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

Gaurav Jaswal[1][0000−0002−3971−0160],
Seshan Srirangarajan[1,2][0000−0003−4424−638X], and
Sumantra Dutta Roy[2][0000−0002−2141−5067]

[1] Department of Electrical Engineering
[2] Bharti School of Telecommunication Technology and Management
Indian Institute of Technology Delhi, New Delhi 110016, India
{gauravjaswal; seshan; sumantra}@ee.iitd.ac.in

Abstract. Recently hand gesture recognition via millimeter-wave radar has attracted a lot of research attention for human-computer interaction. Encouraged by the ability of deep learning models in successfully tackling hand gesture recognition tasks, we propose a deep neural network (DNN) model namely, Res3DTENet that aims to classify dynamic hand gestures using the radio frequency (RF) signals. We propose a scheme that improves the convolutional process of 3DCNNs with residual skip connection (Res3D) to emphasize local-global information and enriches the intra-frame spatio-temporal feature representation. A multi-head attention transformer encoder (TE) network has been trained over the spatio-temporal features to refine the inter-frame temporal dependencies of range-Doppler sequences. The experiments are carried out on the publicly available Soli hand gesture data set. Based on our extensive experiments, we show that the proposed network achieves improved gesture recognition accuracy than the state-of-the-art hand gesture recognition methods.

Keywords: range-Doppler · residual learning · transformer network

1 Introduction

With the increasing prevalence of personal handheld devices in our daily lives, there has been a growing interest in the area of human-computer interaction (HCI) [1, 12]. Recently, millimeter wave radar and other similar sensors are being explored for human-computer interaction applications [3]. The evolution of computing power and miniaturization of the hardware has led researchers to explore new ways of interacting with the electronic devices [6]. The use of millimeter wave (mm-Wave) radar technology for gesture sensing is a step in this direction. Traditionally, radar systems have been used to locate, track, and create two or three-dimensional reconstructions of objects such as airplanes, rockets, and missiles. Real aperture radar (RAR) and synthetic aperture radar (SAR) are examples of the radar being used as image sensing modalities [3]. Radar-based hand gesture recognition systems need to address the following challenges before they can become popular: (i) identifying micro-motion gesture features through a machine learning (ML) model usually requires large labeled data set to avoid over-fitting. However, the acquisition of unobtrusive and low-effort gestures irrespective of diversity is difficult and time-consuming, and (ii) the Doppler signature for hand motion is often also influenced by the motion of other body parts, which leads to distorted motion features, resulting in low recognition accuracy.

Recently, a mmWave solid-state Soli radar technology was reported [6] and its capability in recognizing hand gestures was demonstrated. This was a breakthrough in using mmWave radar as a sensing device for HCI applications and is attractive for several reasons. Firstly, radar can overcome some of the limitations of optical and depth camera systems. Though these cameras have shown good results in recognizing sign and action gestures, these sensing modalities show poor performance when dealing with occluded objects. Hand gestures need to be performed in line of sight (LoS) for these optical systems. The optical camera needs to be placed/fixed in the front panel of the devices and should be visible to the user. This can become a constraint when the form factor of the devices becomes small such as wristwatches. mmWave radar overcomes these drawbacks as it uses radio waves to interact with and recognize the objects. The use of sound waves to recognize hand gestures has also been explored but these methods have been capable of recognized only a limited set of hand gestures. In [10], RF sensing has been used to recognize 11 different hand gestures. It allows the RF sensor to be positioned anywhere within the device and the sensor forms a beam pattern within its range so that the user can perform hand gestures even without being in LoS with the sensor. These sensors have potential to be used in many consumer applications and thus there is a need to work on radar-based sensing to build classifier systems that can recognize hand gestures accurately and efficiently.

1.1 Problem Statement and Contribution

We tackle the challenge of developing a high performance hand gesture recognition system using a small training data set. The existing DNN frameworks fail to fully exploit the available spatio-temporal information w.r.t. hand position and direction in the gesture sequences. This manifests in the form of two problems: (ii) ambiguity: when different gestures are confused as one, and (ii) variability: when the same gesture performed under different conditions gets classified as different gestures. In addition, the sequential networks such as RNN/LSTM have limited ability to learn temporal dynamics of multi-channel range-Doppler sequences and result in high computations during the training and testing phases. In this article we address the above challenges by proposing a robust gesture recognition solution which is invariant to the subject, hand position and direction, and is capable of learning the inherent range-Doppler spatio-temporal features.

