FMCW Radar-Based Hand Gesture Recognition using Spatiotemporal Deformable and Context-Aware Convolutional 5D Feature Representation

Dong, Xichao; Zhao, Zewei; Wang, Yupei; Zeng, Tao; Wang, Jianping; Sui, Yi

DOI
10.1109/TGRS.2021.3122332

Publication date
2022

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Geoscience and Remote Sensing

FMCW Radar-Based Hand Gesture Recognition using Spatiotemporal Deformable and Context-Aware Convolutional 5D Feature Representation

Xichao Dong, Zewei Zhao, Yupei Wang, Tao Zeng, Jianping Wang and Yi Sui

Abstract—Recently, frequency-modulated continuous wave (FMCW) radar-based hand gesture recognition using deep learning has achieved favorable performance. However, many existing methods use extracted features separately, i.e., using one of the range, Doppler, azimuth or elevation angle information, or a combination of any two, to train convolutional neural networks (CNNs), which ignore the interrelation among the 5D time-varying-range-Doppler-azimuth-elevation feature space. Although there have been methods using the 5D information, their mining of the interrelation among the 5D feature space is not sufficient, and there's still room for improvements. This paper proposes a new processing scheme of hand gesture recognition based on 5D feature cubes which are jointly encoded by a 3D fast Fourier transform (3D-FFT) based method. Then a CNN is proposed by building two novel blocks, i.e., spatiotemporal deformable convolution (STDC) block and adaptive spatiotemporal context-aware convolution (ASTCAC) block. Concretely, STDC is designed to cope with hand gestures' large spatiotemporal geometric transformations in the 5D feature space. Moreover, ASTCAC is designed for modeling long-distance global relationships, e.g., relationships between pixels of the feature at upper left corner and lower right corner, and exploring the global spatiotemporal context, in order to enhance the target feature representation and suppress interference. Finally, our presented method is verified on a large radar dataset including 19760 sets of 16 common hand gestures, collected by 19 subjects. Our method obtains a recognition rate of 99.53% on validation dataset, and that of 97.22% on test dataset, which is significantly better than state-of-the-art methods.

***Index Terms*—Frequency-modulated continuous wave (FMCW) radar, hand gesture recognition, spatiotemporal deformable convolution, spatiotemporal context modeling.**

I. Introduction

HAND GESTURE recognition (HGR) has important application value in human-machine interaction [1]-[3], e.g., it can be used in sign language recognition [4], home automation [5] driving automation [6] and many other scenarios.

Researchers can achieve HGR based on visual equipment or other sensors [7]. Visual-equipment-based methods need to acquire hand gestures' images or videos first. However, visual-equipment-based methods are sensitive to light conditions. Sensor-based approaches require the use of sensors, such as WiFi-based sensors, electromyography (EMG), to measure hand gestures' accelerations, positions or velocities. However, WiFi-based methods are susceptible to interference because their waveforms are specially designed according to the purpose of communication [8]. EMG-based methods [9] could only work under contact conditions, which is not convenience and may put the user at risk of exposure to bacteria and virus.

Because of the advantages such as small size of antennas and the ability of working under non-contact and non-light conditions, HGR solutions based on Frequency-modulated continuous wave (FMCW) radars, such as Google's Soli [10], are promising, and have aroused widespread interests in the consumer electronics industry and the microwave communities.

FMCW radar-based HGR methods usually have two key steps: (1) beat signal pre-processing is used to process the complex raw radar data stream for presentation as input of feature extract models, (2) multiple parallel processing feature extraction architectures are used to extract the information separately, and feature-level or decision-level fusion is employed for hand gesture prediction and classification.

However, in terms of beat signal pre-processing, many existing methods extract features separately, such as range, Doppler, azimuth and elevation information, or a combination of any two [11]-[20], and these methods ignore one or several dimensions of information in the time-varying 5D feature space. Although there have been methods such as [21]-[23] which explore using the 5D feature representation, their mining of the

This work was supported by the China Postdoctoral Science Foundation under Grant 2020M670162, and also funded in part by the National Natural Science Foundation of China under Grant Nos. 61960206009, Distinguished Young Scholars of Chongqing (Grant No. cstc2020jcyj-jqX0008) and the Special Fund for Research on National Major Research Instruments (NSFC Grant Nos. 61827901, 31727901).

Xichao Dong, Zewei Zhao, YuPei Wang*, Tao Zeng and Yi Sui are with the School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China, and also with The Key Laboratory of Electronic and Information Technology in Satellite Navigation (Beijing Institute of Technology), Ministry of Education, Beijing 100081, China. (Corresponding author: Yupei Wang, email:wangyupei2019@outlook.com).

Xichao Dong is also with Chongqing Key Laboratory of Novel Civilian Radar and Beijing Institute of Technology Chongqing Innovation Center, Chongqing, 401120, China.

Jianping Wang is with the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology, Delft, 2628CD, the Netherlands.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

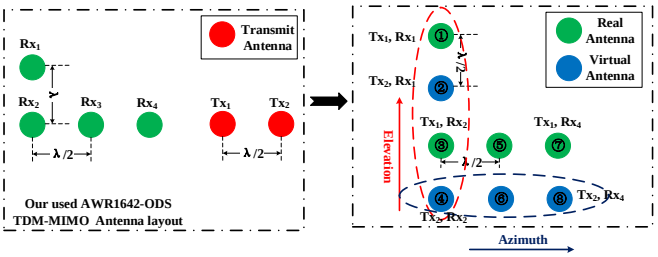


Fig. 1. Our used antenna layout (left) and the equivalent virtual array (right). It's an L-shaped array, and has 2 transmit antennas (Tx1 and Tx2) and 4 receive antennas (Rx1~Rx4). Under TDM-MIMO mode, a virtual array is generated. In the virtual array, the vertically arranged Tx-Rx pairs (surrounded by the red dashed line) are used for elevation angle estimation, while horizontally arranged Tx-Rx pairs (surrounded by the blue dashed line) are used for azimuth angle estimation. Data collected by 8 virtual array elements are represented by ①-⑧.

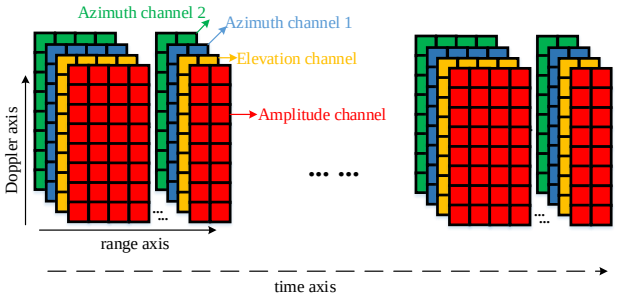


Fig. 2. Diagram of our used 5D feature cubes.

interrelation among the 5D feature space is not sufficient, and there is still room for improvement. For example, [22][23] use a multi feature encoder to encode the selected the first K points' 5D features of the gesture which have the greatest magnitudes in the range-Doppler spectrograms. However, the fixed manually selected K points are susceptible to dynamic interference and difficult to adapt to complex scenes.

In terms of feature extraction and classification, traditional methods, such as Hidden Markov Models (HMM) [25], Support Vector Machines (SVM) [26], can only classify a few simple gestures. More complex gesture categories can be recognized via convolutional neural network (CNN) based methods [22]-[24] or Long Short-Term Memory (LSTM) based methods [27][28] from thousands of data samples. However, using separately extracted range, Doppler and angle information to train CNNs, these methods are usually 2D-CNN based, and they cannot take a 5D feature representation as input. Even inputs in [22][23] consider the 5D feature, their used simple CNN networks have difficult in effectively extracting the key information that characterizes different gestures.

