Real-Time Radar-Based Gesture Detection and Recognition Built in an Edge-Computing Platform — Source link

Yuliang Sun, Tai Fei, Xibo Li, Alexander Warnecke ...+2 more authors

Institutions: Ruhr University Bochum, RWTH Aachen University

Topics: Gesture recognition, Personal computer, Convolutional neural network, Feature extraction and Recurrent neural network

Related papers:

- A Dynamic Continuous Hand Gesture Detection and Recognition Method with FMCW Radar

- TS-I3D Based Hand Gesture Recognition Method With Radar Sensor

- Real-Time Macro Gesture Recognition Using Efficient Empirical Feature Extraction With Millimeter-Wave Technology

- Spectrum-Based Hand Gesture Recognition Using Millimeter-Wave Radar Parameter Measurements

- Micro Gesture Recognition with Terahertz Radar Based on Diagonal Profile of Range-Doppler Map

Share this paper:

Real-Time Radar-Based Gesture Detection and Recognition Built in an Edge-Computing Platform

Yuliang Sun, *Student Member, IEEE,* Tai Fei, *Member, IEEE,* Xibo Li, Alexander Warnecke, Ernst Warsitz, and Nils Pohl, *Senior Member, IEEE*

arXiv:2005.10145v1 [eess.SP] 20 May 2020

Abstract—In this paper, a real-time signal processing framework based on a 60 GHz frequency-modulated continuous wave (FMCW) radar system to recognize gestures is proposed. In order to improve the robustness of the radar-based gesture recognition system, the proposed framework extracts a comprehensive hand profile, including range, Doppler, azimuth and elevation, over multiple measurement-cycles and encodes them into a feature cube. Rather than feeding the range-Doppler spectrum sequence into a deep convolutional neural network (CNN) connected with recurrent neural networks, the proposed framework takes the aforementioned feature cube as input of a shallow CNN for gesture recognition to reduce the computational complexity. In addition, we develop a hand activity detection (HAD) algorithm to automatize the detection of gestures in real-time case. The proposed HAD can capture the time-stamp at which a gesture finishes and feeds the hand profile of all the relevant measurement-cycles before this time-stamp into the CNN with low latency. Since the proposed framework is able to detect and classify gestures at limited computational cost, it could be deployed in an edge-computing platform for real-time applications, whose performance is notedly inferior to a state-of-the-art personal computer. The experimental results show that the proposed framework has the capability of classifying 12 gestures in real-time with a high F_1-score.

Index Terms—AoA information, FMCW radar, Gesture classification, Hand activity detection, Real-time.

I. INTRODUCTION

RADAR sensors are being widely used in many long-range applications for the purpose of target surveillance, such as in aircrafts, ships and vehicles [1], [2]. Thanks to the continuous development of silicon techniques, various electric components can be integrated in a compact form at a low price [2], [3]. Since radar sensors become more and more affordable to the general public, numerous emerging short-range radar applications, e.g., non-contact hand gesture recognition, are gaining tremendous importance in efforts to improve the quality of human life [4], [5]. Hand gesture recognition enables users to interact with machines in a more natural and intuitive manner than conventional touchscreen-based and button-based human-machine-interfaces [6]. For example, Google has integrated a 60 GHz radar into the smartphone Pixel 4, which allows users to change songs without touching the screen [7]. What's more, virus and bacteria surviving on surfaces for a long time could contaminate the interface and cause people's health problems. For instance, in 2020, tens of

thousands of people have been infected with COVID-19 by contacting such contaminate surfaces [8]. Radar-based hand gesture recognition allows people to interact with the machine in a touch-less way, which may reduce the risk of being infected with virus in a public environment. Unlike optical gesture recognition techniques, radar sensors are insensitive to the ambient light conditions; the electromagnetic waves can penetrate dielectric materials, which makes it possible to embed them inside devices. In addition, because of privacy-preserving reasons, radar sensors are preferable to cameras in many circumstances [9]. Furthermore, computer vision techniques applied to extract hand motion information in every frame are usually not power efficient, which is therefore not suitable for wearable and mobile devices [10].

Motivated by the benefits of radar-based touch-less hand gesture recognition, numerous approaches were developed in recent years. The authors in [9], [11], [12] extracted physical features from micro-Doppler signature [1] in the time-Doppler-frequency (TDF) domain to classify different gestures. Li *et al.* [13] extracted sparsity-based features from TDF spectrums for gesture recognition using a Doppler radar. In addition to Doppler information of hand gestures, the Google Soli project [10], [14] utilized the range-Doppler (RD) spectrums for gesture recognition via a 60 GHz frequency-modulated continuous wave (FMCW) radar sensor. Thanks to the wide available bandwidth (7 GHz), their systems could recognize fine hand motions. Similarly, the authors in [15]–[17] also extracted hand motions based on RD spectrums via an FMCW radar. In [18], [19], apart from the range and Doppler information of hand gestures, they also considered the incident angle information by using multiple receive antennas to enhance the classification accuracy of their gesture recognition system. However, none of the aforementioned techniques exploited all the characteristics of a gesture simultaneously, i.e., range, Doppler, azimuth, elevation and temporal information. For example, in [9]–[16], they could not differentiate two gestures, which share similar range and Doppler information. This restricts the design of gestures to be recognized.

