is hard to extract trajectories from noisy radar signal.

## 2 RELATED WORK

Previous work [2], [5], [6] are based on CNN+LSTM. [5] first introduced CNN+LSTM architecture to process radar based gesture recognition. CNN learns spatial patterns inside RDM, then feed into LSTM to learn temporal patterns among RDMs. CNN+LSTM achieved 87% accuracy on 11 gestures. [2] replace CNN with AllConvNet to reduce parameters and inference time. [6] introduced 3D CNN to learn spatial and temporal patterns at the same time.

Solutions above [2], [5], [6] use real value Range-Doppler map as input, they ignore Angle-of-Arrival information. In order to take Angle-Of-Arrival into consideration, we need extra signal processing procedure.

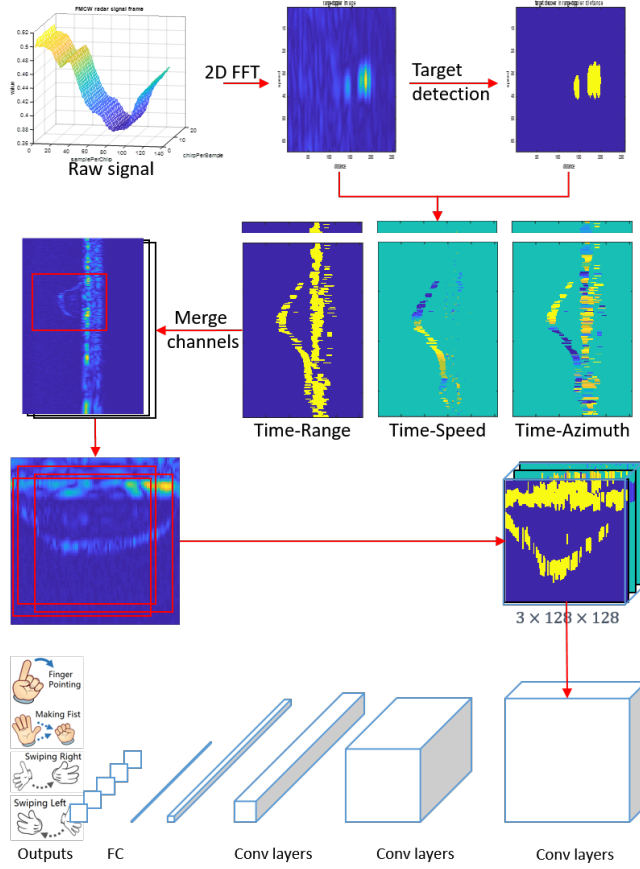


Figure 5: Radar data processing pipeline

### 3 PROPOSED SYSTEM

To explicitly extract Range, Speed and Azimuth trajectory, we merge 128 RDM frames, each frame size is  $64 \times 256$ , into a 3-channel input frame, representing range-time, speed-time, azimuth-time respectively, RSA for short. The merged RSA input shape is  $128 \times 128 \times 3$ , then feed it into CNN. Gesture recognition pipeline is shown in Table 2 and Figure 5.

#### Neural Network Architecture Design

Firstly, we design a VGG-like neural network, called VGG-10 in Figure 6(a), follows {Conv3x3, Conv3x3, MaxPooling} building block, and 2 fully Connected layers. ADAM optimizer, early stopping and reduced learning rate is applied. VGG-10 converged at 10th epoch with validation accuracy 92%.

To improve the performance, we add residual block between convolution layers, batch normalization is also added between each residual block to make back-propagation more robust [3], and we called it ResNet-20 in Figure 6(b). ResNet-20 outperforms VGG-10 achieved 98% validation accuracy.

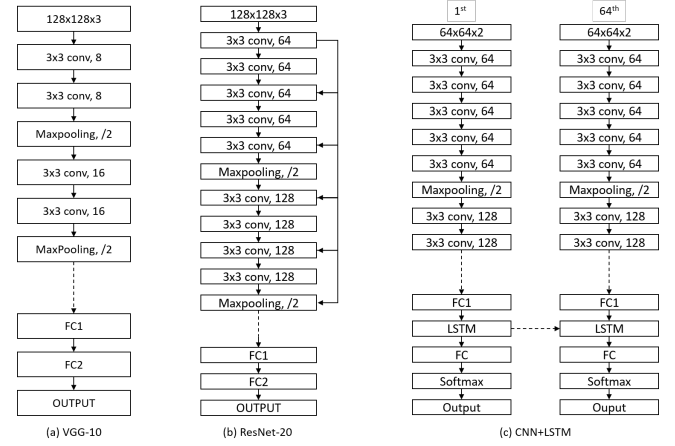


Figure 6: Network architecture, (a) VGG-10, (b) ResNet-20, (c) CNN+LSTM

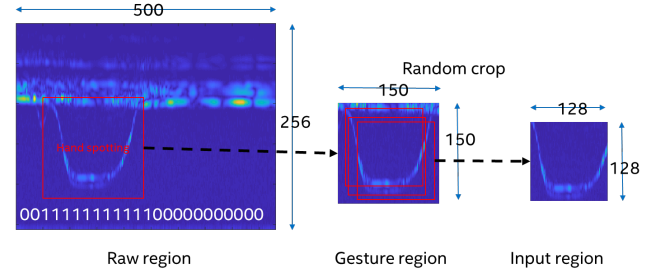


Figure 7: Data augmentation

We also build CNN+LSTM model for comparison, in Figure 6(c). CNN+LSTM needs RDM sequence as input, first we resize original RDM ( $64 \times 256$ ) to  $64 \times 64$ , and feed 64-resized RDM frames into CNN. CNN module follows {Conv5x5, Conv5x5, MaxPooling} building block, and 1 fully Connected layer to encode feature, then put feature encoding into LSTM.