This paper proposes a new processing scheme of HGR based on 5D time-varying-range-Doppler-azimuth-elevation feature cubes which are jointly encoded by using a 3D fast Fourier transform (3D-FFT) based method. And a modified CNN is proposed by building in two novel blocks. We summarize main contributions of this paper as follows.

1) A 3D-FFT based beat signal pre-processing method is introduced for jointly encoding range, Doppler, azimuth and elevation angle information into the time-varying 5D feature space.

2) A spatiotemporal deformable convolution (STDC) block is introduced to improve the ability of recognition network to model spatiotemporal geometric transformations in the 5D feature space by learning extra offsets, drawn inspiration form Dai et al. [29] and Ying et al. [42].

3) An adaptive spatiotemporal context-aware convolution (ASTCAC) block is proposed to improve the ability of recognition network to capture both global and local contextual information.

The rest of this article is organized as follows. First, in Section II we review the related works. Then in Section III we introduce the STDC block and the ASTCAC block, respectively. Moreover, in Section IV we introduce the experimental settings, and analyze the performance of the proposed STDC and ASTCAC block. Finally, in Section V, we give our conclusion and discuss some future works.

II. Related Works

A. FMCW Radar Beat Signal Pre-Processing

Generally, FMCW radar beat signal pre-processing contains estimation of range, Doppler and angle information, and suppression of static and dynamic interference. Typical process is summarized as follows.

First, FMCW radar transmit chirp signals and the range-Doppler maps can be obtained via range-FFT and Doppler-FFT (2D-FFT) for each receiver [48].Then CA-CFAR detectors are used for target detection. After target detection, target's range and radial velocity information can be measured [48].

Fig. 1 show our used antenna layout and the equivalent virtual array under TDM-MIMO mode. The vertically arranged transmit-receive antenna pairs are used for elevation angle estimation, and the horizontally arranged transmit-receive antenna pairs are used for azimuth angle estimation. Azimuth (elevation) information of the target is estimated by azimuth-FFT (elevation-FFT) or other super-resolution algorithms, such as multiple signal classification (MUSIC) algorithm [31].

Note that during the switching time of different transmit antennas, the amount of phase change owing to moving hand gestures' Doppler frequency are coupled to each receive antenna, resulting in a defocusing effect of the spectrum [32]. This phase change would have effect on the elevation angle estimation in our case, and we carry on phase compensation before elevation angle estimation.

In this paper, we choose azimuth-FFT (elevation-FFT) method over other azimuth (elevation) angle estimation methods mainly because, (1) despite of other super-resolution estimation algorithms perform better angle resolution than azimuth-FFT (elevation-FFT) method, however, the large amount of calculation imply that they may be unable to achieve real-time processing, (2) it is difficult to perform 2D-FFT based range-Doppler estimation in parallel with MUSIC-based angle estimation for jointly encoding range, Doppler, azimuth and elevation angle information into the time-varying 5D feature space, hence the interrelation information among range-Doppler-azimuth-elevation domains is neglected, this may degrades the performance of HGR.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

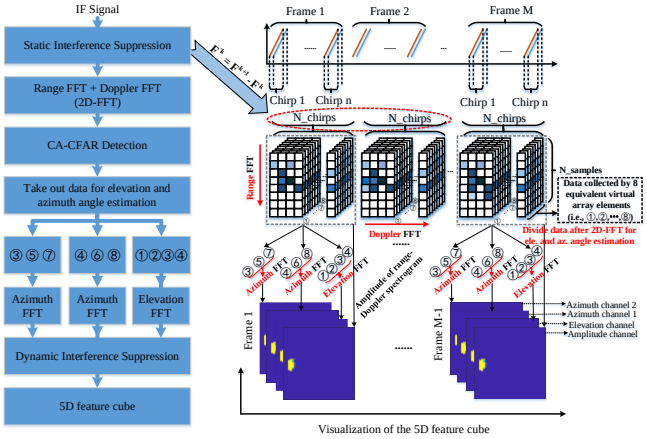


Fig. 3. Pipeline of our signal processing method to generate the 5D feature cubes. ①-⑧ indicate the data collected by 8 array elements (refer Fig. 1 to query which array element each number represents).

To suppress static interference such as walls, we use an inter frame difference method, which can be described as:

$$F^{'k} = F^{k+1} - F^{k} \tag{1}$$

where F^{k} represents the k_{th} radar frame, $F^{'k}$ represents the k_{th} frame after inter frame differential operation. There is still dynamic interference, such as moving bodies or arms, after suppressing static interference. Generally, when a subject is making a gesture, distance between bodies and radar is larger than that of hands, while velocity of bodies is smaller than that of hands. Based on this prior information, we can suppress the dynamic interference by filtering out the scatters with larger distance and velocities.

As input of our CNN-based feature extraction models, the data structure of our used 5D feature cubes are described as follows: as shown in Fig. 2, there are 4 channels in the feature cube of each frame channel (time channel), namely the amplitude channel, elevation channel, the first azimuth channel and the second azimuth channel, representing the amplitude of the range-Doppler spectrograms, the estimated elevation information, the estimated azimuth information from the upper horizontally arranged transmit-receive antenna pairs, and that from the lower horizontally arranged transmit-receive antenna pairs. The x-axis of each channel data is the range axis and the y-axis of each channel is the Doppler axis. That is to say, if we visualize the 4 channels of each frame's feature cube data, we get range-Doppler maps (RDMs), range-Doppler-elevation maps (RDEMs) and range-Doppler-azimuth-maps (RDAMs) accordingly. The feature cubes of each frame are concatenated along the time axis to generate the finally 5D feature cubes, which can be considered as a tensor type with a shape of T×H×W×C, where T is the time dimension, i.e. number of frames, H and W are the dimensions of range and Doppler domain respectively, C is the number of channel, i.e. 4 in our case.

Fig. 3 shows the pipeline of the proposed FMCW radar beat signal pre-processing method.

B. 3D CNNs

Different with 2D spatial convolutional kernels used in 2D CNNs, 3D convolutional kernels used in 3D CNNs [33][35] are extended to three dimensions, and the added dimension is the time dimension. 3D convolutions use a three-dimensional sliding window to scan the spatiotemporal data at the same time, which can capture unified spatiotemporal characteristics. Hence 3D CNNs are suitable for video-related tasks, such as action recognition. However, training 3D CNNs is more difficult than 2D CNNs due to the sharply increased parameters. Until the emergence of large-scale video datasets [50][51], 3D CNN based methods outperform 2D CNN based methods gradually. Ji et al. first proposed C3D network [33] for human action recognition. Later, Hara et al. proposed Res3D network [36] based on ResNet network [37]. Carreira et al. propose I3D network [34] based on the Inception network. Fig. 4 (a) and (b) show an illustration of 2D and 3D convolution.

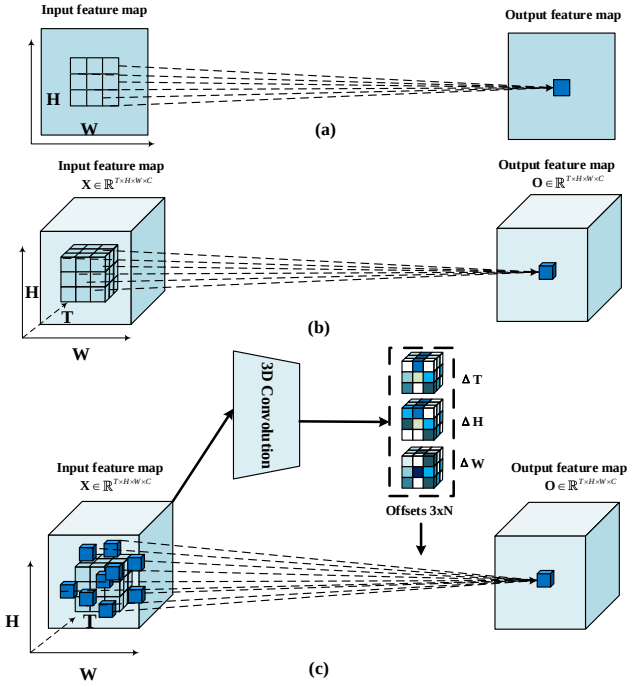


Fig. 4. (a) 2D convolution, (b) 3D convolution and (c) spatiotemporal deformable convolution (STDC). Note that (c) is modified from [42].

