

Range-Doppler Hand Gesture Recognition using Deep Residual-3DCNN with Transformer Network

Gaurav Jaswal¹, Seshan Srirangarajan^{1,2}, and Sumantra Dutta Roy¹ ¹Department of Electrical Engineering

²Bharti School of Telecommunication Technology and Management Indian Institute of Technology Delhi, New Delhi 110016, India E-mail: {gauravjaswal; seshan; sumantra}ee.iitd.ac.in

1. MOTIVATION

- How can we use a small set of fine finger gestures or hand movements to control everything around us?
- Existing keyboard or touch based interaction paradigm is slow.
- Natural, low effort and high precision interaction paradigm is urgently needed. For example, hand gestures.
- Camera or vision sensors for hand gestures acquisition cause sensitivity to lighting conditions, occlusion, and typically require dedicated processing power.
- Radar sensors emerged as new HCI technology that offer micro gesture interaction with low energy consumption.

3. Res3DTENET ARCHITECTURE: HAND GESTURE CLASSIFICATION

- Res3DTENet consists of two main modules in sequential order: * Residual 3D-CNN (Res3D-CNN); * Transformer Encoder Network (TENet).
- SOLI Hand Gesture Dataset: 11 gesture classes, 10 subjects, 25 instances per subject per gesture = 2750 sequences
 Approximately 40 frames per gesture instance/sequence
 - 11 gesture classes, 1 subject, 50 instances per subject per gesture, 5 sessions = 2750 sequences





(a) HGR network to classify spatio-temporal RD features



(b) Feature maps from 6th CNN Layer

2. MAJOR CHALLENGES

- Low Signal-to-Noise Ratio (SNR) environments due to the presence of non-stationary and unexpected back-ground.
- High variability of gestures in terms of scale, nonuniform frame rate, and measured distance.
- Lack of leveled data, high intra-class and low interclass variation in features

4. Res3DTENET ARCHITECTURE: TRANSFORMER ENCODER

- Transformer Encoder N/W consists of 2 encoder layers and FC layer.
- Each encoder block consists of 3 attention heads, feed forward n/w, layer norm.
- Trained over spatio-temporal RD features to refine temporal inter-relationship b/w frames

Res3D-CNN									
Layer name	Residual block	Kernel size	No. of filters	Output size $32 \times 32 \times 40 \times 4$					
Input			-2						
Conv1	-	$3 \times 3 \times 3$	16	$32 \times 32 \times 40 \times 16$					
Conv2	R1	$3 \times 3 \times 3$	16	$32 \times 32 \times 40 \times 16$					
Conv3	R1	$3 \times 3 \times 3$	16	$32 \times 32 \times 40 \times 16$					
Max- pooling		$2 \times 2 \times 1$	80	$16 \times 16 \times 40 \times 16$					
Conv4		$3 \times 3 \times 3$	32	$16 \times 16 \times 40 \times 32$					
Conv5	R2	$3 \times 3 \times 3$	32	$16 \times 16 \times 40 \times 32$					
Conv6	R2	$3 \times 3 \times 3$	32	$16 \times 16 \times 40 \times 32$					
Output			-	$16 \times 16 \times 40 \times 32$					



• Sequential n/w i.e., RNN/ LSTM have limited ability to learn temporal dynamics of multi-channel range-Doppler sequences



(b) Transformer Encoder

5. Important Findings

 Transformer n/w outperforms the LSTM network in terms of faster training and capturing long-term dependencies.

6. EXPERIMENTAL ANALYSIS

□ -120.0
 0.0
 1.0
 1.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0
 0.0<