In order to classify different hand gestures, many research works employed artificial neural networks for this multi-class classification task. For example, the authors in [12], [18]–[20] considered the TDF spectrums or range profiles as images and directly fed them into a deep convolutional neural network (CNN). Whereas, other research works [14], [15], [21] considered the radar data over multiple measurement-cycles

as a time-sequential signal, and utilized both the CNNs and recurrent neural networks (RNNs) for gesture classification. The Soli project [14] employed a 2-dimensional (2-D) CNN with a long short-term memory (LSTM) to extract both the spatial and temporal features, while the Latern [21], [22] replaced the 2-D CNN with 3-D CNN [23] followed by several LSTM layers. Because the 3-D CNN could extract not only the spatial but also the short-term temporal information from the RD spectrum sequence, it results in a better classification accuracy than the 2-D CNN [24]. However, the proposed 2-D CNN, 3-D CNN and LSTM for gesture classification require huge amounts of memory in the system, and are computationally inefficient. Although Choi *et al.* [16] projected the range-Doppler-measurement-cycles into range-time and Doppler-time to reduce the input dimension of the LSTM layer and achieved a good classification accuracy in real-time, the proposed algorithms were implemented on a personal computer with powerful computational capability. As a result, the aforementioned radar-based gesture recognition system in [12], [14]–[16], [18]–[21] are not applicable for most commercial embedded systems such as wearable devices, smartphones, in which both memory and computational power are limited.

In this paper, we present a real-time gesture recognition system using a 60 GHz FMCW radar in an edge-computing platform. The proposed system is expected to be applied in short-range applications (e.g., tablet, display, and smartphone) where the radar is assumed to be stationary to the user. The entire signal processing framework is depicted in Fig. 1. After applying the 2-dimensional finite Fourier transform (2-D FFT) to the raw data, we select a certain number of points from the resulting RD spectrum as an intermediate step rather than directly putting the entire spectrum into deep neural networks. Additionally, thanks to the L-shaped receive antenna array, the angle of arrival (AoA) information of the hand, i.e., azimuth and elevation, can be calculated. For every measurement-cycle, we store this information in a feature matrix with reduced dimensions. By selecting a few points from the RD spectrum, we reduce the input dimension of the classifier and limit the computational cost. Further, we present a hand activity detection (HAD) algorithm called the short-term average/long-term average (STA/LTA)-based gesture detector. It employs the concept of STA/LTA [25] to detect when a gesture comes to an end, i.e., the tail of a gesture. After detecting the tail of a gesture, we arrange the feature matrices belonging to the measurement-cycles, which are previous to this tail, into a feature cube. This feature cube constructs a compact and comprehensive gesture profile which includes the features of all the dominant point scatters of the hand. It is subsequently fed into a shallow CNN for classification. The main contributions are summarized as follows:

- The proposed signal processing framework is able to recognize more gestures (12 gestures) than those reported in other works in the literature. The framework can run in real-time built in an edge-computing platform with limited memory and computational capability.
- We develop a multi-feature encoder to construct the ges-

ture profile, including range, Doppler, azimuth, elevation and temporal information into a feature cube with reduced dimensions for the sake of data processing efficiency.
- We develop an HAD algorithm based on the concept of STA/LTA to reliably detect the tail of a gesture.
- Since the proposed multi-feature encoder has encoded all necessary information in a compact manner, it is possible to deploy a shallow CNN with a feature cube as its input to achieve a promising classification performance.
- The proposed framework is evaluated twofold: its performance is compared with the benchmark in off-line scenario, and its recognition ability in real-time case is assessed as well.

The remainder of this paper is organized as follows. Section II introduces the FMCW radar system. Section III describes the multi-feature encoder including the extraction of range, Doppler and AoA information. In Section IV, we introduce the HAD algorithm based on the concept of the STA/LTA. In Section V, we present the structure of the applied shallow CNN for gesture classification. In Section VI, we describe the experimental scenario and the collected gesture dataset. In Section VII, the performance is evaluated in both off-line and real-time cases. Finally, conclusions are given in Section VIII.

II. SHORT-RANGE FMCW RADAR SYSTEM

Our 60 GHz radar system adopts the linear chirp sequence frequency modulation [26] to design the waveform. After mixing, filtering and sampling, the discrete beat signal consisting of I_T point scatters of the hand in a single measurement-cycle from the z-th receive antenna can be approximated as [27]:

$$b^{(z)}(u,v) \approx \sum_{i=1}^{I_T} a_i^{(z)} \exp\left\{ j2\pi \left(f_{\mathrm{r}i} u T_{\mathrm{s}} - f_{\mathrm{D}i} v T_{\mathrm{c}} \right) \right\},$$

$$u = 0, \cdots, I_{\mathrm{s}} - 1, \quad v = 0, \cdots, I_{\mathrm{c}} - 1, \tag{1}$$

where the range and Doppler frequencies $f_{\mathrm{r}i}$ and $f_{\mathrm{D}i}$ are given as:

$$f_{\mathrm{r}i} = 2\frac{f_{\mathrm{B}}}{T_{\mathrm{c}}}\frac{r_i}{c}, \quad f_{\mathrm{D}i} = 2\frac{v_{\mathrm{r}i}}{\lambda}, \tag{2}$$

respectively, r_i and $v_{\mathrm{r}i}$ are the range and relative velocity of the i-th point scatter of the hand, f_{B} is the available bandwidth, T_{c} is the chirp duration, λ is the wavelength at 60 GHz, c is the speed of light, the complex amplitude $a_i^{(z)}$ contains the phase information, I_{s} is the number of sampling points in each chirp, I_{c} is the number of chirps in every measurement-cycle, and the sampling period $T_{\mathrm{s}} = T_{\mathrm{c}}/I_{\mathrm{s}}$. The 60 GHz radar system applied for gesture recognition can be seen in Fig. 2. It can also be seen that, the radar system has an L-shaped receive antenna array. To calculate the AoA in azimuth and elevation directions, the spatial distance between two receive antennas in both directions is d, where $d = \lambda/2$.