III. Hand Gesture Recognition Network

We can represent the 5D feature cubes as a tensor $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$. We argue that it is more suitable to model spatiotemporal feature extraction based on 3D CNNs than 2D CNNs when using the 5D feature cubes as input. Although 3D CNNs are usually computationally expensive, we can exploit some efficient variations, such as replacing 3D convolutions with separable convolutions or S3D network [38], to tradeoff model complexity with speed and accuracy.

In this paper, we choose S3D-like network as our backbone, and we first perform some adaptive modifications to the original S3D network to make it suitable for 5D feature cube inputs. Then we elaborate the STDC block and the ASTCAC block to further improve the accuracy of HGR.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

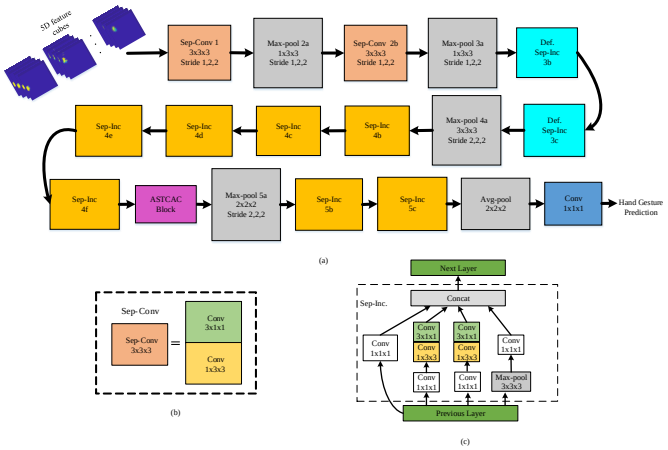


Fig. 5. (a) Architecture of our HGR network. Backbone is S3D [38]. Compared with the original S3D architecture, we replace Sep-Inc 3b and Sep-Inc 3c layers with STDC blocks, and plug the ASTCAC block after Sep-Inc 4f. (b) Our used temporal separable convolution block (Sep-Conv). (c) Our used 3D temporal separable inception block (Sep-Inc).

A. *Modified S3D network*

In original S3D network, 3×3×3 convolutions are replaced with one 3×1×1 convolution in temporal domain and one 1×3×3 convolution in spatial domain. This block is called temporal separable convolution (Sep-Conv) block as shown in Fig. 5 (b). What's more, S3D uses a 3D temporal separable inception (Sep-Inc) block as basic block, which is shown in Fig. 5 (c). These structures are proved to tradeoff model accuracy with complexity better in action recognition tasks compared with conventional 3D CNNs [38].

The original S3D network has two Sep-Conv blocks, nine Sep-Inc blocks, four max pooling layers, one average pooling layer and one 1×1×1 convolution layer.

However, in 5D feature cubes, the input length along the spatial dimensions is much longer than those along the time dimension. When using 5D feature cubes as input, retaining origin S3D network configurations may result in premature down-sampling in the time dimension. Hence we have to modify these configurations. Concretely, strides in the spatial dimensions of the first two Sep-Conv layers and max pool layers are both modified from 1 to 2, while strides in the time dimension remain unchanged at 1, as illustrated in Fig. 5 (a).

B. *STDC Block*

Fig. 6 shows two sets of RDAMs (i.e. visualizations of the 3rd channel of the 5D data cubes) of push right gestures made by different subjects. In order to show the time-varying features contained in multi-frame RDAMs in one figure, we start from the first frame and take 5 of the 31 frames at equal intervals, and draw these 5 frames in the same figure. The curve formed by the red arrows indicates the trajectory of the palm.

It can be seen that the trajectories of the gestures in Fig. 6 (a) and (b) show similar characteristics, the range first decreases and then increases, and the radial velocity changes from negative to positive. This can be seen as the common feature of push right (PS-R) gesture reflected by RDAMs. However, the PS-R gesture in Fig. 6 (a) has a larger range of range-Doppler

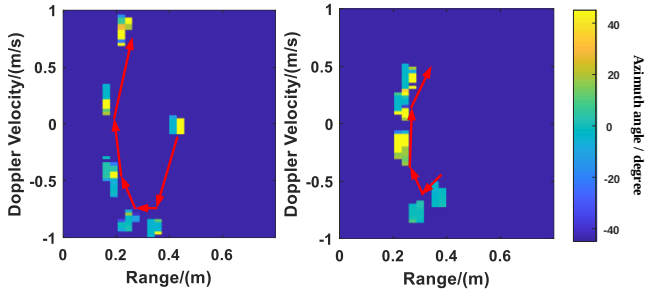


Fig. 6. Two sets of RDAMs of push right (PS-R) made by different subjects.

change, while the range of azimuth angle change in Fig. 6 (b) is larger. This indicates that the large spatiotemporal deformation during gesture movement. It is related to the subject's personal habit of making gestures, e.g., the size of the palm of the subject is different, the speed of the gesture is different, and the angle of the gesture relative to the radar line of sight is different. Similar with PS-R gestures, other types of gestures also have in-class commonalities and differences reflected in RDAMs.

However, it's tough for conventional 3D convolutions to handle abovementioned huge spatiotemporal deformations, because their sampling process on the feature map is usually performed on regular, rectangular sampling grid. To this end, we try to augment conventional 3D convolutions with learnable 3D sampling offsets to model complex geometric transformations, inspired by Dai et al.[29] and Ying et al [42]. Different from [42] which applies deformable convolution in low-level video super resolution task and only performs kernel deformation in spatial dimension, we use STDC block for a high-level, FMCW radar-based HGR task, and for radar data. What's more, we perform kernel deformation in both spatial and temporal dimensions.

Next we introduce STDC in detail. Suppose there are n_k 3D convolution kernels $\mathbf{F} \in \mathbb{R}^{n_k \times t_k \times h_k \times w_k}$, with the kernel size of $t_k \times h_k \times w_k$. Note that for simplifying the presentation, we omit the channel dimension with reference to [30], i.e. for $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$ we assume the channel dimension C=1 here. A 3D convolution operation which takes $\mathbf{X}=\{\mathbf{x}_{t,h,w}\}$ as input and outputs $\mathbf{O}=\{\mathbf{o}_{t,h,w}\}$ can be written as:

$$\mathbf{O} = \mathbf{F} \odot \mathbf{X}, \text{ where}$$

$$\mathbf{o}_{t_0,h_0,w_0} = \left[q^1_{t_0,h_0,w_0}, q^2_{t_0,h_0,w_0}, ..., q^{n_k}_{t_0,h_0,w_0} \right]^T, \qquad (2)$$

$$q^n_{t_0,h_0,w_0} = \sum_{t,h,w} \mathbf{F}^n_{t,h,w} \cdot \mathbf{X}^{t_0 h_0 w_0}_{t,h,w}$$

where t_0, h_0, w_0 represents the start spatiotemporal position of the 3D convolution, $q^n_{t_0,h_0,w_0}$ denotes the output of the n^{th} 3D convolution kernel at position t_0, h_0, w_0. Since we have n_k 3D convolution kernels, the output feature map $\mathbf{O}=\{\mathbf{o}_{t,h,w}\}$ have n_k channels.

