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

Gaurav Jaswal[a], Seshan Srirangarajan[a,b], Sumantra Dutta Roy[a]
[a]Department of Electrical Engineering
[b]Bharti School of Telecommunication Technology and Management
Indian Institute of Technology Delhi, India

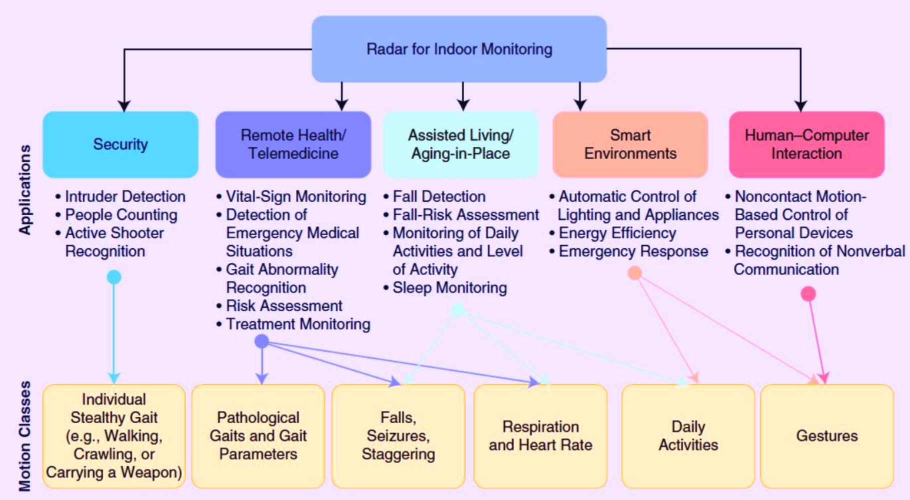
Presented by: Gaurav Jaswal

December 25, 2020

Role of Hand Gestures in HCI

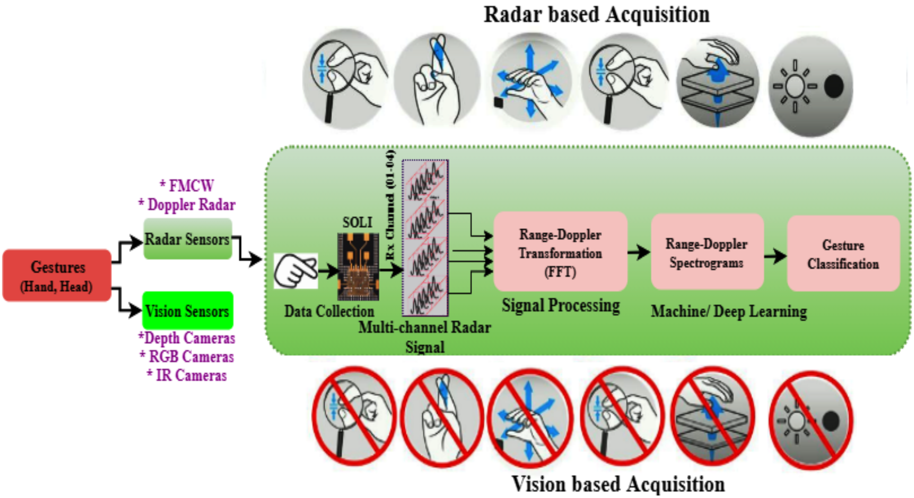
- There is growing popularity of new ways to interact with computer interfaces. For example, wearable devices such as small watches and head mounted display.
- Existing keyboard or touch based interaction paradigm is slow.
- Natural, low effort and high precision interaction paradigm is urgently needed.
- How can we use a small set of fine finger gestures or hand movements to control everything around us?
- Camera or vision sensors 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.

Radar based Motion Sensing Applications



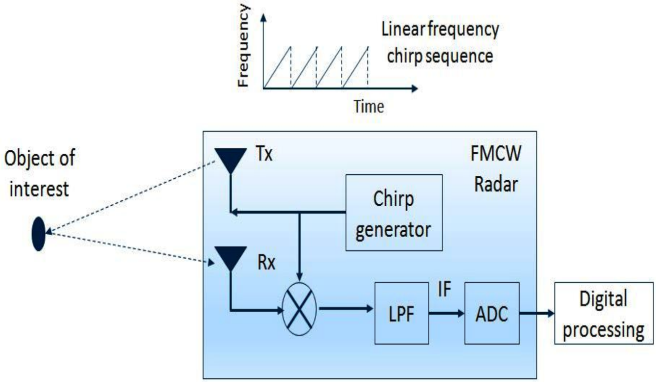
Source: Gurbuz and Amin (2019)