III. MULTI-FEATURE ENCODER

A. 2-D Finite Fourier Transform

A 2-D FFT is applied to the discrete beat signal in (1) to extract the range and Doppler information in every measurement-
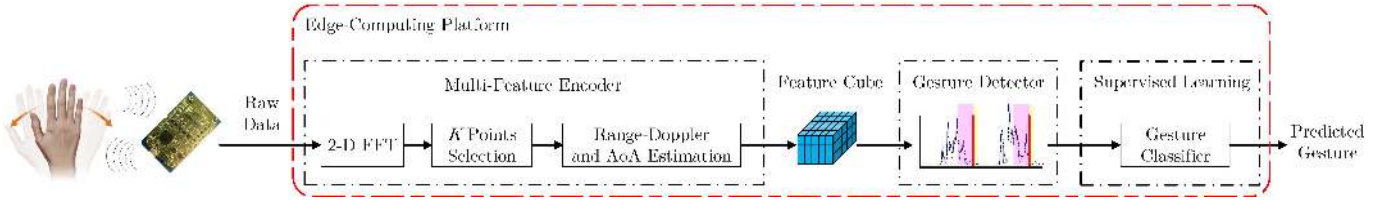
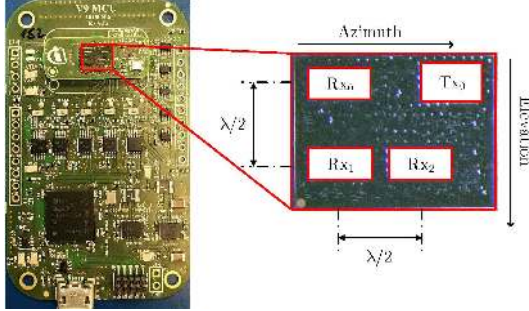Fig. 1. Proposed real-time radar-based gesture recognition framework built in an edge-computing platform.

Fig. 2. Infineon's BGT60TR13C 60 GHz radar system for gesture recognition. The $\{\text{Rx}_z : z = 0, 1, 2\}$ denotes the z-th receive antenna. The pair consisting of Rx_1 and Rx_0 is responsible for elevation angle, and the pair consisting of Rx_1 and Rx_2 is used for azimuth angle calculation.

cycle [28]. The resulting complex-valued RD spectrum for the z-th receive antenna can be calculated as:

$$
\begin{aligned}
B^{(z)}(p,q) = {} & \frac{1}{I_\text{c} I_\text{s}} \sum_{v=0}^{I_\text{c}-1} \sum_{u=0}^{I_\text{s}-1} \left\{ b^{(z)}(u,v) w(u,v) \right\} \\
& \cdot \exp\left(-j2\pi \frac{pu}{I_\text{s}}\right) \cdot \exp\left(-j2\pi \frac{qv}{I_\text{c}}\right), \\
& p = 0, \cdots, I_\text{s} - 1, \quad q = 0, \cdots, I_\text{c} - 1,
\end{aligned}
\tag{3}
$$

where $w(u,v)$ is a 2-D window function, p and q are the range and Doppler frequency indexes. The range and relative velocity resolution can be deduced as:

$$
\Delta r = c \frac{T_\text{c}}{2f_\text{B}} \cdot \Delta f_\text{r} = \frac{c}{2f_\text{B}}, \quad \Delta v_\text{r} = \frac{\lambda}{2} \cdot \Delta f_\text{D}, \tag{4}
$$

where the range and Doppler frequency resolution Δf_r and Δf_D are $1/T_\text{c}$ and $1/(I_\text{c} T_\text{c})$, respectively. To improve the signal-to-noise ratio (SNR), we sum the RD spectrums of the three receive antennas incoherently, i.e.,

$$
RD(p,q) = \sum_{z=0}^{2} |B^{(z)}(p,q)|. \tag{5}
$$

B. Range-Doppler Estimation

To obtain the range, Doppler and AoA information of the hand in every measurement-cycle, we select K points from $RD(p,q)$, which have the largest magnitudes. The parameter K is predefined, and its choice will be discussed in Section VII-A. Then, we extract the range, Doppler frequencies and the magnitudes of those K points, which are denoted as $\hat{f}_{\text{r}k}$, $\hat{f}_{\text{D}k}$ and A_k, respectively, where $k = 1, \cdots, K$.

C. Azimuth and Elevation Angle Estimation

The AoA can be calculated from the phase difference of extracted points in the same positions of complex-valued RD spectrums belonging to two receive antennas. The AoA in azimuth and elevation of the k-th point can be calculated as:

$$
\hat{\phi}_k = \arcsin\left(\frac{\left(\psi\left(a_k^{(1)}\right) - \psi\left(a_k^{(2)}\right)\right)\lambda}{2\pi d} \right), \tag{6}
$$

$$
\hat{\theta}_k = \arcsin\left(\frac{\left(\psi\left(a_k^{(1)}\right) - \psi\left(a_k^{(0)}\right)\right)\lambda}{2\pi d} \right), \tag{7}
$$

respectively, where $\psi(\cdot)$ stands for the phase of a complex value, $a_k^{(z)}$ is the complex amplitude $B^{(z)}\left(\hat{f}_{\text{r}k}, \hat{f}_{\text{D}k}\right)$ from the z-th receive antenna.