Our Contribution: Motivated by the success of deep residual learning and recent developments in transformer networks, we propose a 3D residual transformer encoder network (Res3DTENet) which is an end-to-end trainable network for classification of hand gestures using the range-Doppler sequences. In this article, we consider a radar gesture sequence as input to a Res-3DCNN with dense skip connections for allowing the flow of spatial information to the deeper layers. We train the transformer encoder (TE) to learn the positional transitions of input embedding across the temporal dimension. Compared to the traditional RNN [10] or LSTM [5, 15], the proposed architecture will be shown to be better at capturing the dynamic long-range temporal dependencies in the input-output sequences while speeding up the training time. Experimental results based on the Soli data set [10] demonstrate that the proposed network has a strong generalization ability for classifying micro hand gestures.

2 Literature Survey

Various radar-based gesture recognition techniques using Doppler radar [2], FMCW radar [14], and ultra wideband radar [7] have been studied. In addition, ultrasound, RFID, and Wi-Fi based gesture sensing technologies have been explored, however these are significantly affected by the propagation speed, diffraction of sound, or other interference in the environment. In contrast, mmWave radar [12] based gesture recognition solutions have the advantage of a small form factor, sensing through smoke, dust, non-metallic materials, and can provide good range resolution. Soli [6] is among the first mmWave radar solutions for gesture based applications on consumer electronic devices. Various gesture recognition methods based on the mmWave radar have been proposed with the gesture features being extracted from range–Doppler images [10, 5], range-Doppler sequences [2], micro-Doppler images [12], and range-time images [15]. In Soli [6], a mmWave radar hardware set-up and gesture acquisition pipeline were described. The authors collected a gesture set with four hand gestures and extracted low-dimensional features such as range, velocity, acceleration, velocity, centroids, and total energy, and used a random forest classifier for gesture recognition. In another work on Soli [10], a sequence of range-Doppler images is given as input to train a CNN-RNN network and the recognition capabilities were demonstrated over large a gesture data set with eleven hand gesture classes and a total of 2750 gesture sequences. Similarly in [4], range-Doppler image features are extracted to recognize hand gestures. The authors used CFAR algorithm to detect hand gestures, and then employed an LSTM network over the motion profile sequences for gesture recognition. In [8], continuous-wave radar at 2.4 GHz is employed and I-Q trajectory images are generated from the time-domain I and Q plots of the radar echos. Authors employed a CNN based approach using the I-Q images for recognizing the hand gestures and reported a recognition accuracy of 96%. In [13], a Doppler radar with dual receiver channels operating from 5-8 GHz is used and scalogram images are generated using time-frequency analysis tools (STFT and CWT). Finally, a CNN is trained to classify the hand gestures and a recognition accuracy of 98% has been reported. In [5], authors proposed a few-shot learning using 3DCNN which accepts video files of radar-Doppler image sequences as input and the model is trained using the triplet loss function for classifying the hand gestures. CNN ensures that the gestures have minimum intra-class variance and maximum inter-class variance. In [1], authors demonstrated gesture recognition using radar micro-Doppler signature envelopes. The authors collected a data set of 15 hand gestures and represent the gestures using PCA and canonical angle metric features. They also extract envelope features using the spectrograms and demonstrate hand gesture recognition using K-nearest neighbor (KNN) and support vector machine (SVM) algorithms. A summary of the state-of-the-art gesture recognition methods is listed in Table 1.

3 Network Architecture

We propose a novel deep learning framework consisting of two main modules in sequential order: Res3DCNN and TENet, as shown in Fig. 1. This end-to-end trainable network integrates residual learning and temporal relationship learning for fine-grained gesture recognition.

3.1 Residual 3DCNN (Res3D)

The motivation for incorporating residual learning in 3DCNN is to capture the contextual information and receptive fields. In addition, residual skip connections provide easy gradient flow through the network and addresses the problem of vanishing gradients effectively. CNN tends to lose the local feature representation of the range-Doppler sequences at deeper layers and only preserves the aggregated global features for classification. In order to discover more local-global salient spatial representation from the range-Doppler sequences, we introduce residual 3DCNN with dense skip connections for fine grained classification. The architecture of Res3D consists of two CNN blocks (B1, B2), where each block contains three consecutive 3D CNN layers with ReLU activation. The input (**I**) provided to the network is of size $32 \times 32 \times 40 \times 4$ representing height, width, frames, and channels, respectively. In particular, the first 3D CNN layer (filter size $3 \times 3 \times 3$ and 16 filters) in block B1 is given I and produces a feature map $\mathrm{B1}_1$ with size $32 \times 32 \times 40 \times 16$. Subsequently, the other two convolutional layers of the B1 block are applied and the output is $\mathrm{B1}_2$. To incorporate residual learning, a skip connection from input to the second convolutional layer ($\mathrm{B1}_1$) is followed by a concatenation with the feature map of the third 3DCNN layer ($\mathrm{B1}_2$) resulting in the aggregated feature map $\mathrm{B1}_R$. Consequently, a max pooling operation is applied over the aggregated feature map, which reduces the overall size of $\mathrm{B1}_R$ to $16 \times 16 \times 40 \times 16$ and provides contextual features. However, the introduction of additional pooling layers may reduce the spatial information significantly and increase the number of parameters in Res3D. This will lead to false positives or errors during classification. The B2 block with three 3DCNN layers is employed in a similar manner as block B1, but with 32 filters for each CNN layer. The final