cy bas 96.99 93.03 78.92	sed on 96.77 93.55	50:50 99.19	0 train	ing a	nd tes	ting s	plit of	the	lata	12	<u>.</u>
96.99 93.03 78.92	96.77 93.55	99.19	91.13	01 04	-			enc c	iata		
93.03 78.92	93.55			91.94	100	100	100	96.77	100	99.19	91.94
78.92	A CONTRACTOR OF A CONTRACTOR O	100	77.42	78.23	91.94	100	100	91.13	100	99.19	91.94
	62.22	64.88	45.07	89.30	67.20	94.62	94.89	87.34	92.50	85.78	84.39
91.12	87.90	96.77	83.87	83.06	87.10	95.16	88.71	91.13	97.58	91.94	99.19
87.17	67.72	71.09	77.78	94.48	84.84	98.45	98.63	88.89	94.85	89.56	92.63
77.71	60.35	62.25	38.72	89.45	66.77	92.52	94.93	86.89	91.39	85.52	86.22
96.77	97.58	100	98.38	83.06	100	100	99.19	99.19	96.77	98.38	91.93
leave	one s	subjec	t out:	cross	Subj	ect Va	lidatio	on on	10 su	bjects	0
92.25	89.12	93.34	92.20	83.43	84.66	93.50	97.68	100	95.78	93.22	91.84
79.06	58.71	67.62	64.80	91.82	72.31	72.91	93.40	89.99	95.16	82.80	80.24
91.38	84.80	98.40	88.00	78.40	87.60	99.20	90.00	99.20	96.40	93.99	89.20
leave	one s	session	i out:	cross	Sessio	n Val	idatio	n on 1	L Subj	ect	
92.98	92.03	100	92.00	60.15	98.24	100	100	100	100	100	80.42
85.75	56.69	61.98	76.43	96.83	92.73	81.38	98.42	97.79	95.33	96.92	89.10
97.75	94.33	99.33	97.76	90.33	97.66	100	100	99.66	100	99.66	96.66
	77.71 96.77 leave 92.25 79.06 91.38 leave 92.98 85.75 97.75	77.71 60.35 96.77 97.58 leave one s 92.25 89.12 79.06 58.71 91.38 84.80 leave one s 92.98 92.98 92.03 85.75 56.69 97.75 94.33	77.71 60.35 62.25 96.77 97.58 100 leave one subject 92.25 89.12 93.34 79.06 58.71 67.62 91.38 84.80 98.40 leave one session 92.98 92.03 100 85.75 56.69 61.98 97.75 94.33 99.33	77.71 60.35 62.25 38.72 96.77 97.58 100 98.38 leave one subject out: 92.25 89.12 93.34 92.20 79.06 58.71 67.62 64.80 91.38 84.80 98.40 88.00 leave one session out: 92.98 92.03 100 92.00 85.75 56.69 61.98 76.43 97.76	77.71 60.35 62.25 38.72 89.45 96.77 97.58 100 98.38 83.06 leave one subject out: cross 92.25 89.12 93.34 92.20 83.43 79.06 58.71 67.62 64.80 91.82 91.38 84.80 98.40 88.00 78.40 leave one session out: cross 92.98 92.03 100 92.00 60.15 85.75 56.69 61.98 76.43 96.83 97.75 94.33 99.33 97.76 90.33	77.71 60.35 62.25 38.72 89.45 66.77 96.77 97.58 100 98.38 83.06 100 leave one subject out: cross Subject out: subject out: cross Subject out:	77.71 60.35 62.25 38.72 89.45 66.77 92.52 96.77 97.58 100 98.38 83.06 100 100 leave one subject out: cross Subject Va 92.25 89.12 93.34 92.20 83.43 84.66 93.50 79.06 58.71 67.62 64.80 91.82 72.31 72.91 91.38 84.80 98.40 88.00 78.40 87.60 99.20 leave one session out: cross Session Val 92.98 92.03 100 92.00 60.15 98.24 100 85.75 56.69 61.98 76.43 96.83 92.73 81.38 97.75 94.33 99.33 97.76 90.33 97.66 100	77.71 60.35 62.25 38.72 89.45 66.77 92.52 94.93 96.77 97.58 100 98.38 83.06 100 100 99.19 leave one subject out: cross Subject Validation 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 79.06 58.71 67.62 64.80 91.82 72.31 72.91 93.40 91.38 84.80 98.40 88.00 78.40 87.60 99.20 90.00 leave one session out: cross Session Validation 92.98 92.03 100 92.00 60.15 98.24 100 100 85.75 56.69 61.98 76.43 96.83 92.73 81.38 98.42 97.75 94.33 99.33 97.76 90.33 97.66 100 100	77.71 60.35 62.25 38.72 89.45 66.77 92.52 94.93 86.89 96.77 97.58 100 98.38 83.06 100 100 99.19 99.19 leave one subject out: cross Subject Validation on 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 92.06 58.71 67.62 64.80 91.82 72.31 72.91 93.40 89.99 91.38 84.80 98.40 88.00 78.40 87.60 99.20 90.00 99.20 leave one session out: cross Session Validation on 1 92.98 92.03 100 92.00 60.15 98.24 100 100 100	77.71 60.35 62.25 38.72 89.45 66.77 92.52 94.93 86.89 91.39 96.77 97.58 100 98.38 83.06 100 100 99.19 99.19 96.77 leave one subject out: cross Subject Validation on 10 su 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 95.78 79.06 58.71 67.62 64.80 91.82 72.31 72.91 93.40 89.99 95.16 91.38 84.80 98.40 88.00 78.40 87.60 99.20 90.00 99.20 96.40 leave one session out: cross Session Validation on 1 Subj 92.98 92.03 100 92.00 60.15 98.24 100 100 100 85.75 56.69 61.98 76.43 96.83 92.73 81.38 98.42 97.79 95.33 97.75 94.33 99.33 97.76 90.33 97.66 100 100 99.66 100 <td>77.71 60.35 62.25 38.72 89.45 66.77 92.52 94.93 86.89 91.39 85.52 96.77 97.58 100 98.38 83.06 100 100 99.19 99.19 96.77 98.38 leave one subject out: cross Subject Validation on 10 subjects 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 95.78 93.22 79.06 58.71 67.62 64.80 91.82 72.31 72.91 93.40 89.99 95.16 82.80 91.38 84.80 98.40 88.00 78.40 87.60 99.20 90.00 99.20 96.40 93.99 leave one session out: cross Session Validation on 1 Subject 92.98 92.03 100 92.00 60.15 98.24 100 100 100 100 85.75 56.69 61.98 76.43 96.83 92.73 81.38 98.42 97.79 95.33 96.92 97.75 94.33 99.33</td>	77.71 60.35 62.25 38.72 89.45 66.77 92.52 94.93 86.89 91.39 85.52 96.77 97.58 100 98.38 83.06 100 100 99.19 99.19 96.77 98.38 leave one subject out: cross Subject Validation on 10 subjects 92.25 89.12 93.34 92.20 83.43 84.66 93.50 97.68 100 95.78 93.22 79.06 58.71 67.62 64.80 91.82 72.31 72.91 93.40 89.99 95.16 82.80 91.38 84.80 98.40 88.00 78.40 87.60 99.20 90.00 99.20 96.40 93.99 leave one session out: cross Session Validation on 1 Subject 92.98 92.03 100 92.00 60.15 98.24 100 100 100 100 85.75 56.69 61.98 76.43 96.83 92.73 81.38 98.42 97.79 95.33 96.92 97.75 94.33 99.33