### 4 DATASET AND EXPERIMENT RESULTS

We collected 50 subjects' gesture data, each subject did 4 gestures 10 times with left hand and right hand, total 3652 valid records, validation-train split ratio is 0.3.

We also do data-augmentation to enrich the dataset. First, we draw a block containing gesture movements, then crop the area randomly to generate training data, shown in Figure 7. At last, we obtained more than 400k training data.

We compare VGG-10, ResNet-20 and CNN+LSTM on a same dataset, and calculate average accuracy. CNN+LSTM has the lowest accuracy on LEFT/RIGHT, due to lack of Angle-Of-Arrival information; Deep CNN outperforms shallow CNN, which is aligned with experiment results in [5].

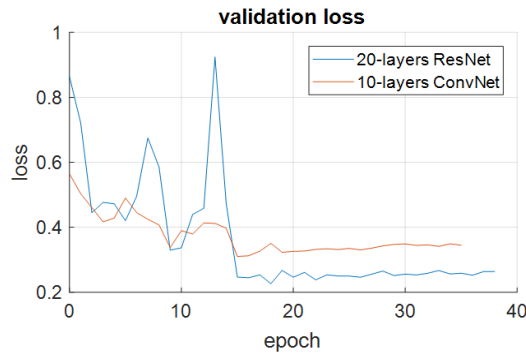
In Figure 9, our model achieves 98% average accuracy on test set.

**Table 2: Radar signal pre-Processing**

Step	input shape	procedure	output shape
1	$16 \times 128$	2D FFT, to convert raw signal into RDM	$64 \times 256$
2	$64 \times 256$	Do Constant False Alarm Rate (CFAR) [1] on RDM to detect hand and body	$64 \times 256$
3	$64 \times 128$	Crop RDM to keep body and hand, generate subset of RDM	$64 \times 128$
4	$64 \times 128$	For each range bin, calculate maximum speed, average azimuth, generate a frame	$1 \times 128 \times 3$
5	$1 \times 128 \times 3$	merge 128 frames of above output into one frame	$128 \times 128 \times 3$

**Table 3: Accuracy comparison among models**

Network Architecture	Avg. Acc.	LEFT	RIGHT	CLICK	WRIST
VGG-10	91.0%	94.9%	80.7%	95.5%	97.0%
ResNet-20	98.7%	99.1%	99.0%	97.9%	98.9%
CNN+LSTM	78.0%	69.0%	49.5%	84.6%	90.1%

**Figure 8: Validation loss of VGG-10 and ResNet-20**

Confusion matrix on Test set		gesture prediction					
		none	left	right	click	wrist	prec
Gesture ground truth	none	2189	18	5	6	5	98.5%
	left	9	2118	10	0	0	99.1%
	right	1	19	2055	0	0	99.0%
	click	1	5	0	2150	40	97.9%
	wrist	10	0	0	14	2215	98.9%
	recall	99.0%	98.1%	99.3%	99.1%	98.0%	98.7%

**Figure 9: Confusion matrix on test set**

## Error Analysis

- LEFT is misclassified as RIGHT, and RIGHT is misclassified as LEFT. The gesture movement consists of three temporally overlapping phases: preparation, nucleus and retraction. The retraction phase of LEFT is a RIGHT, and the retraction phase of RIGHT is a LEFT. A better gesture segmentation preprocessing may help to reduce this kind of error.
- CLICK is misclassified as LEFT and WRIST is misclassified as RIGHT. Recall Figure 4, the biggest difference between LEFT/RIGHT and CLICK/WRIST is the angle difference, when CLICK/WRIST gesture has large angle changes, they are easily misclassified as LEFT/RIGHT. More accurate angle estimation may help reduce this error.

## 5 CONCLUSION

FMCW radar is a low cost/high spatial/speed resolution sensors, it can detect anonymous object movements, suitable for privacy-concerned interaction application. We designed a 20-layer residual network to recognize gestures, and the model could achieve 96% accuracy on real-time test. In the future, we plan to support user defined gestures.

## REFERENCES

- [1] Stephen Blake. 1988. OS-CFAR theory for multiple targets and nonuniform clutter. *IEEE transactions on aerospace and electronic systems* 24, 6 (1988), 785–790.
- [2] Souvik Hazra and Avik Santra. 2018. Robust gesture recognition using millimetric-wave radar system. *IEEE sensors letters* 2, 4 (2018), 1–4.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [4] Sujeet Milind Patole, Murat Torlak, Dan Wang, and Murtaza Ali. 2017. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine* 34, 2 (2017), 22–35.
- [5] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. 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*. ACM, 851–860.
- [6] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. 2018. Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sensors Journal* 18, 8 (2018), 3278–3289.