D. Feature Cube

As a consequence, in every measurement-cycle, the k-th point in $RD(p,q)$ has five attributes, i.e., range, Doppler, azimuth, elevation and magnitude. As depicted in Fig. 3, we encode the range, Doppler, azimuth, elevation and magnitude of those K points with the largest magnitudes in $RD(p,q)$ along I_L measurement-cycles into the feature cube $\mathcal{V}$ with dimension $I_L \times K \times 5$. The $\mathcal{V}$ has five channels corresponding to five attributes and each element in $\mathcal{V}$ at the l-th measurement-cycle can be described as:

$$
\begin{aligned}
\mathcal{V}(l,k,1) = \hat{f}_{\text{r}k}, \quad \mathcal{V}(l,k,2) = \hat{f}_{\text{D}k}, \quad \mathcal{V}(l,k,3) = \hat{\phi}_k, \\
\mathcal{V}(l,k,4) = \hat{\theta}_k, \quad \mathcal{V}(l,k,5) = A_k,
\end{aligned}
\tag{8}
$$

where $l = 1, \cdots, I_L$.

Fig. 3. Structure of feature cube $\mathcal{V}$.
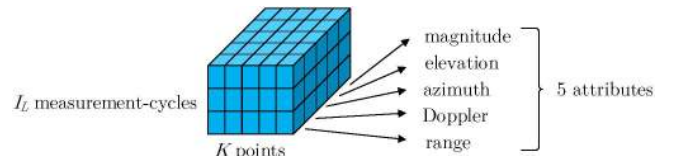
IV. HAND ACTIVITY DETECTION

A. Problem Statement

Similar to voice activity detection in the automatic speech recognition system, our gesture recognition system also needs to detect some hand activities in advance, before forwarding the data to the classifier. It helps to design a power-efficient gesture recognition system, since the classifier is only activated
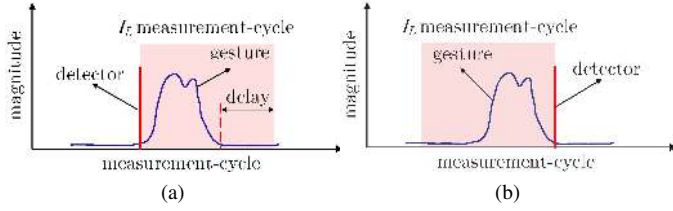
Fig. 4. (a) A delay occurs before forwarding the gesture data to the classifier when we detect the start time-stamp of the gesture. (b) The gesture data is directly forwarded to the classifier without delay when we detect the tail of the gesture.

when a gesture is detected rather than keeping it active for every measurement-cycle. The state-of-the-art event detection algorithms usually detect the start time-stamp of an event. For example, the authors in [25] used the STA/LTA and power spectral density methods to detect when a micro-seismic event occurs. In the case of radar-based gesture recognition, we could also theoretically detect the start time-stamp of a gesture and consider that a gesture event occurs within the following I_L measurement-cycles. However, detecting the start-stamp and forwarding the hand data in the following I_L measurement-cycles to the classifier could cause a certain time delay, since the time duration of designed gestures is usually different. As illustrated in Fig. 4(a), due to the facts that the proposed multi-feature encoder requires I_L measurement-cycles and the duration of the gesture is usually shorter than I_L, a delay occurs, if we detect the start time-stamp of the gesture. Therefore, as depicted in Fig. 4(b), to reduce the time delay, our proposed HAD algorithm is designed to detect when a gesture finishes, i.e., the tail of a gesture, rather than detecting the start time-stamp.

B. STA/LTA-based Gesture Detector

We propose a STA/LTA-based gesture detector to detect the tail of a gesture. The exponential moving average (EMA) is used to detect the change of the magnitude signal at the l-th measurement-cycle, which is given as:

$$M(l) = (1 - \alpha)M(l - 1) + \alpha x(l), \tag{9}$$

where $\alpha \in [0, 1]$ is the predefined smoothing factor, $x(l)$ is the range-weighted magnitude (RWM), and it is defined as:

$$x(l) = A_{\max} f_{\mathrm{r_{max}}}^{\beta}, \quad A_{\max} = \max_k \{A_k\}, \tag{10}$$

where $A_{\max}$ represents the maximal magnitude among K points in $RD(p, q)$ at l-th measurement-cycle, $f_{\mathrm{r_{max}}}$ denotes the range corresponding to $A_{\max}$, and the predefined coefficient β denotes the compensation factor. The radar cross section (RCS) of a target is independent of the propagation path loss between the radar and the target. According to the radar equation [29], the measured magnitude of a target is a function of many arguments, such as the path loss, RCS, etc. As deduced in (10), we have built a coarse estimate of the RCS by multiplying the maximal range information with its measured magnitude to partially compensate the path loss. Furthermore, we define the $STA(l)$ and $LTA(l)$ as the mean

