Real-Time Multi-Gesture Recognition using 77 GHz FMCW MIMO Single Chip Radar

Piyali Goswami[*], Sandeep Rao[*], Sachin Bharadwaj[*], Amanda Nguyen[§]
Texas Instruments (India) Pvt. Ltd. [*], Texas Instruments[§]
piyali_g@ti.com, s-rao@ti.com, s-bhardwaj@ti.com, a-nguyen@ti.com

Abstract*—Innovations in CMOS radar has paved way for new functions like gesture-based human-machine interaction using radar for consumer and automotive electronics. Single chip radars which integrate the RF front-end and digital processing logic are fit for such applications due to their cost and form factor but are constrained in angular resolution, memory, and processing power. In this paper, we propose low complexity radar-based multi-gesture classification solution which overcomes these constraints to achieve 96% accuracy for 6 gestures generalized across 8 users. The algorithm developed was found to consume only 8.4% DSP cycles and 256KiB memory on Texas Instrument's AWR1642.

***Keywords*—Gesture Recognition, 77 GHz CMOS Radar, Single Chip Radar**

I. Introduction

Natural user interfaces have been a field of active research over the past years in academia and industry. Multiple sensing technologies have been used to seamlessly interact with computing devices. Camera-based systems have made great advancements in the last few years in the field of natural user interfaces [1][2]. However, these techniques are challenged by higher processing latency, dark light conditions, shadows, and occlusions. Gesture sensing using 3D Time of flight (ToF) sensors have been shown to achieve higher accuracy and faster response times versus traditional stereo-camera based systems [3][4]. However, 3D ToF systems are challenged by environmental conditions (like fog, dust and bright light) and do not scale to very small form factors required by wearable computing devices. Radar sensors have been used in the past in conjunction with other vision and ToF sensors for gesture detection and recognition [5][6]. The role of the radar in these systems was limited to detecting the velocity and depth of the hand. RF sensing for gesture recognition, particularly millimeter wave (mmWave) Radar, has gathered significant interest amongst researchers through the works highlighted in [7] and [8].

Automotive radar systems are starting to extensively use 77 GHz millimeter wave radar for long-range radar (LRR) and short-range radar (SRR) vehicle safety applications [9]. In this paper, we use 77 GHz mmWave Radar for gesture sensing applications and discuss an innovative algorithm and software pipeline which enables detection of dynamic gestures in real time on single chip low power CMOS Radar devices with very limited memory and compute resources. The choice of 77 GHz was inspired by the re-purposing of the same modules used to achieve SRR and LRR functionality on the vehicle when stationary to allow hands free interaction in applications such as swipe to open the door, kick to open trunk etc. This work can easily be extended to 60/24 GHz Radar.

Frequency modulated continuous wave (FMCW) radar can measure the range (i.e. radial distance from the radar), velocity (relative velocity w.r.t radar) and direction of arrival (DOA) of objects in front of the radar. Range resolution, velocity resolution, and angle resolution are key parameters affecting the performance of the radar. The range resolution of radar depends on the RF bandwidth of the transmitted signal (chirp). Thus velocity resolution is largely independent of hardware constraints and can be improved by extending the frame time, which in practice implies adding more chirps per frame. The angle resolution of radar depends on the number of antennas. Since each receive (Rx) antenna requires its own Rx signal chain, increasing angle resolution entails a significant increase in area and cost of the radar solution. Consequently, limited angle resolution is one of the biggest limitations of single-chip radar solutions. Therefore, gesture recognition algorithm for radar should leverage the velocity and range resolution while minimizing the dependence on angle resolution.

II. Gesture Recognition with Single Chip Radar

A. Gesture Recognition Pipeline

Finding the position of an obstacle requires estimating both the range and angle of the object (w.r.t the radar). Thus any gesture recognition approach that works for detecting/tracking the location of the hand (or fingers within a hand) will require a radar sensor with both high range resolution and angular resolution. With its limited angle resolution, such an approach will not work on single-chip radar. Instead, we recommend the approach highlighted in Fig. 1. As the first step, a 2D-FFT is performed to resolve the scene in range and Doppler. The 2D-FFT output across multiple antennas is non-coherently added to create a range-doppler heat map. Several specific features are then extracted from the range-doppler heat map.