Table 1: Overview of state-of-the-art radar-based gesture recognition methods.

Network model	**Proposed framework**	**Sensor/ Prototype**	**Data set**	**Training scheme and accuracy**
CNN-LSTM [10]	Deployed deep learning framework including CNN and RNN models for dynamic hand gesture classification	FMCW mmWave radar (at 60 GHz)	2750 gesture sequences with an average of 40 frames per sequence: 10 users, 11 gesture classes, 25 instances	87% (50%-50%), 79.06% (leave one subject out), 85.75% (leave one session out)
3DCNN-LSTM-CTC [6]	LSTM-CTC, fusion algorithm to classify spatio-temporal features of gesture sequences	Frequency modulated continuous wave (FMCW) radar (at 24 GHz)	3200 sequences with duration of 3 s: 4 subjects, 8 gesture classes, 100 instances	95% (3 s), 92.7% (2 s), 85.2% (1 s)
TS-I3D [11]	Range-time and Doppler-time feature sequences using I3D and two LSTM networks	FMCW radar with bandwidth of 4 GHz	4000 sequences each with 32 frames: 10 gesture classes, 400 instances	96.17% (TS-I3D), 94.72% (without IE), 93.05% (I3D-LSTM)
3DCNN with triplet loss [5]	Feature embedding using 3DCNN in conjunction with triplet loss	FMCW radar (at 24 GHZ)	9000 sequences each with 100 frames: 6 gesture classes, 10 subjects, 150 instances	94.50% (triplet loss), 86.30% (cross entropy loss), 88.10% (2DCNN-LSTM)
Gesture-VLAD [2]	Frame level representation and frame aggregation for extracting temporal information	Soli sensor [6]	Soli data set [6]	91.06% (frame-level), 98.24% (sequence-level)
Autoencoder [14]	Autoencoder network to learn micro hand motion representation	FMCW radar (at 24GHz)	3200 sequences each of duration 1.5 s: 4 subjects, 8 gesture classes, 100 instances	95% (0.5m), 87% (0.3m), 76% (0.1m)

output feature map of Res3D is $\mathrm{F}_{\mathrm{res}}$ (with size $16 \times 16 \times 40 \times 32$), which is given to a transformer module for fine grained classification. Further, in order to visualize the activation layer feature map representation, we select the pull gesture class (G8) to explain the learning process. The original frames and the learned feature maps from Res3D are shown in Fig. 2. The detailed parametrization of Res3D is presented in Table 2.

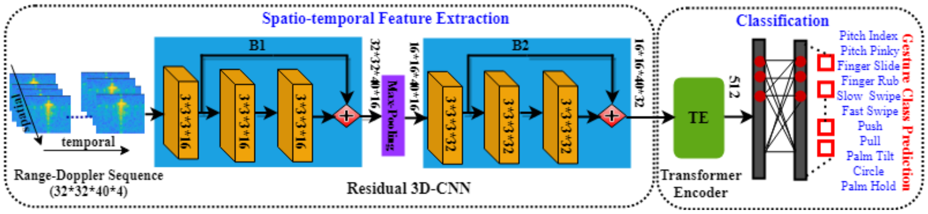
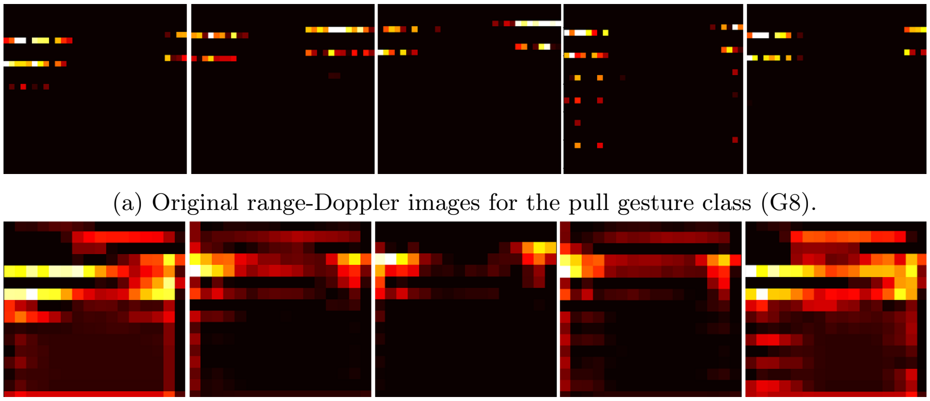