EMA in short and long windows at the l-th measurement-cycle:

$$\mathrm{STA}(l) = \frac{1}{L_1} \sum_{ll=l+1}^{l+L_1} M(ll), \quad \mathrm{LTA}(l) = \frac{1}{L_2} \sum_{ll=l-L_2+1}^{l} M(ll), \tag{11}$$

respectively, where L_1 and L_2 are the length of the short and long window. The tail of a gesture is detected, when the following conditions are fulfilled:

$$\sum_{ll=l-L_2+1}^{l} x(ll) \geq \gamma_1 \quad \text{and} \quad \frac{\mathrm{STA}(l)}{\mathrm{LTA}(l)} \leq \gamma_2, \tag{12}$$

where γ_1 and γ_2 are the predefined detection thresholds. Fig. 5 illustrates that the tails of two gestures are detected via the proposed STA/LTA gesture detector. According to (12), one condition of detecting the tail of a gesture is that, the average of RWM in the long window exceeds the threshold γ_1. It means that a hand motion appears in the long window. The other condition is that, the ratio of the mean EMA in the short window and that in the long window is lower than the threshold γ_2. In other words, it detects when the hand movement finishes. In practice, the parameters β, γ_1 and γ_2 in our HAD algorithm should be thoroughly chosen according to different application scenarios.
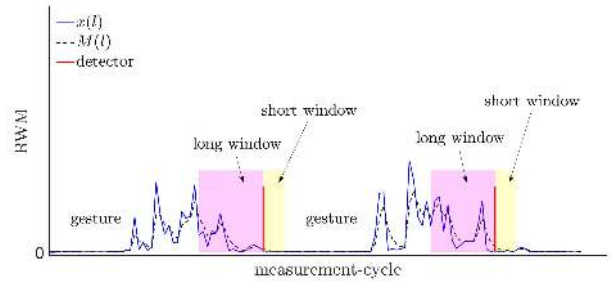

Fig. 5. The tails of two gestures are detected via the proposed gesture detector.

V. SUPERVISED LEARNING

As discussed in Section III-D, the feature cube obtained by the multi-feature encoder has a dimension of $I_L \times K \times 5$. Thus, we could simply use the CNN for classification without any reshaping operation. The structure of the CNN can be seen in Fig. 6. We employ four convolutional (Conv) layers, each of that has a kernel size 3×3 and the number of kernels in each Conv layer is 64. In addition, the depth of the first kernel is five, since the input feature cube has five channels (i.e., range, Doppler, azimuth, elevation and magnitude), while that of the other kernels in the following three Conv layers is 64. We choose the rectified linear unit (ReLU) [30] as activation function, since it solves the problem of gradient vanishing and is able to accelerate the convergence speed of training [31]. Then, the last Conv layer is connected by two fully-connected (FC) layers, either of which has 256 hidden units and is followed by a dropout layer for preventing the network from overfitting. The third FC layer with a softmax function is utilized as the output layer. The number of hidden units in the

third FC layer is designed to be in accordance with the number of classes in the dataset. The softmax function normalizes the output of the last FC layer to a probability distribution over the classes.

Through thoroughly network tuning (e.g., number of hidden layers, number of hidden units, depth number), we construct the CNN structure as shown in Fig. 6. The designed network should (a) take the feature cube as input, (b) achieve a high classification accuracy, (c) consume few computational resources, and (d) be deployable in the edge-computing platform. In Section VII, we will show that the designed network in Fig. 6 fulfills these criteria.

VI. SCENARIO AND GESTURE DATASET DESCRIPTION

As illustrated in Fig. 7, we used the 60 GHz FMCW radar in Fig. 2 to recognize gestures. Our radar system has a detection range up to 0.9 m and an approx. $120°$ antenna beam width in both azimuth and elevation directions. The parameter setting used in the waveform design is presented in Table I, where the pulse repetition interval (PRI) is 34 ms. The radar is

TABLE I
TRANSMIT SIGNAL DESIGN AND RADAR PARAMETERS

Transmit signal design	f_c 60 GHz	f_B 5 GHz	T_c 432 µs	I_s 32	I_c 32	PRI 34 ms
Resolution	Δr 3 cm	Δv_r 18 cm/s				

connected with an edge-computing platform, i.e., NVIDIA Jetson Nano, which is equipped with Quad-core ARM A57 at 1.43 GHz as central processing unit (CPU), 128-core Maxwell as graphics processing unit (GPU) and 4 GB memory. We have built our entire radar-based gesture recognition framework described in Fig. 1 in the edge-computing platform in C/C++. The proposed multi-feature encoder and HAD have been implemented in a straightforward manner without any runtime optimization, while the implementation of the CNN is supported by TensorRT developed by NVIDIA. In addition, as depicted in Fig. 8, we designed 12 gestures, which are (a) Check, (b) Cross, (c) Rotate clockwise (CW), (d) Rotate counterclockwise (CCW), (e) Moving fingers (FG), (f) Pinch index, (g) Pull, (h) Push, (i) Swipe backward (BW), (j) Swipe forward (FW), (k) Swipe left (LT) and (l) Swipe right (RT). We invited 20 human subjects including both genders with various heights and ages to perform these gestures. Among 20 subjects, the ages range from 20 to 35 years old, and the heights are from 160 cm to 200 cm. We divided the 20 subjects into two groups. In the first group, ten subjects were taught how to perform gestures in a normative way. Whereas, in the second group, in order to increase the diversity of the dataset, only an example for each gesture was demonstrated to the other ten subjects and they performed gestures using their own interpretations. Self-evidently, their gestures were no longer as normative as the ones performed by the ten taught subjects. Furthermore, every subject repeated each gesture 30 times. Therefore, the total number of realizations in our gesture dataset is (12 gestures)×(20 people)×(30 times), namely 7200. We also found out that the gestures performed in our dataset

take less than 1.2 s. Thus, to ensure that the entire hand movement of a gesture is included in the observation time, we set I_L to 40, which amounts to a duration of 1.36 s (40 measurement-cycles × 34 ms).