Feature extraction is a common technique for any classification problem. The innovation in our methodology is that the feature extracted is a single number from the current heat map whose value reflects the weighted average of a radar specific parameter. This significantly reduces the compute and memory complexity in the further classification stages. Examples of such features are average doppler, average range, Doppler spread etc. It is important to note that a single frame of the radar yields a single value for each feature. A sequence of frames thus yields a time series for each feature. Features extracted over a sliding time window are sent to a classification algorithm (such as an Artificial Neural Network (ANN)) to determine the gesture. In the following sections, we deep dive into the features which were extracted and the neural network architecture used to classify 6 gestures as shown in Fig. 2.

B. Feature Extraction

The output of the 2D FFT is a time-varying set of complex numbers in the range and doppler dimension. A lot of information can be derived from this range-doppler image based on the features described in this section.

Magnitude Based Features: These features are derived after calculating the magnitude of the complex 2D FFT of each antenna followed by non-coherent addition to create a range doppler image (RDI). Here, ‘i’ corresponds to the index of the RDI data, Z_i is the magnitude value, R_i is the range value, and D_i is the Doppler value corresponding to the i^{th} index.

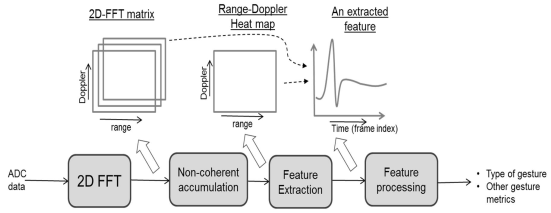


Fig. 1 Gesture Recognition Processing Pipeline

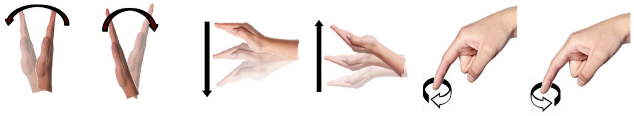


Fig. 2 Six gestures used for the gesture evaluation.

Weighted Range: This feature can be used to detect the position of the hand (range centroid) and can be calculated using (1). $U = \frac{\sum_i Zi.Ri}{\sum_i Zi}$ (1)

Weighted Doppler: This feature can be used to detect the velocity centroid of the hand and can be calculated using (2).

$$V = \frac{\sum_i Zi.Di}{\sum_i Zi} \qquad (2)$$

Instantaneous Energy: This feature can be used to detect the presence of the hand and can be calculated using (3).

$$I = \sum_i Zi \qquad (3)$$

Range Dispersion: This feature measures the variation in the weighted range value. Equation (4) is used to calculate this dispersion.

$$RD = \sqrt{\left(\frac{\sum_i Zi.(Ri-U)^2}{\sum_i Zi}\right)} \qquad (4)$$

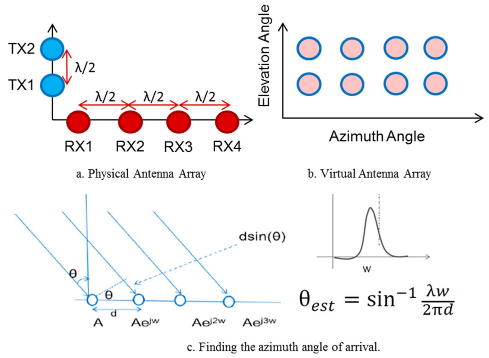
Doppler Dispersion: This captures the variation in weighted doppler and can be calculated using (5).

$$DD = \sqrt{\left(\frac{\sum_i Zi.(Di-V)^2}{\sum_i Zi}\right)} \qquad (5)$$

Phase-Based Features: These features are derived from the phase values corresponding to range and doppler coordinates of the strongest object in 2D FFT across all antennas after ignoring the zero Doppler lines (i.e. static clutter in the environment is ignored). Consider 2 transmit antennas and 4 receive antennas arranged as shown in Fig. 3a. With both transmit antennas operating in a time/code division multiplexed fashion; we get an equivalent 2-dimensional synthesized antenna array as shown in Fig. 3b. The basic principle of one-dimensional angle estimation for an antenna array is shown in Fig. 3c. We first take a snapshot of the signal across all antennas and take a 3rd dimension FFT of the captured signals and estimate the index ‘*w*’ corresponding to the peak. The angle of arrival can then be estimated using ‘*w*’. We extend this principle to two dimensions to find the azimuth angle (θ) and elevation angle (φ) and add this to our feature vector.

Statistical Features: In order to capture the relations between the different features, statistical features can be used to differentiate between different gestures.

Doppler-Azimuth Correlation: This feature captures the variation of Doppler with the azimuthal angle in time. We create a sliding time window of the last *n* frames at time *t* to capture the weighted doppler and azimuth angle vectors.



a. Physical Antenna Array

b. Virtual Antenna Array

c. Finding the azimuth angle of arrival.