From (2) we can know that the conventional 3D convolutional operation is restricted to fixed spatiotemporal sampling grid when sampling input feature cubes, which degrades the ability of feature representation.

To better perceive the long-distance and diverse-change motion characteristics in 5D feature cubes, the proposed STDC block learns offsets $\{ (\Delta T_t, \Delta H_h, \Delta W_w) | t,h,w = 1,...,k \}$ (k is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

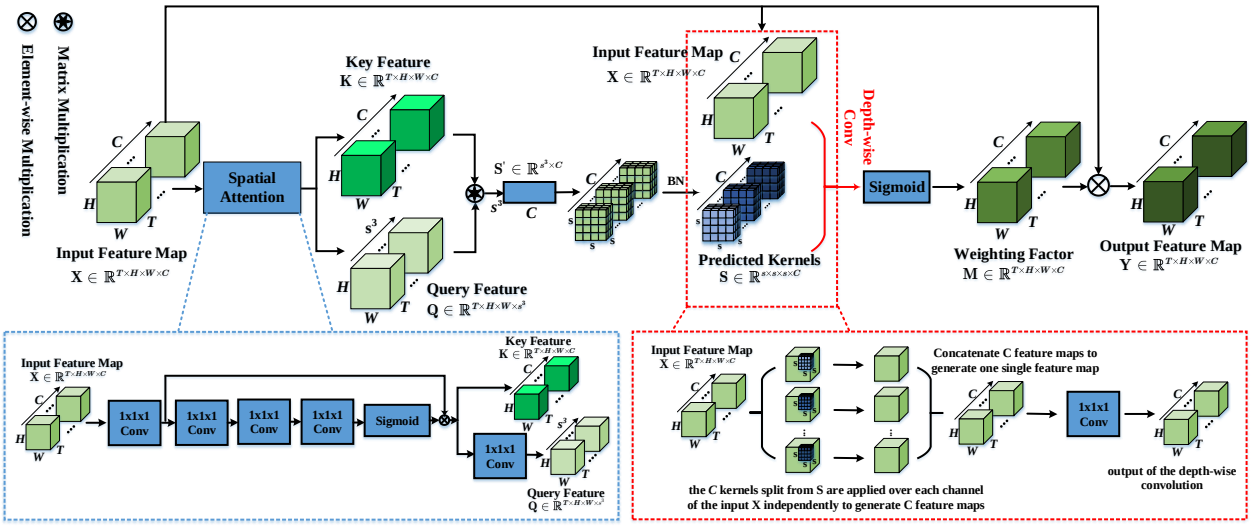


Fig. 7. Our proposed adaptive spatial temporal context aware convolution (ASTCAC) kernels. We first use a spatial attention sub-module (the blue dashed box) to generate the key feature map $\mathbf{K}$ and the query feature map $\mathbf{Q}$. Then we use a matrix multiplication method to generate the predicted ASTCAC kernels. After that, to generate the weighting map $\mathbf{M}$, the predicted kernels are used to convolved with $\mathbf{X}$ by a depth-wise convolution (the red dashed box). Finally, by element-wise multiplication (Hadamard product) between $\mathbf{M}$ and $\mathbf{X}$, we get the output feature map $\mathbf{Y}$.

kernel size) to deform the standard sampling grid instead of using regular sampling grids. STDC can be described as:

$$\mathbf{O} = \mathbf{F} \odot \mathbf{X}, \text{ where}$$

$$\mathbf{o}_{t_0,h_0,w_0} = \left[q^1_{t_0,h_0,w_0}, q^2_{t_0,h_0,w_0}, ..., q^{n_k}_{t_0,h_0,w_0} \right]^T, \qquad (3)$$
$$q^n_{t_0,h_0,w_0} = \sum_{t,h,w} \mathbf{F}^n_{t,h,w} \cdot \mathbf{X}^{t_0 h_0 w_0}_{(t+\Delta T, h+\Delta H, w+\Delta W)}$$

Compared with conventional convolution in (2), corresponding convolution field of STDC has been modified form $\mathbf{X}^{t_0 h_0 w_0}_{t,h,w}$ to $\mathbf{X}^{t_0 h_0 w_0}_{(t+\Delta T, h+\Delta H, w+\Delta W)}$.

Take a STDC block with $3\times3\times3$ kernel size as an example, we explain the calculation process of the STDC block as follows. First, a conventional $3\times3\times3$ convolution, as shown in the upper branch of Fig. 4 (c), is used to obtain the 3D spatiotemporal offsets, taking $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$ as input. It should be noted that the learned 3D spatiotemporal offsets have 3*N* sets of channels, which represent the deformation of STDC's sliding windows among spatiotemporal domains [42]. Then the deformation of the conventional sampling grid is guided by these learned offsets. Finally, we obtain outputs with these deformable sampling grid following (3). Specially, we use a trilinear interpolation method [44] to generate exact values of the offsets, since they are usually fractional.

C. ASTCAC Block

In addition to the obvious spatiotemporal deformations of gestures, local differences of different gesture types in the 5D feature cubes are not obvious and there are still clutter and interference after signal processing. To enhance the target feature representation and suppress interference, we need to mine global spatiotemporal context information. However, conventional convolutions can only extract local context and cannot model long-distance global relationships, e.g., relationships between upper left corner and lower right corner pixels of the feature.

Recently, previous channel-wise feature re-weighting methods (such as SE-Net [45]) are proved efficient to perceive global context for re-weighting feature channels. However, previous methods usually model the long-distance global relations with global-consistent feature re-weighting vector. Different with previous methods, we propose the ASTCAC block, a spatiotemporally-varying feature weighting factors based method, to mine higher level spatiotemporal contextual information.

We show the detailed ASTCAC architecture in Fig. 7. Instead of using the fully-connected-layers (FC-layers) to predict all ASTCAC kernel parameters as previous dynamic kernels do [46], we generate the ASTCAC kernel parameters via matrix multiplication to reduce computation. Next is the specific implementation of the ASTCAC block.

Assume the size of kernel is $s\times s\times s$. First, we transform input feature $\mathbf{X}$ into the key feature map $\mathbf{K} \in \mathbb{R}^{T\times H\times W\times C}$ and the query feature map $\mathbf{Q} \in \mathbb{R}^{T\times H\times W\times s^3}$ via a spatial attention block (SAB). In detail, first, by transmission $\mathbf{T}_E$, $\mathbf{X}$ is used to generate the spatial-attention feature map $\mathbf{A}$. $\mathbf{T}_E$ is implemented with non-linear project function and Sigmoid activation function. Then we perform Hadamard product between $\mathbf{A}$ and $\mathbf{X}$ to get $\mathbf{E}$. Let key feature map $\mathbf{K} = \mathbf{E}$. At the same time, by another transformation $\mathbf{T}_Q$ which is implemented by independent non-linear project function, we get the query feature map $\mathbf{Q}$ which can capture spatiotemporal distributions of $\mathbf{K}$. Particularly, the non-linear project functions can be implemented with $1\times1\times1$ convolutions. The pipeline of SAB is shown in Fig. 7.

After generating $\mathbf{K}$ and $\mathbf{Q}$, we reshape them to 2D vectors $\mathbf{K}' \in \mathbb{R}^{(T\times H\times W)\times C}$ and $\mathbf{Q}' \in \mathbb{R}^{(T\times H\times W)\times s^3}$, respectively. We argue that each column of $\mathbf{K}'$ represents one of the C-dimensional characteristics of $\mathbf{X}$, and each column of $\mathbf{Q}'$ captures one of the s^3-dimensional spatiotemporal features.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

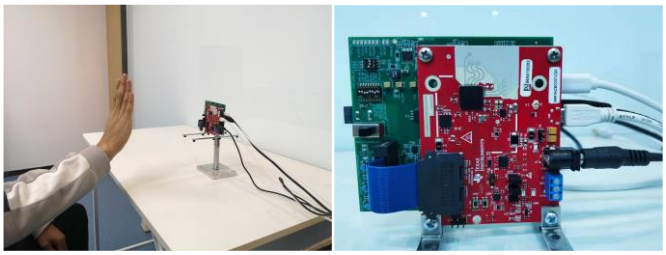


Fig. 8. Experimental environment and equipment.