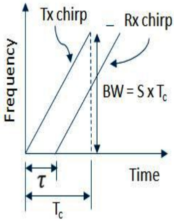
Motivation: Radar based Gesture Recognition



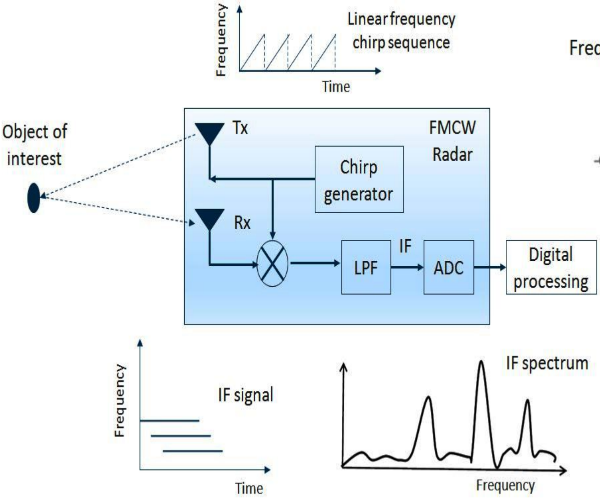
Frequency Modulated Continuous Wave (FMCW) Radar: Basic Principle

- Send a chirp whose frequency changes linearly over time.
- Estimate the frequency difference.

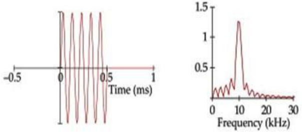




FMCW Radar....Contd



Frequency resolution and range resolution



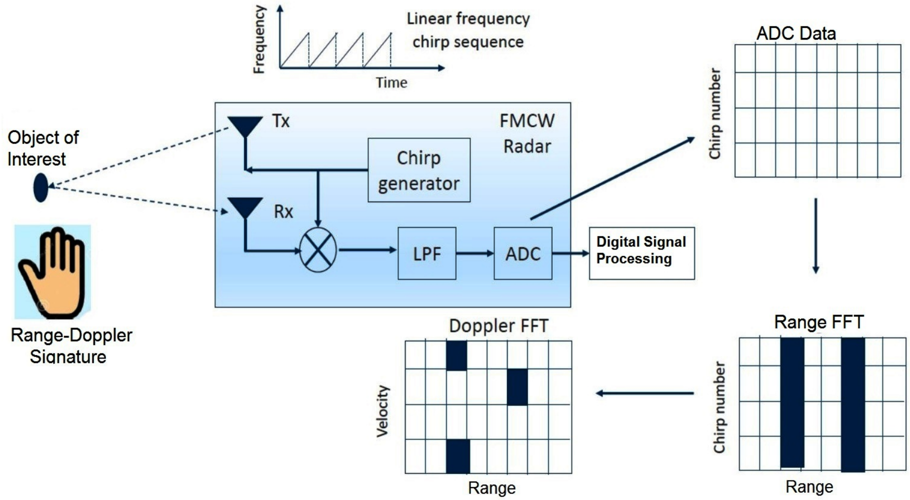
$$\text{Main lobe width} \propto \frac{1}{\text{Observation window}}$$

$$\Delta f > \frac{1}{T_c}$$

$$\frac{S(2\Delta d)}{c} > \frac{1}{T_c} \qquad \boxed{d_{res} = \frac{c}{2BW}}$$

Maximum range of the radar

FMCW Radar.... Contd.



Radar based Gesture Recognition Methods

Network model	Proposed framework	Sensor	Data set	Training scheme and accuracy
CNN-LSTM[2]	Deployed deep learning framework including CNN and RNN models for dynamic HGR	FMCW radar (at 60 GHz)	2750 gesture sequences: 10 users, 11 gesture classes	87% (50%-50%), 79.06% and 85.75% (cross validation)
Gesture-VLAD[3]	Frame level aggregation for extracting temporal information	-	Soli data set [2]	91.06% (frame-level), 98.24% (sequence-level)
Triplet loss [4]	Feature embedding using 3DCNN in conjunction with triplet loss	FMCW radar (at 24 GHZ)	9000 sequences: 6 gesture classes, 10 subjects	94.50% (triplet loss)
3DCNN-LSTM-CTC[5]	LSTM-CTC, fusion algorithm to classify spatio-temporal features of gesture sequences	FMCW radar (at 24 GHz)	3200 sequences: 4 subjects, 8 gesture classes, 100 instances	95% (3 s), 92.7% (2 s), 85.2% (1 s)
TSI3D [6]	Range-Doppler time feature sequences using I3D and two LSTM n/ws	FMCW radar (at 4 GHz)	4000 sequences: 10 gesture classes, 400 instances	96.17% (TS-I3D), 94.72% (without IE), 93.05% (I3D-LSTM)
AE [7]	Autoencoder network to learn micro hand motion representation	FMCW radar (at 24GHz)	3200 sequences: 4 subjects, 8 gesture classes	95% (0.5m), 87% (0.3m), 76% (0.1m)