Fig. 1: Proposed network architecture for hand gesture classification.



(a) Original range-Doppler images for the pull gesture class (G8).

(b) Feature maps extracted by the Res3D-CNN model.

Fig. 2: Input sequence and corresponding feature map visualization.

We can mathematically represent the Res3D model as:

$$\mathrm{Res}(\mathbf{I}) = \mathbf{\Phi}\left(\mathbf{W}^{3} \circledast \left(\mathbf{\Phi}\left(\mathbf{W}^{2} \circledast \left(\mathbf{\Phi}\left(\mathbf{W}^{1} \circledast \mathrm{I}\right)\right)\right)\right)\right) + \mathbf{I} \tag{1}$$

where $\mathbf{W}^1$, $\mathbf{W}^2$, and $\mathbf{W}^3$ are the weights of layers L1, L2, and L3, respectively.

3.2 Transformer Encoder Network (TENet)

Sequence modelling or time series tasks require us to exploit temporal features and when the task involves spatio-temporal data, such as in the case of ges-

Table 2: Architecture of the Residual 3DCNN model.

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	-	-	-	$16 \times 16 \times 40 \times 32$

ture recognition, spatial as well as temporal features have to be extracted. To achieve this 3DCNN and RNN type architectures such as long short-term memory (LSTM) or gated recurrent unit (GRU) have been employed so far. Although, RNN type architectures are good at capturing long-term dependencies but as the sequence length increases they can suffer degradation in performance. Secondly, RNNs are slow to train as they lack parallelism due to their sequential information processing architecture. In order to address these problems of RNNs, transformer networks have been proposed recently.

Introduction: Transformer networks are a recent architecture capable of addressing multiple downstream sequential tasks and work on the basis of the attention mechanism to capture very long-term dependencies [9]. When in operation, all the sequential inputs are processed in parallel instead of time-step based processing. The network architecture consists of mainly two units: encoder and decoder. As the name suggests, encoder based on the learnt attention, processes information whereas the decoder uses the knowledge passed by the encoder to generate outputs and it also employs attention for learning representations. The overall architecture involves passing the input through encoder which is a stack of multiple encoder blocks. Subsequently, the processed output is given to different decoder blocks stacked together forming the decoder layer.

Implementation: This network includes several multi-head attention modules along with fully connected layers. In our experiment, we have only considered the transformer's encoder module (TENet) while leaving out the decoder module. The details of this TENet are depicted in Fig. 3. The TENet consists of 2 blocks with each block having a multi-head attention layer along with a position-wise fully connected feed-forward network. In addition, the residual connection with layer normalization has been utilized in each block. Initially, the input $\mathrm{f}_{\mathrm{Res3D}}$ is provided to the first block of the transformer. However, before being given to this block, positional encoding is computed and added to the input $\mathrm{f}_{\mathrm{Res3D}}$. Positional embedding incorporates order of the sequence, and provides unique encoding for each time-step. During computation of the positional encoding, sinusoidal function has been utilized, where $\sin(ut)$ and $\cos(ut)$ have been used for even and odd positions of the input, respectively. In this manner, two consecutive encoder blocks are used resulting in a feature vector $\mathrm{E2}_{\mathrm{out}}$ which is then given to the linear layer followed by a soft-max classifier for gesture classification.