Then, for extracting the interrelation between each column of $\mathbf{Q}^{'}$ and $\mathbf{K}^{'}$ among the overall $T \times H \times W$ spatiotemporal positions, a dot product is performed:

$$\mathbf{S}^{'}(i,j) = \sum_{q=1}^{(T\times H\times W)} \mathbf{Q}^{'}(q,i) \times \mathbf{K}^{'}(q,j) \tag{4}$$

where $i = 1,2,...,s^3$, $j = 1,2,...,C$. Because we have s^3 query vectors, we can capture s^3 features of the overall spatiotemporal distributions of $\mathbf{K}^{'}$ which has C feature channels. (4) can also be rewritten as a form of matrix multiplication [47]:

$$\mathbf{S}^{'} = \mathbf{Q}^{'T}\mathbf{K}^{'} \tag{5}$$

where $\mathbf{S}^{'} \in \mathbb{R}^{s^3 \times C}$, $\mathbf{Q}^{'T}$ denotes the transpose matrix of $\mathbf{Q}^{'}$.

After that, to obtain the final predicted ASTCAC kernels, we reshape $\mathbf{S}^{'} \in \mathbb{R}^{s^3 \times C}$ into $\mathbf{S} \in \mathbb{R}^{s \times s \times s \times C}$, and a batch normalization layer is used to modulate $\mathbf{S}$. $\mathbf{S}$ is the predicted ASTCAC kernels. We can use $\mathbf{S}$ to generate the global spatiotemporally-varying weighting factor $\mathbf{M} \in \mathbb{R}^{T \times H \times W \times C}$ for all $T \times H \times W$ locations.

Importantly, when generating $\mathbf{M}$, a depth-wise convolution is operated between $\mathbf{S}$ and $\mathbf{X}$ to ensure each channel of $\mathbf{S}$ can independently modulate the corresponding channel of $\mathbf{X}$, and save computation at the same time. The depth-wise convolution is first introduced by [52] to reduce computation. As shown in the red dashed box of Fig. 7, the depth-wise convolution works as follows, first $\mathbf{S} \in \mathbb{R}^{s \times s \times s \times C}$ is split into C kernels, and each kernel has a dimension of $s \times s \times s$. Subsequently, these C kernels are applied over each channel of the input $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$ independently to obtain an intermediate feature map, followed by a $1 \times 1 \times 1$ convolution to project the intermediate feature map's channels onto a new channel space and get the output of the depth-wise convolution. The output of the depth-wise convolution is further passed through one Sigmoid activation function to obtain $\mathbf{M} \in \mathbb{R}^{T \times H \times W \times C}$. Finally, $\mathbf{M}$ and $\mathbf{X}$ are multiplied by elements to get the output feature map $\mathbf{Y}$.

IV. EXPERIMENTS AND ANALYSES

A. Experimental Platform

As shown in Fig. 8, we use the Texas Instruments (TI) single chip AWR1642BOOST-ODS radar system to collect hand gesture data [49], which is equipped with two transmit antennas and four receive antennas. Descriptions about the antenna layouts are already shown in Fig. 1. We set the ADC sampling frequency to 10MHz under complex 2x sampling module. Under this configuration, we obtain complex raw radar data. We list detail radar configuration parameters in Table I.

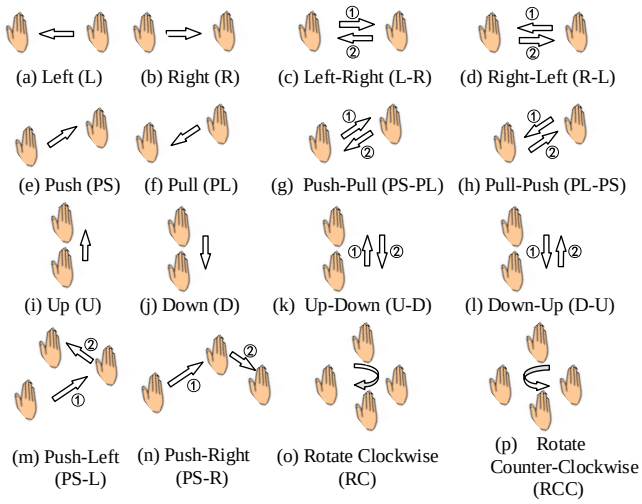


Fig. 9. Our used 16 kinds of gestures.

TABLE I

RADAR SYSTEM PARAMETERS FOR HAND GESTURE RECOGNITION

Parameters	Values
Number of transmit antennas N_{Tx}	2
Number of transmit antennas N_{Rx}	4
Duty cycle	44.2%
Chirp Bandwidth (B)	4 GHz
Frequency modulation rate	105.22 MHz/us
Sampling mode	complex 2x
Time duration of the chirp (T_S)	38 us
Actual chirp duration (Tc)	138 us
Time duration of per frame (T_f)	40 ms
Number of chirps per frame (N_{chirp})	64
Number of sampling points per chirp ($N_{samples}$)	256
Total number of frames (N_{frame})	32

A spacious indoor environment, as shown in Fig. 8, is used for data collection. The subject sits directly in front of the radar and collects data according to the prescribed gestures. After collecting data, we use MATLAB to implement the 3D-FFT based FMCW radar beat signal pre-processing algorithm to generate 5D feature cubes dataset. Then we use the PyTorch frame work to build the HGR network.

B. Dataset

We organize 19 graduate students for basic skill training on radar systems and radar signal processing. The training included an experiment operation and data processing based on this platform. With their help, we built a dataset with rich diversity. 16 kinds of hand gestures containing both azimuth and elevation movements are used, as shown in Fig. 9, and the total number of realizations is (16 classes) × (19 subjects) × (65 times), namely 19760.

The collected dataset is divided into train dataset, validation dataset and test dataset according to a certain proportion of 4:1:3. The train dataset and validation dataset are used to train

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

TABLE II

RECOGNITION ACCURACY (%) COMPARISON OF DIFFERENT METHODS ON VALIDATION DATASET

Methods	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	Avg.
2D-CNN (RTM+DTM)	92.47	91.95	91.00	92.67	97.33	93.19	92.34	90.23	94.42	90.35	96.48	91.44	88.74	97.38	90.93	96.10	92.93
2D-CNN (RTM+DTM +ATM+ETM)[20]	99.06	99.17	99.35	99.22	99.3	99.28	97.77	99.17	99.17	99.26	99.39	99.96	94.18	99.17	99.46	99.15	98.87
3D-CNN (MPCA)[22]	95.21	95.18	95.99	95.08	97.87	95.47	96.78	97.14	97.64	94.56	99.02	94.46	89.00	94.96	93.68	95.16	95.45
3D-CNN +LSTM[23]	98.38	98.38	98.44	99.54	99.14	100.00	99.54	97.83	98.48	97.87	99.57	99.02	92.38	97.34	98.36	95.49	98.11
2D-CNN (Multi-feature encoder)[23]	98.35	98.35	98.40	99.45	98.92	100.0	99.45	97.80	98.38	97.85	99.45	98.95	92.34	97.27	98.28	95.40	98.04
S3D (5D feature cubes)	98.51	99.20	99.70	99.55	99.42	99.14	97.76	99.55	99.50	99.69	99.46	98.80	93.65	99.47	99.04	98.50	98.80
S3D +STDC **(ours)**	99.51	99.70	99.27	99.42	99.74	99.62	98.04	99.47	99.60	99.90	99.86	99.35	93.98	99.70	99.35	99.55	99.12
S3D +ASTCAC **(ours)**	100.0	99.58	98.88	99.39	99.87	97.27	97.27	99.32	99.92	99.65	98.88	99.71	93.68	99.48	99.94	99.79	99.01
S3D +STDC +ASTCAC **(ours)**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**98.11**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.51**	**100.0**	**100.0**	**100.0**	**99.53**

our HGR model, and the generated model is used to classify the gestures in test dataset after training.