Fig. 3 Angle of Arrival Estimation to extract phase based features.

$$\vec{\boldsymbol{V}} = \{V_{t-n}, V_{t-n+1}, \dots \dots, V_{t-1}\}$$

$$\vec{\boldsymbol{\theta}} = \{\theta_{t-n}, \theta_{t-n+1}, \dots \dots, \theta_{t-1}\}$$

The correlation is derived using (6) where μ is the mean, σ is the standard deviation, and E is the expectation operation on these vectors.

$$\rho_{\mathbf{V\theta}} = \frac{E[(\mathbf{V}-\mu_{\mathbf{V}})(\boldsymbol{\theta}-\mu_{\boldsymbol{\theta}})]}{\sigma_{\mathbf{V}}.\sigma_{\boldsymbol{\theta}}} \qquad (6)$$

Number of detected points: This captures the number of range, doppler coordinate pairs in the RDI which have a magnitude higher than a multiple of the noise floor. The multiplication factor is heuristically derived based on measurements.

C. *Gesture Classification*

We stack all the extracted features into a feature vector. The size of the feature vector is a function of a sliding time window across which the gestures are to be classified.

We use an artificial neural network to classify the gestures. The choice of the feature set, sliding time window and ANN architecture, hyperparameter selection is described in detail in Section III.B.

III. SYSTEM CONFIGURATION

A. *FMCW Configuration*

The FMCW chirp configuration used in our evaluation is shown in TABLE I. The configuration has been chosen to leverage the superior velocity and range resolution achievable with 77GHz AWR1642 radar. The large inter-frame time (400us) between adjacent chirps and a large number of chirps (128) within a frame serve to increase the dwell time and thus improve the velocity resolution. The total dwell time was chosen to be nearly 50 ms. The chirp bandwidth is configured to 4GHz, which results in the best range resolution achievable.

TABLE I. FMCW CHIRP PARAMETERS

Parameter	**Value**
Chirp BW	~4GHz
Chirp repetition interval	400us
Number of chirps per frame	128
Range Resolution	~4cm
Velocity resolution	0.037 m/s

B. Gesture Classifier: ANN Configuration

Multiple hyper-parameters need to be identified for the ANN classifier to balance complexity (memory and processing requirement) and classification accuracy. We discuss the steps to identify the values of these hyperparameters.

Input Feature Vector Size: For the gestures under consideration, we found that the gestures complete with 10 dwells or 500 ms. Based on this observation, we chose the sliding time interval across which gestures are analyzed to be 10 dwells. We additionally optimized the size of the feature set to 6 features, wherein we chose only those features which showed differentiating characteristics. The optimized feature set consisted of Weighted range, Weighted doppler, Instantaneous Energy, Azimuth Angle, Elevation Angle and Azimuth-Doppler Correlation.

Supervised Training: We use supervised training method to train the parameters (weights and biases) of the ANN classifier. The labeled dataset had 1.68 million data points (feature vectors) collected across 8 users.

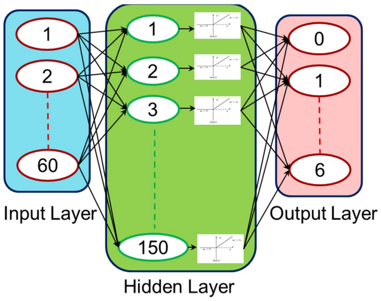


Fig. 6 ANN Classifier architecture

The dataset was split into 3 parts: Training (65 %): used to train the network for given hyper-parameter set. Validation (15 %): used to choose the best hyper-parameters. Test (20 %): which served as a proxy for real life and used to find the generalized performance of the ANN classifier. We evaluated