VII. EXPERIMENTAL RESULTS

In this section, the proposed approach is evaluated regarding a twofold objective: first, its performance is thoroughly compared with benchmarks in literature through an off-line cross-validation, and secondly, its real-time capability is investigated with an on-line performance test. In Section VII-A, we discuss how the parameter K affects the classification accuracy. In Section VII-B, we compare our proposed algorithm with the state-of-the-art radar-based gesture recognition algorithms in terms of classification accuracy and computational complexity based on leave-one-out cross-validation (LOOCV). It means that, in each fold, we use the gestures from one subject as test set, and the rest as training set. In addition, Section VII-C describes the real-time evaluation results of our system. The performances of taught and untaught subjects are evaluated separately. We randomly selected eight taught and eight untaught subjects as training sets, while the remaining two taught and two untaught subjects are test sets. In real-time performance evaluation, we performed the hardware-in-the-loop (HIL) test, and fed the raw data recorded by the radar from the four test subjects into our edge-computing platform.

A. Determination of Parameter K

As described in Section III, we extract K points with the largest magnitudes from $RD(p, q)$, to represent the hand information in a single measurement-cycle. We define the average (avg.) accuracy as the avg. classification accuracy across the 12 gestures based on LOOCV. In Fig. 9, we let K vary from 1 to 40, and compute the avg. accuracy in five trials. It can be seen that the mean avg. accuracy over five trials keeps increasing and reaches approx. 95%, when K is 25. After that, increasing K can barely improve the classification accuracy. As a result, in order to keep low computational complexity of the system and achieve a high classification accuracy, we set K to 25. It results that the feature cube $\mathcal{V}$ in our proposed multi-feature encoder has a dimension of $40 \times 25 \times 5$.

B. Off-line Performance Evaluation

In the off-line case, we assumed that each gesture is perfectly detected by the HAD algorithm and compared our proposed multi-feature encoder + CNN with the 2-D CNN + LSTM [14], the 3-D CNN + LSTM [21], 3-D CNN + LSTM (with AoA) and shallow 3-D CNN + LSTM (with AoA) in terms of the avg. classification accuracy and computational complexity based on LOOCV. In our proposed multi-feature encoder + CNN, the feature cube $\mathcal{V}$, which has the dimension of $40 \times 25 \times 5$, was fed into the CNN described in Fig. 6. The input of the 2-D CNN + LSTM [14] and the 3-D CNN + LSTM [21] is the RD spectrum sequence over 40 measurement-cycles, which has the dimension of $40 \times 32 \times 32 \times 1$. Since [21] did not include any
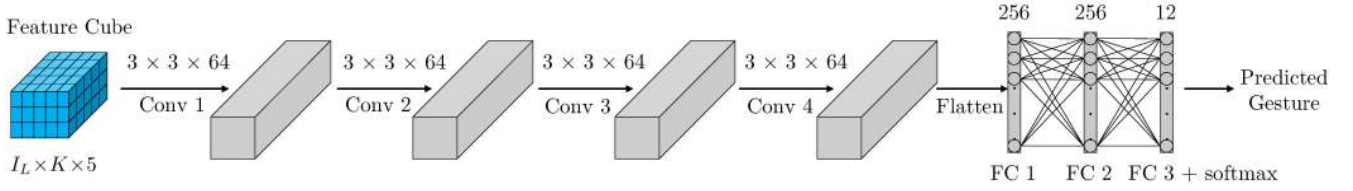
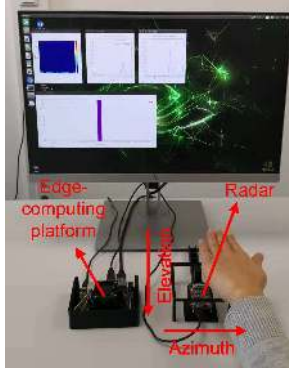Fig. 6. Structure of the shallow CNN taking the feature cube as input.

Fig. 7. Experiment scenario of the radar-based gesture recognition system.