C. Experimental Results and Analysis

We implement our models on a server which is equipped with 3NVIDIA TITAN RTX graphics card. We use Adam optimizer and cross entropy error function for back propagation. Our used batch-size is 32 and learning rate is 0.0001. What's more, we train each of the models for 100 epochs.

Fig. 10 (a) and (b) show visualizations of multi-frame RDAMs of a left and right gesture respectively, the azimuth angle of the right gesture increases over time, while that of the left gesture decreases over time. Fig. 10 (c) and (d) show multi-frame RDEMs of an up and down gesture respectively. The elevation angle of the up gesture increases over time, while that of the down gesture decreases over time. These angular changes can be considered as the feature enjoyed by specific kind of gestures, especially by those with obvious angular changes. This kind of information are learned by feature extraction models and are guided for classification.

In order to effectively evaluate the performance of the proposed STDC and ASTCAC blocks, we make the following ablation studies. We first train the modified S3D network on our built dataset. The recognition accuracy on validation dataset is 98.80%, the accuracy on test dataset is 95.33%, as listed in Table II and Table III, respectively.

Then, to analyze the rationality of the proposed STDC block, we plug STDC block in the modified S3D network. We argue that it is unreasonable to directly replace all conventional convolutional layers with STDC blocks, because too many STDC blocks will inevitably bring additional parameters, thereby affecting the convergence of the model [44]. In order to find out in which part of the network structure is most suitable to replace the conventional convolution with STDC blocks, we gradually replace the conventional convolution layers with STDC blocks from shallow layers to deep layers [44]. As listed in Table IV, when using more STDC blocks from shallow layer to deep layers gradually, the recognition accuracy is improved steadily, and we get the best accuracy when conventional convolutions in Sep-Inc 3b layer and 3c layer are replaced with STDCs. According to the experimental results, two STDC blocks are good enough for this architecture. We call this model S3D+STDC, which achieve an accuracy of 99.12% on validation dataset, 96.31% on test dataset, as shown in Table II and Table III respectively. Similarly, to demonstrate the effectiveness of the ASTCAC block, we gradually embed ASTCAC block after conventional convolutional layers on the basis of S3D. As listed in Table II and Table III, the best recognition accuracy (99.01% on validation dataset and 96.47% on test dataset) is obtained when we plug one ASTCAC module after Sep-Inc 4f layer. We called this architecture S3D+ASTCAC model.

Combination of these two complementary blocks can further improve the recognition accuracy. According to above analysis, we replace Sep-Inc 3b and Sep-Inc 3c layers with deformable versions, and plug one ASTCAC block after Sep-Inc 4f. We call this structure S3D+STDC+ASTCAC model. The recognition accuracy on test dataset is significantly improved from 95.33% to 97.22%, as listed in Table III. Fig. 11 illustrates the confusion matrix of the recognition result on the test dataset with this network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3122332, IEEE Transactions on Geoscience and Remote Sensing

TABLE III

RECOGNITION ACCURACY (%) COMPARISON OF DIFFERENT METHODS ON TEST DATASET

Methods	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	Avg.
2D-CNN (RTM+DTM)	92.64	80.11	73.83	75.86	81.80	81.24	78.68	98.29	95.36	91.74	93.77	87.08	63.94	58.98	65.12	62.80	80.08
2D-CNN (RTM+DTM +ATM+ETM)[20]	91.05	91.81	97.95	94.64	96.36	97.32	86.1	97.56	96.64	92.27	94.23	97.63	87.90	94.54	88.56	92.36	93.56
3D-CNN (MPCA)[22]	92.25	88.46	95.24	86.01	86.81	92.18	83.13	97.59	95.26	92.75	94.87	93.09	87.39	85.10	85.78	87.68	90.22
3DCNN+LSTM[23]	93.17	93.88	94.01	91.81	97.61	95.45	89.43	99.67	94.71	97.99	97.56	94.64	88.00	90.70	94.04	96.52	94.32
2D-CNN (Multi-feature encoder)[23]	95.28	97.13	97.73	89.54	96.20	92.71	89.65	98.03	96.31	94.43	95.17	97.17	91.04	94.67	94.64	95.89	94.72
S3D (5D feature cubes)	92.78	98.34	98.75	96.03	96.63	98.10	88.24	98.63	95.72	93.57	97.91	96.6	91.34	96.05	93.21	93.34	95.33
S3D +STDC **(ours)**	95.89	98.74	98.98	98.31	96.72	97.57	91.04	98.23	98.88	95.96	96.18	96.99	93.24	94.58	96.89	92.72	96.31
S3D +ASTCAC **(ours)**	95.76	98.69	99.41	97.55	96.81	97.60	89.68	98.55	99.59	95.43	97.65	97.64	93.18	95.83	95.95	94.16	96.47
S3D +STDC +ASTCAC **(ours)**	**96.74**	**98.91**	**99.46**	**98.38**	**97.80**	**98.91**	**91.26**	**98.90**	**100.0**	**96.20**	**98.35**	**98.35**	**94.54**	**96.31**	**97.13**	**94.24**	**97.22**

TABLE III

EFFECTS OF DIFFERENT POSITIONS FOR STDC BLOCK. THE 'POSITION' COLUMN IN THE TABLE REPRESENTS THE POSITION WHERE THE CONVENTIONAL CONVOLUTION LAYER IN S3D NETWORK IS REPLACED BY THE STDC BLOCK.

Position	Accuracy on test dataset(%)
None	95.33
Sep-Cov 2	96.00
Sep-Inc 3b	96.23
Sep-Inc 3b, Sep-Inc 3c	**96.31**
Sep-Inc 3b~3c, Sep-inc4b~4f	96.31

TABLE VI

EFFECTS OF DIFFERENT POSITIONS FOR ASTCAC BLOCK. THE 'POSITION' COLUMN IN THE TABLE INDICATES WHERE THE ASTCAC BLOCK IS INSERTED IN THE S3D NETWORK.

Position	Accuracy on test dataset(%)
None	95.33
Sep-Cov 2	96.08
Sep-Inc 3c	96.39
Sep-Inc 4f	**96.47**
Sep-Inc 3c, Sep-Inc 4f	96.47

We summarize the recognition accuracies on the validation dataset and test dataset of different network settings in Table II and Table III. Through the above quantitative analysis, it can be seen that STDC and ASTCAC can help to improve recognition accuracy, especially when we use them into a deeper layer.