Major Challenges

- Acquisition of unobtrusive and low-effort gestures irrespective of people diversity is difficult and time-consuming.
- Range-Doppler signature of hand motion is often influenced by other flicks of body parts, which leads to distorted motion features.
- Low signal-to-noise ratio (SNR) environments due to the presence of non-stationary and unexpected background.
- High variability of gestures in terms of scale among different subjects.
- Non-uniform frame rate and measured distance.
- Lack of labeled data, high intra-class and low inter-class variation in the features.
- Sequential networks such as RNN/LSTM have limited ability to learn temporal dynamics of multi-channel range-Doppler sequences.

Dataset Used: SOLI Hand Gesture Dataset[1]

– Approximately 40 frames per gesture instance/sequence

Dataset	Subjects	Classes	Sequences
SOLI Phase-1	10 subjects	11 gesture classes recorded per subject in 25 times	2,750 = (11*25*10)
SOLI Phase-2	1 subject	11 gesture classes recorded by 1 subject in 50 times and 5 sessions	2,750 = (11*50*5)

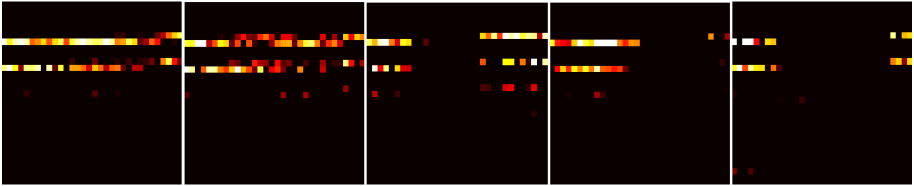


Figure: Sample range-Doppler frames

[1] https://github.com/simonwsw/deep-soli

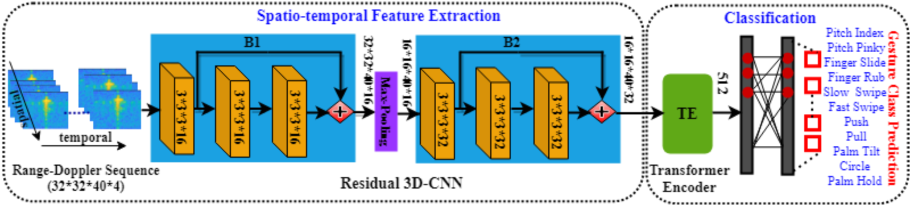
Model Architecture: Res3DTENet

– **Res3DTENet** consists of two main modules in sequential order:

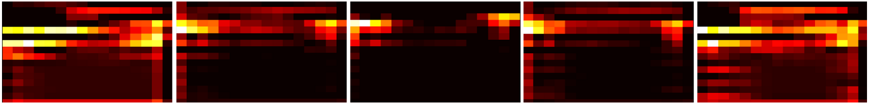
* Residual 3D-CNN (Res3D-CNN)

* Transformer Encoder Network (TENet)

– 11 Gesture Class Classification



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



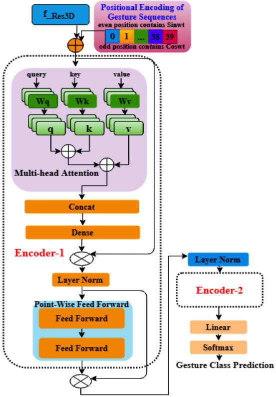
(b) Feature map visualization from the 6th CNN layer

Model Architecture......Contd

- Res3DCNN: 6 3D-Conv layers, 1 max-pooling, 2 residual blocks.
- Res3DCNN learns spatio-temporal RD features.
- Transformer Encoder N/W: 2 encoder layers (each encoder consists 3 attention heads, feed forward n/w, layer norm).
- Transformer Encoder refines inter-frame temporal dynamics.