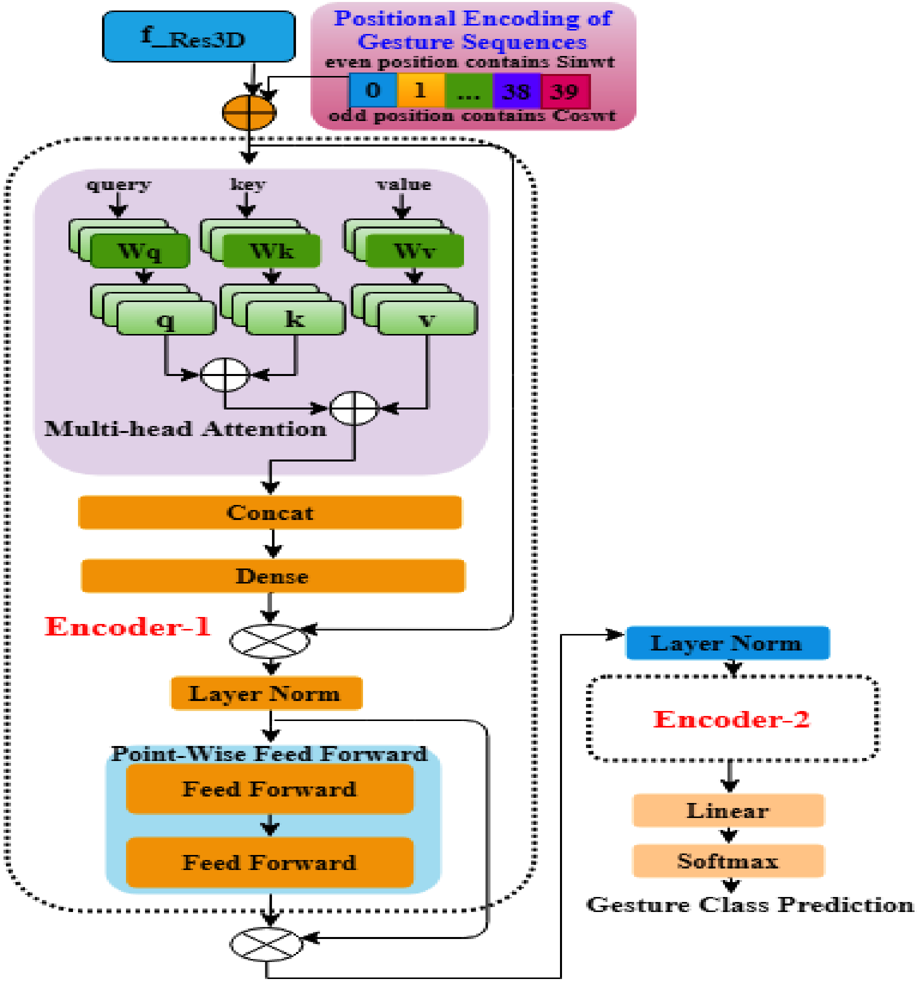


Fig. 3: Proposed network architecture for the transformer encoder network.

Multi-head Attention: This module contains several attention layers that run in parallel. In the attention layers, we define three input vectors query (q), keys (k), and values (v). Thus, the attention can be defined as:

$$\text{atten}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{m}}\right) v \tag{2}$$

We consider the dot product attention mechanism which is time-space efficient for our experiments. In our case, both query (q) and key (k) have the same dimension m whereas values has a different dimension n. Taking the dot product of q and k, followed by division by the square root of m, results in an m dimension vector A. Subsequently, softmax function is applied on A to obtain the attention vector with different weights for the different entries of v with the output as A_{atten}. Finally, the vector v is multiplied by A_{atten} to obtain the feature vector atten. In the multi-head attention module, we have parallel attention layers, where each layer is given an independent subset of the full input. Suppose we have 10 parallel attention layers then each layer will accept input of dimension $\frac{m}{10}$. Finally, the output of each attention layer is concatenated to form an m-dimensional vector. Subsequently, feed forward network consisting of two consecutive fully connected layers is applied. This feed forward network would be used separately at each position of the input.

The network architecture employed here consists of only the encoder blocks. There are a total of 3 encoder layers, with each layer consisting of 4 attention heads. We have experimented with 2 and 6 heads as well and observed that 3 encoders each with 4 attention heads produce optimal results. The outputs of the encoder blocks are fed into a dense layer and finally to a softmax layer for final classification. We the cross entropy loss function and stochastic gradient descent as the optimizer. We also experimented with different learning rates and observed that a value of 0.001 gave the best performance.

4 Experiment and Discussion

In this section, we describe in detail our experiments and the results obtained using the Soli data set. The quantitative evaluation of the proposed network will be presented in terms of classification accuracy of the gestures. We consider the classification accuracy per class and report the average accuracy.

4.1 Data Set

The performance of the Res3DTENet model is tested on publicly available Soli data set[3]. It contains 2750 gesture sequences with 11 hand gestures performed by 10 users in the first phase and a second set of 2750 sequences are collected in a second phase. Each hand gesture is available as a sequence of processed range-Doppler images in the h5 format. The data set was collected using a FMCW mmWave radar with four receive channels (named as channels 0, 1, 2 and 3).