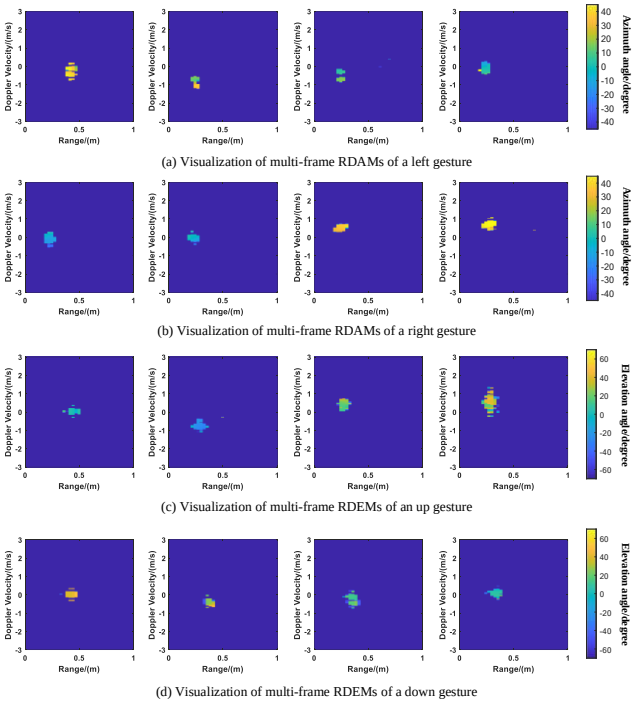
V. DISCUSSION AND CONCLUSION

In this paper, we try to improve existing radar-based HGR methods from two perspectives of radar signal processing and recognition network designing.

To evaluate our methods, we first collect a large dataset of 16 kinds of gestures containing both azimuth and elevation movements, and the total number of realizations is (16 classes) × (19 subjects) × (65 times), namely 19760. Compared with model performances evaluated on small datasets with just hundreds or thousands of realizations or collected by several subjects such as [8][22][23], our dataset is more challenging and results performed on our dataset are more convincing.

Since verification results on a large and independent test dataset can assess the model's generalization ability and its robustness, we focus on analyzing results on the test dataset. From experimental results we can know that, first, HGR methods with combination of range, Doppler, azimuth and elevation angle information as inputs (such as RTM+DTM+ATM+ETM or the 5D feature cubes, where RTM DTM, ATM and ETM represent range-time-maps, Doppler-time-maps, azimuth-time-maps and elevation-time-maps respectively, note that in [20] they only use RTM+DTM+ATM because their linear antenna array can't estimate azimuth and elevation angle at the same time, and here we use RTM+DTM+ATM+ETM for comparisons) outperform methods with single range, Doppler, azimuth or elevation information as inputs, or combination of any two (such as RTM+DTM). This shows that providing more dimensions of information is beneficial for radar based HGR, in line with the expected conclusion. Actually, for gestures with similar features in range-Doppler domain, it's necessary to introduce angular information for better recognition results.

Moreover, although there're already methods considering using the 5D feature representation, such as 3D-CNN (MPCA) [22], 3D-CNN+LSTM [23] and 2D-CNN (multi-feature encoder) [22]. However, our methods outperform all of these



(a) Visualization of multi-frame RDAMs of a left gesture

(b) Visualization of multi-frame RDAMs of a right gesture

(c) Visualization of multi-frame RDEMs of an up gesture

(d) Visualization of multi-frame RDEMs of a down gesture

Fig. 10. Examples of visualization of multi-frame RDAMs and RDEMs.

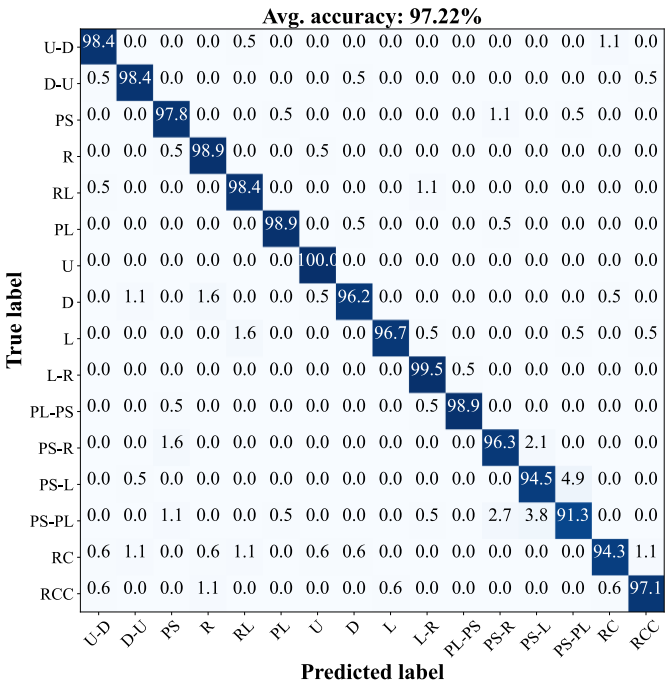


Fig. 11. Confusion matrix of recognition results of S3D+STDC network on test dataset.

methods. The 3D-CNN (MPCA) only gets an accuracy of 90.22%, and this may be explained by that this method suffers from the high dimensionality of the extracted 5D feature tensor [22]. The 2D-CNN (multi-feature encoder) method gets an accuracy of 94.72% on the test dataset, 4.5% higher than that of 3D-CNN (MPCA), a little (0.4%) higher than that of 3D-CNN+LSTM, comparable with that of S3D (5D feature cubes), but still lower than that of our methods, such as S3D+STDC, S3D+ASTCAC and S3D+STDC+ASTCAC. Note that the multi-feature encoder used in [22] and [23] directly extract the first K points' range, Doppler, azimuth and elevation information with the greatest amplitudes in the incoherently integrated range-Doppler spectrogram of different receive antennas to represent the features of the gesture, while the selected K points may not only encode features of gesture targets but also that of dynamic interference. Although their multi-feature encoder reduces the dimensionality of the hand gestures' features and reduces the amount of calculation, this has an impact on recognition accuracy.

In conclusion, owing to effective 3D-FFT based beat signal pre-processing method, and the carefully designed STDC and ASTCAC blocks, our methods improves by 2.50% (97.22% versus 94.72%) on the test dataset than the best result of other methods using 5D feature representation, and this improvement may help the HGR systems play robustly in practical application scenarios, especially in high-risk application scenarios such as autonomous driving, where small increase in recognition accuracy may have a chance to avoid driving accidents and ensure driving safety.

ACKNOWLEDGMENT

Thanks to Shengyuan Wang for setting up and commissioning the radar system, and thanks to the subjects for helping us collect radar data.

REFERENCES

[1] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.

[2] Y. Zhang, S. Dong, C. Zhu, M. Balle, B. Zhang and L. Ran, "Hand Gesture Recognition for Smart Devices by Classifying Deterministic Doppler Signals," *IEEE Trans. Microw. Theory Tech.*, doi: 10.1109/TMTT.2020.3031619.

[3] J. Le Kernec et al., "Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 29–41, Jul. 2019.

[4] T. Starner, J. Weaver and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371-1375, Dec. 1998.

[5] Wan Q et al., "Gesture recognition for smart home applications using portable radar sensors," in *36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Chicago, IL, USA, 2014, pp. 6414-6417.

[6] Faheem K, Seong L, Sung C, "Hand-Based Gesture Recognition for Vehicular Applications Using IR-UWB Radar," *Sensors*, vol 17, no. 4, pp. 833-850, Apr. 2017.

[7] Cheok M J, Omar Z, Jaward M H, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol 10, pp. 1-23, Jan. 2019.

[8] Z. Xia, Y. Luomei, C. Zhou and F. Xu, "Multidimensional Feature Representation and Learning for Robust Hand-Gesture Recognition on Commercial Millimeter-Wave Radar," *IEEE Trans. Geosci. Remote Sensing*, pp. 1-16, July 2020.

[9] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang and J. Yang, "A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors," *IEEE Trans. Syst. Man Cybern. A Syst.*, vol. 41, no. 6, pp. 1064-1076, Nov. 2011.

[10] J. Lien et al., "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, p. 142, 2016.

[11] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018.

[12] M. Ritchie, A. Jones, J. Brown, and H. D. Griffiths, "Hand Gesture Classification using 24 GHz FMCW Dual Polarised Radar," in *Int. Conf. Radar Sys. (Radar 2017)*, Belfast, UK, 2017, pp. 1–6.