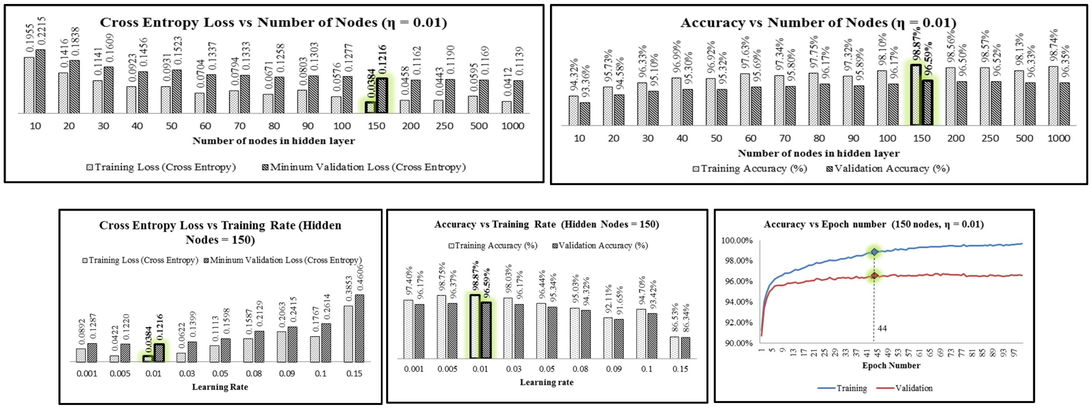


Fig. 4 ANN Classifier Hyper-Parameter evaluation

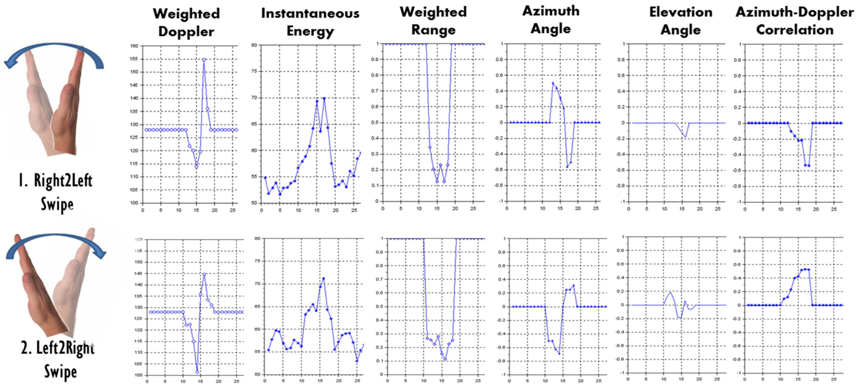


Fig. 5 Feature Variation over time for Right to Left and Left to Right Hand Swipes

TABLE IV. ANN CLASSIFIER TEST DATA CONFUSION MATRIX

Predicted Actual	BACK GROUND	RIGHT TO LEFT	LEFT TO RIGHT	UP TO DOWN	DOWN TO UP	FINGER CLOCKWISE	FINGER ANTICLOCKWISE
BACK GROUND	**94.94%**	1.12%	0.45%	0.67%	0.11%	0.90%	1.80%
RIGHT TO LEFT	1.60%	**96.92%**	0.00%	0.25%	0.62%	0.12%	0.49%
LEFT TO RIGHT	1.27%	0.00%	**96.96%**	0.25%	0.51%	0.25%	0.76%
UP TO DOWN	0.41%	0.00%	0.14%	**98.64%**	0.27%	0.27%	0.27%
DOWN TO UP	1.61%	0.12%	0.00%	0.50%	**97.40%**	0.12%	0.25%
FINGER CLOCKWISE	1.76%	0.00%	0.12%	0.12%	0.12%	**96.36%**	1.53%
FINGER ANTICLOCKWISE	1.44%	0.48%	0.00%	0.24%	0.48%	1.56%	**95.79%**

the performance of the learning using cross-entropy loss as defined in (7) and optimized it using the stochastic gradient descent algorithm as defined in (8).

$$\text{Cross-entropy loss}= \frac{1}{N}\sum_i \text{loss}(i) \qquad (7)$$

where $loss(i)$ is the loss for i^{th} data-point, N is the number of data-points in the batch, and

$$loss(i) = -label(i)\log[p(i)]$$

Here $p(i)$ is ANN classifier output, $label(i)$ is the true label for i^{th} data-point.

$$\mathrm{w}_j^t = \mathrm{w}_j^{t-1} - \alpha\frac{\partial \mathrm{loss}}{\partial \mathrm{w}_j^{t-1}} \qquad (8)$$

where w_j^t denotes j^{th} weight at iteration t and α is the learning rate.

Network Architecture: The network architecture used is shown in Fig. 6. We used single hidden layer and rectified linear unit (Relu) activation function in order to minimize the compute requirements when implementing the classifier on the C67x DSP. The trade-offs involved in selecting multiple hyper-parameters (i.e, number of nodes in the hidden layer, learning rate and number of epochs used for training) is shown in Fig. 4.

Using 150 nodes in the hidden layer and a learning rate of 0.01 gave the minimum cross-entropy loss and the maximum validation set accuracy.