Res3D-CNN				
Layer name	Residual block	Kernel size	No. of filters	Output size
Input	-	-	-	$32 \times 32 \times 40 \times 4$
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	-	$2 \times 2 \times 1$	-	$16 \times 16 \times 40 \times 16$
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	-	-	-	$16 \times 16 \times 40 \times 32$

(a) Res3D-CNN



(b) Transformer Encoder N/W

Experimental Analysis: Res3DTENet

- Network Training
  - Batch size 16
  - Cross entropy loss

– To validate the utility of Res3DTENet, we have performed two experimental analysis:

- Evaluation using 50:50 Training and Testing Split
  - Res3DTENet; 3DCNNTENet; TENet; Res3DLSTM
  - SOLI [2]
  - GVLAD [3]
- Evaluation Using Cross Validation.
  - Leave One Subject Out
  - Leave One Session Out

[2] https://doi.org/10.1145/2984511.2984565

[3] https://doi.org/10.1109/ACCESS.2019.2942305

Experimental Analysis......Contd.

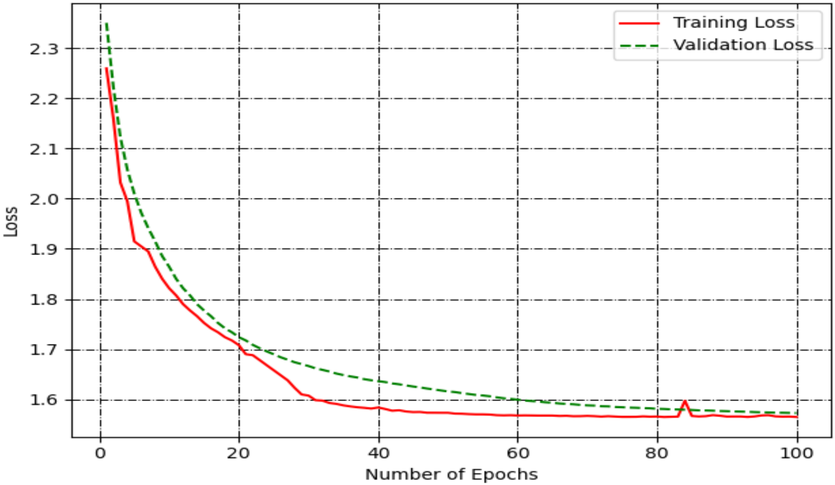


Figure: Training and validation loss curves

Experimental Analysis.....Contd.

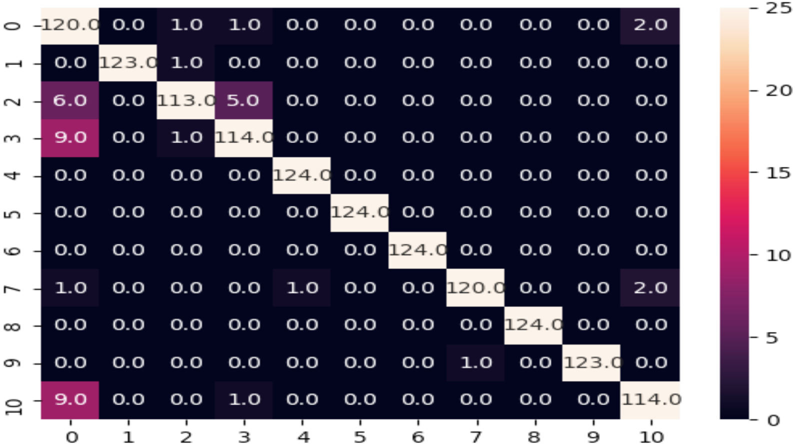


Figure: Confusion matrix for the proposed Res3DTENet model based on a 50:50 training and testing strategy

Experimental Analysis.....Contd.

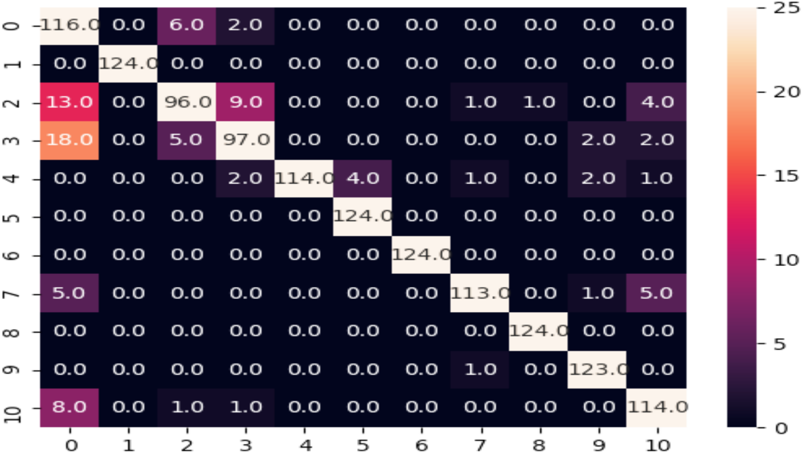


Figure: Confusion matrix for the proposed 3DCNNTENet model based on a 50:50 training and testing strategy

Experimental Analysis......Contd.