Fig. 9. Determination of the number of extracted points K from $RD(p, q)$.
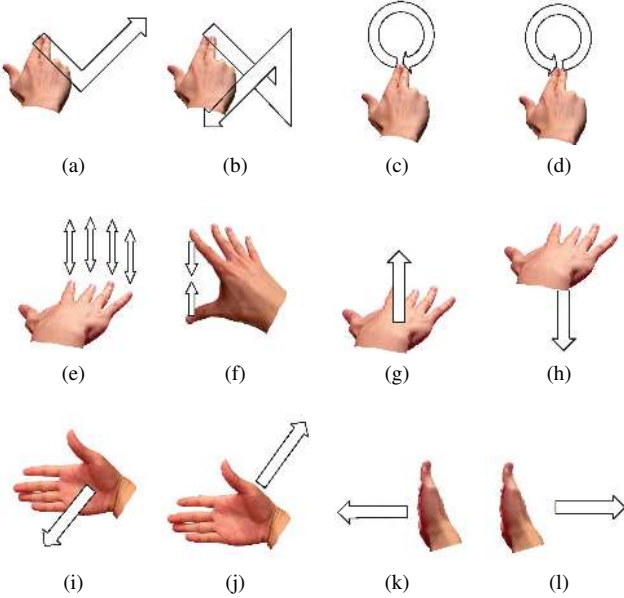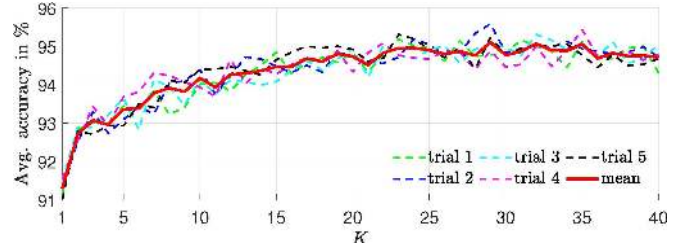

Fig. 8. (a) Check. (b) Cross. (c) Rotate CW. (d) Rotate CCW. (e) Moving fingers. (f) Pinch index. (g) Pull. (h) Push. (i) Swipe BW. (j) Swipe FW. (k) Swipe LT. (l) Swipe RT.

AoA information in their system for gesture classification, the comparison might not be fair. Thus, we added the AoA information according to (6) and (7) to the RD spectrum sequence. It results in the input of the 3-D CNN + LSTM (with AoA) and shallow 3-D CNN + LSTM (with AoA) having the dimension of $40 \times 32 \times 32 \times 3$, where the second and the third channel contain the AoA information in azimuth and elevation dimension, respectively. The shallow 3-D CNN + LSTM (with AoA) is designed to having comparable computational complexity as that of the proposed multi-feature encoder +

CNN but with reduced classification accuracy. To achieve a fair comparison, we optimized the structures and the hyper-parameters as well as the training parameters of those models. The CNN demonstrated in Fig. 6 in the proposed approach was trained for 15000 steps based on the back propagation [32] using the Adam optimizer [33] with an initial learning rate of 1×10^{-4}, which degraded to 10^{-5}, 10^{-6} and 10^{-7} after 5000, 8000 and 11000 steps, respectively. The batch size is 128.

1) Classification Accuracy and Training Loss Curve: In Table II, we present the classification accuracy of each type of gesture based on the algorithms mentioned above. The avg. accuracies of the 2-D CNN + LSTM [14] and 3-D CNN + LSTM [21] are only 78.50% and 79.76%, respectively. Since no AoA information is utilized, the Rotate CW and Rotate CCW can hardly be distinguished, and similarly the four Swipe gestures can hardly be separated, either. On the contrary, considering the AoA information, the multi-feature encoder + CNN, the 3-D CNN + LSTM (with AoA) and the shallow 3-D CNN + LSTM (with AoA) are able to separate the two Rotate gestures, and the four Swipe gestures. It needs to be mentioned that the avg. accuracy of our proposed multi-feature encoder is almost the same as that of the 3-D CNN + LSTM with (AoA). However, it will be shown in the following section that our approach requires much less computational resources and memory than those of the other approaches.

What's more, in Fig. 10, we plot the training loss curves of the three structures of neural networks. It can be seen that the loss of the proposed CNN in Fig. 6 has the fastest rate of convergence among the three structures of neural networks and approaches to zero at around the 2000-th training step. Unlike the input of the 3-D CNN + LSTM (with AoA) and shallow 3-D CNN + LSTM (with AoA), the feature cube contains sufficient gesture characteristics in spite of its compact form ($40 \times 25 \times 5$). It results that the CNN in Fig. 6 is easier to be trained than the other neural networks, and it achieves a high classification accuracy.

TABLE II
CLASSIFICATION ACCURACY IN % OF EACH GESTURE OBTAINED BY DIFFERENT GESTURE RECOGNITION FRAMEWORKS

	avg. acc.	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)
2-D CNN + LSTM [14]	78.50	85.17	82.67	60.67	55.50	93.33	95.00	90.67	91.17	66.83	75.33	67.50	78.17
3-D CNN + LSTM [21]	79.76	83.17	87.17	62.83	57.17	93.50	97.17	93.17	92.17	63.67	77.33	69.33	80.50
3-D CNN + LSTM (with AoA)	95.57	96.67	95.17	97.17	96.33	92.00	95.17	94.67	95.17	94.50	93.50	98.50	98.0
Shallow 3-D CNN + LSTM (with AoA)	94.36	95.33	92.0	96.67	96.83	93.33	94.83	95.83	93.0	90.0	89.17	98.83	96.50
Multi-feature encoder + CNN	95.79	96.50	97.83	95.83	96.83	96.17	95.50	93.17	96.50	92.67	92.67	98.17	97.67