IV. Results

We implemented the gesture recognition pipeline on Texas Instrument's AWR1642 single chip CMOS Radar [10]. The AWR1642 device is a highly integrated 76–81-GHz radar-on-chip solution for SRR applications. The device comprises of the entire millimeter wave (mmWave) radio-frequency (RF) and analog baseband signal chain for two transmitters (TX) and four receivers (RX), as well as two customer-programmable processor cores in the form of a C674x digital signal processor (DSP) @ 600 MHz and an ARM® Cortex®-R4F microcontroller (MCU). The AWR1642 device uses complex baseband architecture and provides in-phase (I-channel) and quadrature (Q-channel) outputs. TI's AWR1642 single-chip radar has a best in class synthesizer which can span a bandwidth of 4GHz, which translates to a range resolution of <4cm. The AWR1642 has a total available memory of 768 KiB.

TABLE II. Measured Processing Latency

Processing Stage	Latency	DSP Utilization
2D – FFT	7.03 µs per chirp	1.79%
Feature Extraction	2.88 µs per chirp per feature	4.42%
Feature Classification	1.1 ms	2.2 %
Total DSP Utilization		**8.41 %**

TABLE III. Gesture Processing Pipeline Memory Consumption

Processing Stage	Memory
2D – FFT Output Buffer	192 KiB
Code Section	6.5 KiB
Read-Only Data	17 KiB
Global Data	40 KiB
Total	**255.5 KiB**

The features extracted were analyzed over time for different gestures. A snapshot of the features for two gestures (Left to Right swipe and Right to Left swipe) is shown in Fig. 5. These features were then subsequently normalized to pass it through the classification stage. The processing pipeline latencies are highlighted in TABLE II. The memory consumption of the algorithm pipeline is highlighted in TABLE III. The classification accuracy was evaluated on the test dataset. The confusion matrix highlighting the accuracy per class of gesture is highlighted in TABLE IV. Here, "Back Ground" includes other hand movements, actions of people walking in front of the radar and no movement.

The overall solution is shown to have higher levels of integration where the RF front end, the ADC converter, and the digital data processing is handled in a single chip in comparison with the solution in [7] and [8].

Conclusion

We presented a novel gesture recognition algorithm and discussed the configuration and implementation details which overcome the constraints of single-chip 77GHz CMOS Radar. The solution drives highly integrated small form-factor systems which can be used for pervasive human-machine interaction using radar. The utilization of 77 GHz enables re-purposing existing LRR and SRR systems mounted on the body of the vehicle for gesture based vehicle control. It opens the way for future systems to detect more complex human gestures with a high accuracy and also maintain a low processing and memory footprint.

References

[1] David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. "Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor." In *Proceedings of the 25th annual ACM symposium on User interface software and technology* (UIST '12). ACM, New York, NY, USA, 167-176.

[2] T. Sharp, C. Keskin, D. P. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. W. Fitzgibbon, and S. Izadi, "Accurate, Robust, and Flexible Real-time Hand Tracking.," in CHI, 2015, pp. 3633–3642.

[3] T. Kopinski, S. Geisler and U. Handmann, "Gesture-based human-machine interaction for assistance systems," *2015 IEEE International Conference on Information and Automation*, Lijiang, 2015, pp. 510-517.

[4] A. V. Laack, J. Blessing, G.D. Tuzar, O. Kirsch, "Time of Flight Technology for Gesture Interaction", Visteon White Paper 2016. Online: https://www.visteon.com/products/documents/Time_of_Flight_Technology_for_Gesture_Interaction.pdf.

[5] P. Molchanov, S. Gupta, K. Kim and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, 2015, pp. 1-8.

[6] P. Molchanov, S. Gupta, K. Kim and K. Pulli, "Short-range FMCW monopulse radar for hand-gesture sensing," *2015 IEEE Radar Conference (RadarCon)*, Arlington, VA, 2015, pp. 1491-1496.

[7] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. "Soli: ubiquitous gesture sensing with millimeter wave radar". *ACM Trans. Graph.* 35, 4, Article 142 (July 2016).

[8] 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* (UIST '16). ACM, New York, NY, USA, 851-860.

[9] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel and C. Waldschmidt, "Millimeter-Wave Technology for Automotive Radar Sensors in the 77 GHz Frequency Band," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 3, pp. 845-860, March 2012.

[10] J. Singh, B. Ginsburg, S. Rao and K. Ramasubramanian, "AWR1642 mmWave sensor: 76–81-GHz radar-on-chip for short-range radar applications," Texas Instruments, May 2017. Online: http://www.ti.com/lit/wp/spyy006/spyy006.pdf