[13] B. Dekker et al., "Gesture recognition with a low power fmcw radar and a deep convolutional neural network," in *Proc. Eur. Radar Conf. (EuRAD)*, Nuremberg, Germany, pp. 163–166, 2017.

[14] J. S. Suh et al., "24 GHz FMCW Radar System for Real-Time Hand Gesture Recognition Using LSTM," in *Asia-Pacific Micro. Conf. (APMC)*, Singapore, pp. 860–862, 2018.

[15] J. Yu, L. Yen and P. Tseng, "mmWave Radar-based Hand Gesture Recognition using Range-Angle Image," in *IEEE 91st Veh. Tech. Conf. (VTC2020-Spring)*, Antwerp, Belgium, 2020, pp. 1-5.

[16] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3041-3048, Apr. 2019.

[17] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke and V. M. Lubecke, "Hand Gesture Recognition Using a Radar Echo I–Q Plot and a Convolutional Neural Network," *IEEE Sensors Letters*, vol. 2, no. 3, pp. 1-4, Sept. 2018.

[18] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz and F. Fioranelli, "Sequential Human Gait Classification with Distributed Radar Sensor Fusion," *IEEE Sensors Journal*, doi: 10.1109/JSEN.2020.3046991.

[19] Y. Wang, S. Wang, M. Zhou, Q. Jiang, and Z. Tian, "TS-I3D based hand gesture recognition method with radar sensor," *IEEE Access*, vol. 7, pp. 22902-22913, 2019.

[20] Y. Wang et al., "Gesture Recognition with Multi-dimensional Parameter Using FMCW Radar", *Journal of Electronics and Information Technology*, vol. 41, no. 4, pp. 822-829, 2019.

[21] Y. Sun, T. Fei, S. Gao, and N. Pohl, "Automatic radar-based gesture detection and classification via a region-based deep convolutional neural network," in *Proc. IEEE Int. Conf. Acoust*., Speech, Signal Process., 2019, pp. 4300–4304.

[22] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz, and N. Pohl, "Multi-feature encoder for radar-based gesture recognition," in *Proc. IEEE Int. Radar Conf. (RadarConf)*, 2020, pp. 351–356.

[23] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz and N. Pohl, "Real-Time Radar-Based Gesture Detection and Recognition Built in an Edge-Computing Platform," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10706-10716, Sept. 2020.

[24] A. D. Berenguer, M. C. Oveneke, H. Khalid, M. Alioscha-Perez, A. Bourdoux and H. Sahli, "GestureVLAD: Combining Unsupervised Features Representation and Spatio-Temporal Aggregation for Doppler-Radar Gesture Recognition," *IEEE Access*, vol. 7, pp. 137122-137135, 2019.

[25] G. Malysa, D. Wang, L. Netsch, and M. Ali, "Hidden Markov model-based gesture recognition with FMCW radar," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 1017–1021.

[26] Huang D Y, Hu W C, Chang S H, "Vision-based hand gesture recognition using PCA+ Gabor filters and SVM" in *IEEE Int. Conf. Intelligent Inf. Hiding Multimedia Signal Process (IIH-MSP)*, Kyoto, Japan, Sep. 2009, pp. 1-4.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, 1997.

[28] Liu J, Wang G, Duan L Y, et al., "Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks,". *IEEE Trans. Image Process*., vol. 27, no. 99, pp. 1586-1599, Apr., 2018.

[29] Jifeng Dai et al., "Deformable convolutional networks" in *Int. Conf. Com. Vis. (ICCV)*, Venice, Italy, 2017, pp. 764-773.

[30] Y. Zhou, X. Sun, Z. Zha, andW. Zeng, "MiCT: Mixed 3D/2D convolutional tube for human action Recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.(CVPR)*, 2018, pp. 449–458.

[31] Ralph O Schmidt, "Multiple Emitter Location and signal Parameter Estimation", *IEEE Trans. On Antennas and Propagation*, vol. 34, pp. 276-280, March 1986.

[32] C. M. Schmid, R. Feger, C. Pfeffer and A. Stelzer, "Motion compensation and efficient array design for TDMA FMCW MIMO radar systems," in *Proc. Eur. Conf. Antennas Propag. (EUCAP)*, 2012, pp. 1746-1750.

[33] S. Ji, W. Xu, M. Yang, K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.

[34] J. Carreira, A. Zisserman, Quo Vadis, "Action Recognition? A New Model and the Kinetics Dataset," in *Proc. Conf. Com. Vis. Pat. Rec. (CVPR)*, Honolulu, Hawaii, 2017, pp. 6299–6308.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks, in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 4489–4497.

[36] K. Hara, H. Kataoka, Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in *Proc. Conf. Com. Vis. Pat. Rec. (CVPR*), Salt Lake City, Utahp, 2018, pp. 6546–6555.

[37] K. He, X. Zhang, S. Ren, J. Sun, "Identity Mappings in Deep Residual Networks" in *Proc. Euro. Conf. Com. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 630–645.

[38] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-accuracy Trade-offs in Video Classification," in *Proc. Euro. Conf. Com. Vis. (ECCV)*, Munich, Germany, 2018, pp. 305–321.

[39] Z. Qiu, T. Yao, T. Mei, "Learning deep spatio-temporal dependence for semantic video segmentation", *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 939-949, 2017.

[40] Bertasius, G., Torresani, L., and Shi, J, "Object detection in video with spatiotemporal sampling networks," in *Proc. Euro. Conf. Com. Vis. (ECCV)*, Munich, Germany, 2018, pp. 342-357.

[41] Tian, Yapeng, et al. "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution," in *Proc. Conf. Com. Vis. Pat. Rec.*, Jun. 2020, pp. 3360-3369.

[42] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," 2020, arXiv:2004.02803. [Online]. Available: http://arxiv.org/abs/2004.02803.

[43] Wang Y, Yang J, Wang L, et al., "Light field image super-resolution using deformable convolution," *IEEE Trans. Image Process*., vol. 30, pp. 1057-1071, 2020.

[44] Zhang Y, Shi L, Wu Y, et al. "Gesture recognition based on deep deformable 3D convolutional neural networks," *Pattern Recognit*., 2020.

[45] Hu J, Shen L, Sun G., "Squeeze-and-excitation networks" in *Proc. Conf. Com. Vis. Pat. Rec. (CVPR)*, Salt Lake City, Utah, June 2018, pp. 7132-7141.

[46] Jia X, De Brabandere B, Tuytelaars T, et al., "Dynamic filter networks," in Adv. Neul. Inf. Proc. Sys. (NIPS), Barcelona, Spain, Dec. 2016, pp. 667-675.

[47] Liu J, He J, Qiao Y, et al., "Learning to Predict Context-adaptive Convolution for Semantic Segmentation" in *Proc. Euro. Conf. Com. Vis. (ECCV)*, Aug. 2020, pp. 769-786.

[48] M. Jankiraman, FMCW Radar Design. London, U.K.: Artech House, 2018.

[49] Texas Instruments. Robust traffic and intersection monitoring using millimeter wave sensors, Available: http://www.ti.com/cn/lit/wp/spyy002b/spyy002b.pdf, 2017.

[50] Karpathy A et al., "Large-scale video classification with convolutional neural networks," in *Proc. Conf. Com. Vis. Pat. Rec. (CVPR)*, Columbus, OH, USA, 2014, pp. 1725-1732.

[51] Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T, "HMDB: a large video database for human motion recognition," in 15th *Res. Rev. Wor. High. Perf. Com. Sci. Eng.(HLRS)*, Stuttgart, 2012, pp. 571-582.

[52] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.