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

4.2 Training

To evaluate the performance of different networks, we trained the proposed network and its variants and employed the same testing conditions for all the models. We used a uniform learning rate of 0.001 throughout the training phase and the stochastic gradient descent optimizer with a momentum of 0.9. A dropout of 0.5 is applied in the transformer encoder layer to avoid overfitting. It is observed that the network loss decreases within the first few epochs and converges by 35 epochs when batch-wise learning (with size 16) is carried out. The training and validation loss curves are shown in Fig. 4. It can be observed that both training and validation errors decrease and converge after around 35 or 40 epochs. To understand the network training, we use the cross-entropy loss that measures the difference between two probability distributions. Finally, confusion matrices are computed to evaluate the classification accuracy of our proposed end-to-end architecture.

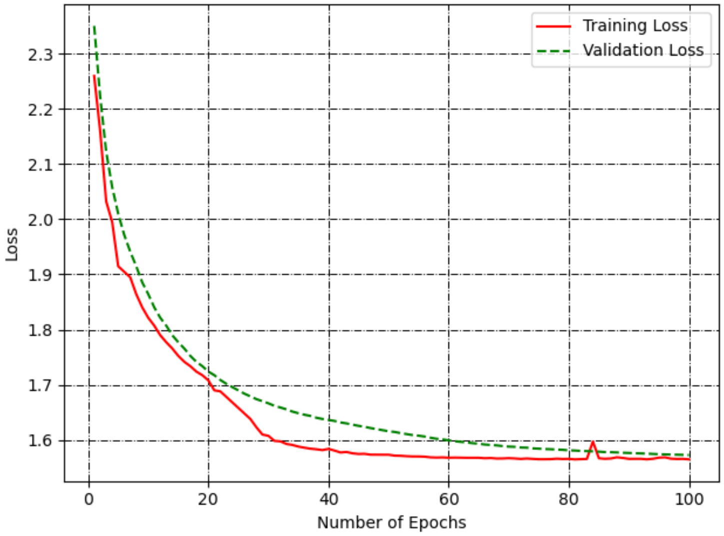
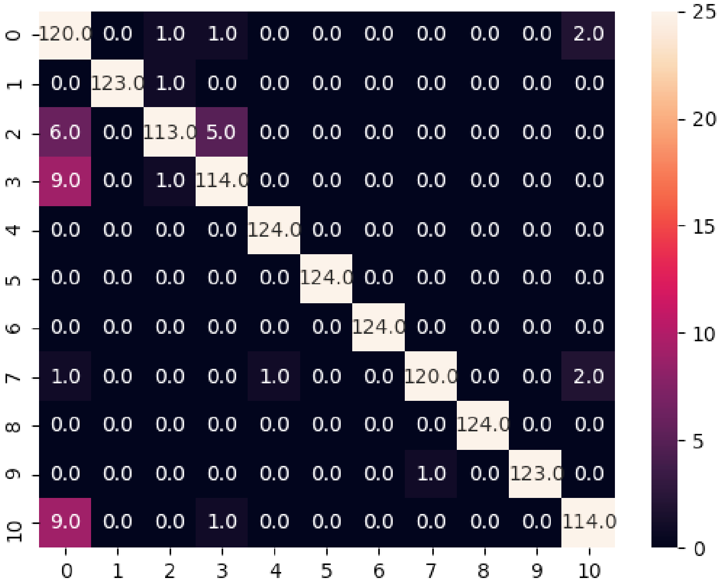


Fig. 4: Training and validation loss curves for the proposed gesture recognition network.

4.3 Evaluation using 50:50 Training and Testing Split

The proposed Res3DTENet and its three variants 3DCNNTENet, TENet, and Res3D-LSTM are trained using a 50:50 training and testing split, and the network parameters are selected empirically. Assuming the average time duration of each gesture to be the same, the length of each sequence is set to 40 frames.



	0	1	2	3	4	5	6	7	8	9	10
0	120.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
1	0.0	123.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	6.0	0.0	113.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	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
5	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0
7	1.0	0.0	0.0	0.0	1.0	0.0	0.0	120.0	0.0	0.0	2.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	123.0	0.0
10	9.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	114.0

Fig. 5: Confusion matrix for the proposed Res3DTENet model based on a 50:50 training and testing strategy.