(a) Res3D-CNN



20

10

- Residual learning helps in training deep network more easily and leads to better generalization.
- Micro motion gestures like finger slide and finger rub are least classifying gesture classes.
- Smaller training data may reduce n/w convergence and its generalization capability.

· · · ·	9.0	0.0	1.0	114.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4 -	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0	0.0	0.0	
<u>ہ</u>	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0	0.0	
9 -	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0	
L -	1.0	0.0	0.0	0.0	1.0	0.0	0.0	120.0	0.0	0.0	2.0	
∞ -	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	
ი -	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	123.0	0.0	
9 ⁻	9.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	114.0	
	ò	i	2	3	4	5	6	7	8	9	10	

(a) Confusion matrix: 50:50 training and testing

(b) Classification Performance

REFERENCES

 Wang, S., Song, J., Lien, J., Poupyrev, I., Hilliges, O.: 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, pp. 851-860, 2016.
 Berenguer, A.D., Oveneke, M.C., Alioscha-Perez, M., Bourdoux, A., Sahli, H., et al.: Gesturevlad: Combining unsupervised features representation and spatio- temporal aggregation for doppler-radar gesture recognition. IEEE Access 7, pp. 137122-137135, 2019.
 Chen, K.S.: Principles of Synthetic Aperture Radar Imaging: A System Simulation Approach, vol. 2. CRC Press (2016)
 Choi., Ryu, S.J., Kim, J.H.: Short-Range Radar Based Real-Time Hand Gesture Recognition using LSTM Encoder. pp. 33610- 33618 (2019)