Network	Avg.	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11
Accuracy based on 50:50 training and testing split of the data												
Res3DTENet	96.99	96.77	99.19	91.13	91.94	100	100	100	96.77	100	99.19	91.94
3DCNNTENet	93.03	93.55	100	77.42	78.23	91.94	100	100	91.13	100	99.19	91.94
TENet	78.92	62.22	64.88	45.07	89.30	67.20	94.62	94.89	87.34	92.50	85.78	84.39
Res3D-LSTM	91.12	87.90	96.77	83.87	83.06	87.10	95.16	88.71	91.13	97.58	91.94	99.19
Soli (CNN-LSTM)[2]	87.17	67.72	71.09	77.78	94.48	84.84	98.45	98.63	88.89	94.85	89.56	92.63
Soli (RNN-shallow) [2]	77.71	60.35	62.25	38.72	89.45	66.77	92.52	94.93	86.89	91.39	85.52	86.22
GVLAD (without CG) [3]	96.77	97.58	100	98.38	83.06	100	100	99.19	99.19	96.77	98.38	91.93
GVLAD [3]	98.24	91.12	99.19	99.19	95.96	100	100	100	100	100	100	95.16
Accuracy using leave one subject out: cross subject validation on 10 subjects												
Res3DTENet	92.25	89.12	93.34	92.20	83.43	84.66	93.50	97.68	100	95.78	93.22	91.84
Soli [2]	79.06	58.71	67.62	64.80	91.82	72.31	72.91	93.40	89.99	95.16	82.80	80.24
GVLAD [3]	91.38	84.80	98.40	88.00	78.40	87.60	99.20	90.00	99.20	96.40	93.99	89.20
Res3DTENet	92.25	89.12	93.34	92.20	83.43	84.66	93.50	97.68	100	95.78	93.22	91.84
Accuracy using leave one session out: cross session validation on 10 subjects												
Res3DTENet	92.98	92.03	100	92.00	60.15	98.24	100	100	100	100	100	80.42
Soli [2]	85.75	56.69	61.98	76.43	96.83	92.73	81.38	98.42	97.79	95.33	96.92	89.10
GVLAD [3]	97.75	94.33	99.33	97.76	90.33	97.66	100	100	99.66	100	99.66	96.66

Important Findings

- Applications of automated HGR range from micro electronic wearable devices, sign language recognition to driver assistance system.
- Unlike camera, RF sensor (Doppler-radar or FMCW) easily detects hand movements in short range, independent of light conditions and security issues.
- Transformer n/w outperforms the LSTM network in terms of faster training and capturing long-term dependencies.
- 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.

Bibliography

1. Gurbuz, S. Z., Amin, M. G.: Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring. IEEE Signal Processing Magazine, 36(4), pp. 16-28, 2019.
2. Wang et al.: 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.
3. Berenguer et al.: Gesturevlad: Combining unsupervised features representation and spatiotemporal aggregation for doppler-radar gesture recognition. IEEE Access 7, 137122-137135, 2019.
4.Amin, M.G., Zeng, Z., Shan, T.: Hand Gesture Recognition Based On Radar micro-Doppler Signature Envelopes. In: IEEE Radar Conference (RadarConf). pp. 1-6, 2019.
5. Lien et al.: Soli: Ubiquitous Gesture Sensing With millimeter Wave Radar. vol. 35, pp. 1-19, 2016.
6. Wang et al.: Ts-i3d based hand gesture recognition method with radar sensor. IEEE Access 7, pp. 22902-22913, 2019.
7. Zhang et al.: u-deephand: Fmcw radar based unsupervised hand gesture feature learning using deep convolutional autoencoder network. IEEE Sensors Journal 19(16), pp. 6811-6821, 2019.
8. Zhang et al.: Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. IEEE Sensors Journal 18(8), 3278-3289, 2018.