Residual learning allows direct connections between alternate CNN layers that helps in training deep network more easily and leads to better generalization. Thus, Res3DTENet achieves an average classification accuracy of 96.99% with an improvement of 5.87% and 3.96% over the 3DCNNTENet and Res3DLSTM shown in Table 3. The confusion matrices for Res3DTENet and 3DCNNTENet are shown in Fig. 5 and Fig. 6, respectively. We also notice that the transformer networks (Res3DTENet, 3DCNNTENet) outperform the LSTM network (Res3DLSTM). However, the basic transformer architecture i.e., TENet (without CNN embedding) does not perform as well as the other DNN models. In this case, TENet receives the raw sequence as input and without a feature extractor it is unable to find robust features for classification. In terms of the individual gesture classes, seven gesture classes namely, pinch index (G1), pinch pinky (G2), slow swipe (G5), fast swipe (G6), push (G7), pull (G8), palm tilt (G9), and circle (G10) are recognized with more than 95% accuracy using the proposed network. On the other hand, using 3DCNNTENet, finger slide (G3) and finger rub (G4) have the lowest average recognition accuracy of 77.42% and 78.22%, respectively. However, Res3DTENet improves the recognition performance for finger slide (G3) and finger rub (G4). This result can be interpreted by the fact that finger slide and finger rub are associated with small motions. Finally, we can summarize that the introduction of transformer network in hand gesture recognition leads to better performance compared to the LSTM network.

Table 3: Classification accuracies obtained using different testing protocols.

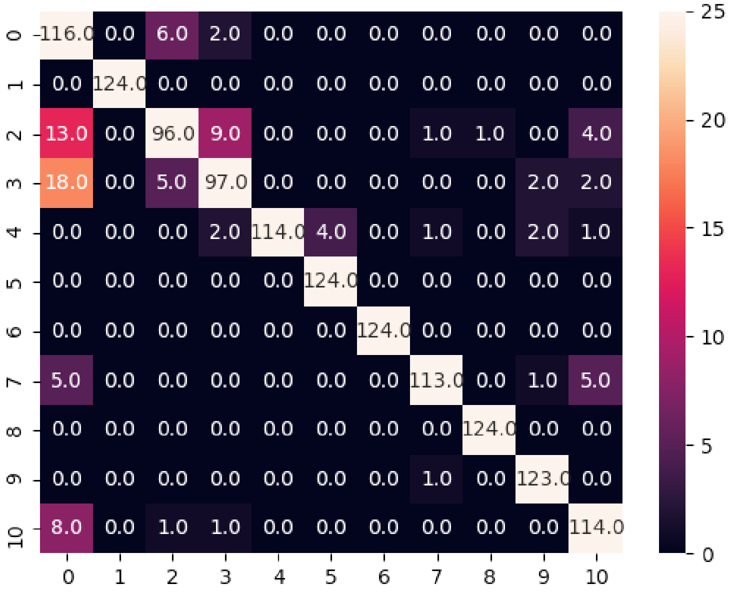
Network	**Avg.**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**
Accuracy based on 50:50 training and testing split of the data												
Res3DTENet	96.99	96.77	99.19	91.13	91.94	100	100	100	96.77	100	99.19	91.94
3DCNNTENet	93.03	93.55	100	77.42	78.23	91.94	100	100	91.13	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) [6]	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) [6]	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) [2]	96.77	97.58	100	98.38	83.06	100	100	99.19	99.19	96.77	98.38	91.93
GVLAD [2]	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 [6]	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 [2]	91.38	84.80	98.40	88.00	78.40	87.60	99.20	90.00	99.20	96.40	93.99	89.20

4.4 Evaluation Using Cross Validation

Next "leave one subject out" cross validation is performed using the proposed Res3DTENet architecture. The Soli data set consists of 2750 gesture sequences obtained from ten subjects. The accuracy results are reported in Table 3. Data for one gesture class is randomly removed and the network is trained by using the rest of the data set. This process is repeated for all the subjects. Finally, the average classification accuracy of 92.25% is achieved using the proposed network which is lower compared to the accuracy obtained using the 50:50 split of the data. In terms of class-wise performance for the different gesture classes, the micro-motion gestures such as pinch index (G1), finger rub (G4), and slow swipe (G5) result in lower recognition accuracy.

4.5 Performance Analysis

We also compare the results using the proposed architecture with the results reported in [10] and [2] as they use the same set of gestures and evaluation metrics. We consider two testing strategies: (i) 50:50 split of the data, and (ii) leave one subject out cross validation. Table 3 presents results in terms of average as well as individual classification accuracies for the different networks. It is observed that the proposed network outperforms the state-of-art CNN-LSTM [10] and GVLAD (without context gating) [2] models in terms of improved classification accuracy using a 50:50 split of the data. For the leave one subject out strategy, performance of the proposed network is comparable to both the CNN-LSTM [6] and GVLAD [2] networks. Since these employ 2D and 3D convolution, respectively, for extracting the features their networks are deeper but convergence rates are poor. Based on 3DCNN, the proposed network uses residual learning in 3D convolution followed by the transformer network in order to extract the long range spatial-temporal dependencies between frames in the gesture sequences.



	0	1	2	3	4	5	6	7	8	9	10
0	116.0	0.0	6.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	124.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	13.0	0.0	96.0	9.0	0.0	0.0	0.0	1.0	1.0	0.0	4.0
3	18.0	0.0	5.0	97.0	0.0	0.0	0.0	0.0	0.0	2.0	2.0
4	0.0	0.0	0.0	2.0	114.0	4.0	0.0	1.0	0.0	2.0	1.0
5	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0	0.0	0.0
7	5.0	0.0	0.0	0.0	0.0	0.0	0.0	113.0	0.0	1.0	5.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	124.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	123.0	0.0
10	8.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	114.0

Fig. 6: Confusion matrix for the proposed 3DCNNTENet based on a 50:50 training and testing strategy.

Therefore, the classification accuracy as well as the convergence of our network is better. This justifies the combination of Res3DCNN and transformer network for capturing more discriminative spatial-temporal features for gesture recognition. We also notice that the TENet results in better average classification accuracy than the RNN-Shallow network [6].

5 Conclusion

In this paper, we presented a residual 3D-CNN transformer network (Res3DTENet) for gesture recognition. We validated the performance of the proposed network on publicly available Soli hand gesture data set. The proposed network performs better than the state-of-the-art networks presented in [6] and [2] as well as three variants of the network. The proposed model achieved gesture recognition accuracy of 96.99% and 92.25% based on a 50:50 data set split and cross validation strategies, respectively. In the future, we will explore unseen gesture data and diversity of subjects for evaluating performance of the proposed network.

6 Acknowledgement

This research was partially supported by the Semiconductor Research Corporation (SRC) Grant 2019-IR-2924.

References

1. 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)
2. 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**, 137122–137135 (2019)
3. Chen, K.S.: Principles of Synthetic Aperture Radar Imaging: A System Simulation Approach, vol. 2. CRC Press (2016)
4. Choi, J.W., Ryu, S.J., Kim, J.H.: Short-Range Radar Based Real-Time Hand Gesture Recognition using LSTM Encoder. vol. 7, pp. 33610 – 33618 (2019)
5. Hazra, S., Santra, A.: Short-Range Radar-Based Gesture Recognition System Using 3D CNN With Triplet Loss. vol. 7, pp. 125623 – 125633 (2019)
6. Lien, J., Gillian, N., Karagozler, M.E., Amihood, P., Schwesig, C., Olson, E., Raja, H., Poupyrev, I.: Soli: Ubiquitous Gesture Sensing With millimeter Wave Radar. vol. 35, pp. 1 – 19 (2016)
7. Park, J., Cho, S.H.: Ir-uwb radar sensor for human gesture recognition by using machine learning. In: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). pp. 1246–1249. IEEE (2016)
8. Sakamoto, T., Gao, X., Yavari, E., Rahman, A., Boric-Lubecke, O., Lubecke, V.M.: Radar-Based Hand Gesture Recognition Using IQ Echo Plot and Convolutional Neural Network. In: IEEE Conference on Antenna Measurements & Applications (CAMA). pp. 393 – 397 (2017)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
10. 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(UIST). pp. 851 – 860 (2016)
11. Wang, Y., Wang, S., Zhou, M., Jiang, Q., Tian, Z.: Ts-i3d based hand gesture recognition method with radar sensor. IEEE Access **7**, 22902–22913 (2019)
12. Xia, Z., Luomei, Y., Zhou, C., Xu, F.: Multidimensional feature representation and learning for robust hand-gesture recognition on commercial millimeter-wave radar. IEEE Transactions on Geoscience and Remote Sensing (2020)
13. Zhang, J., Tao, J., Shi, Z.: Doppler-Radar Based Hand Gesture Recognition System using Convolutional Neural Networks. In: International Conference in Communications, Signal Processing, and Systems. pp. 1096 – 1113 (2017)
14. Zhang, Z., Tian, Z., Zhang, Y., Zhou, M., Wang, B.: u-deephand: Fmcw radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network. IEEE Sensors Journal **19**(16), 6811–6821 (2019)
15. Zhang, Z., Tian, Z., Zhou, M.: Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. IEEE Sensors Journal **18**(8), 3278–3289 (2018)