TABLE III
DIFFERENT NEURAL NETWORK STRUCTURES FOR RADAR-BASED GESTURE RECOGNITION

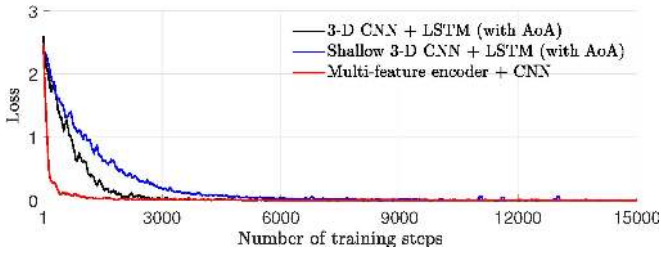| Layers	3-D CNN + LSTM (with AoA)	Shallow 3-D CNN + LSTM (with AoA)	Multi-feature encoder + CNN
input	$40 \times 32 \times 32 \times 3$	$40 \times 32 \times 32 \times 3$	$40 \times 25 \times 5$
1	3-D Conv1 $3 \times 3 \times 3 \times 16$	3-D Conv1 $3 \times 3 \times 3 \times 16$	Conv1 $3 \times 3 \times 64$
2	3-D Conv2 $3 \times 3 \times 3 \times 32$	3-D Conv2 $3 \times 3 \times 3 \times 32$	Conv2 $3 \times 3 \times 64$
3	3-D Conv3 $3 \times 3 \times 3 \times 64$	3-D Conv3 $3 \times 3 \times 3 \times 64$	Conv3 $3 \times 3 \times 64$
4	FC1 512	FC1 64	Conv4 $3 \times 3 \times 64$
5	FC2 512	LSTM 32	FC1 256
6	LSTM 512	FC2 12 - softmax	FC2 256
7	FC3 12 - softmax	-	FC3 12 - softmax
GFLOPs	2.89	0.34	0.26
Size	109 MB	101 MB	4.18 MB

Fig. 10. Comparison of training loss curves.

2) Confusion Matrix: In Fig. 11, we plotted two confusion matrices for ten taught and ten untaught subjects based on our proposed multi-feature encoder + CNN. It could be observed that, for the normative gestures performed by the ten taught subjects, we could reach approx. 98.47% avg. accuracy. Although we could observe an approx. 5% degradation in avg. accuracy in Fig. 11(b), where the gestures to be classified are performed by ten untaught subjects, it still has 93.11% avg. accuracy.

3) Computational Complexity and Memory: The structures of the 3-D CNN + LSTM (with AoA), shallow 3-D CNN + LSTM (with AoA) and the proposed multi-feature encoder + CNN are presented in Table III. We evaluated their computational complexity and required memory in line with the giga floating point operations per second (GFLOPs) and the model size. The GFLOPs of different models were calculated by the built-in function in TensorFlow, the model size is observed through TensorBoard [34]. Although the 3-D CNN + LSTM (with AoA) offers almost the same classification accuracy as that of the proposed multi-feature encoder + CNN, it needs much more GFLOPs than that of the multi-feature encoder + CNN (2.89 GFLOPs vs. 0.26 GFLOPs). Its model size is also much larger than that of the proposed approach (109 MB vs. 4.18 MB). Although we could reduce its GFLOPs using a shallow network structure, such as the shallow 3-D CNN + LSTM (with AoA) in Table III, it results in the degradation

of classification accuracy (94.36%), as can be seen in Table II. We also found out that the CNN used in our approach has the least model size, since its input dimension is much smaller than that of other approaches. On the contrary, the input of the 3-D CNN + LSTM (with AoA) contains lots of zeros due to the sparsity of RD spectrums. Such large volumes usually need large amounts of coefficients in neural networks. Whereas, we exploit the hand information in every measurement-cycle using only 25 points, and the input dimension of the CNN is only $40 \times 25 \times 5$, which requires much less computational complexity than the other approaches.

C. Real-time Performance Evaluation

As mentioned above, subjects are divided into taught and untaught groups, and each has ten subjects. In each group, eight subjects are randomly selected as training set, and the remaining two subjects constitute the test set, resulting in either group having 720 true gestures in the test set. In the HIL context, we directly fed the recorded raw data from the four test subjects into the edge-computing platform. In the real-time case, the system should be robust enough to distinguish true gestures from random motions (RMs). Thus, we also included a certain amount of RMs as negative samples during the training phase. The scale of RMs and true gestures is around 1:3.

1) Precision, Recall and F_1-score: To quantitatively analyze the real-time performance of our system, we introduce the precision, recall and F_1-score, which are calculated as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$F_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (13)$$

where TP, FP and FN denote the number of true positive, false positive, and false negative estimates. For two subjects in the test set, we have 60 realizations for each gesture. It

Avg. accuracy: 98.47 %

True	a	b	c	d	e	f	g	h	i	j	k	l
a)	95.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	2.0
b)	0.2	98.0	0.0	0.7	0.0	0.3	0.0	0.3	0.0	0.0	0.0	0.3
c)	0.0	0.0	98.3	0.0	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0
d)	0.0	0.3	0.0	99.0	0.3	0.0	0.0	0.3	0.0	0.0	0.0	0.0
e)	0.0	0.0	0.0	0.0	98.3	1.3	0.0	0.0	0.0	0.3	0.0	0.0
f)	0.0	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.3	0.0	0.0	0.0
g)	0.7	0.0	0.0	0.3	0.0	0.0	98.7	0.3	0.0	0.0	0.0	0.0
h)	0.0	0.0	0.0	0.0	0.3	1.0	0.0	98.3	0.0	0.3	0.0	0.0
i)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.0	2.0	0.0	0.0
j)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	99.3	0.0	0.0
k)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.7	99.0	0.0
l)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Predicted

(a)

Avg. accuracy: 93.11 %

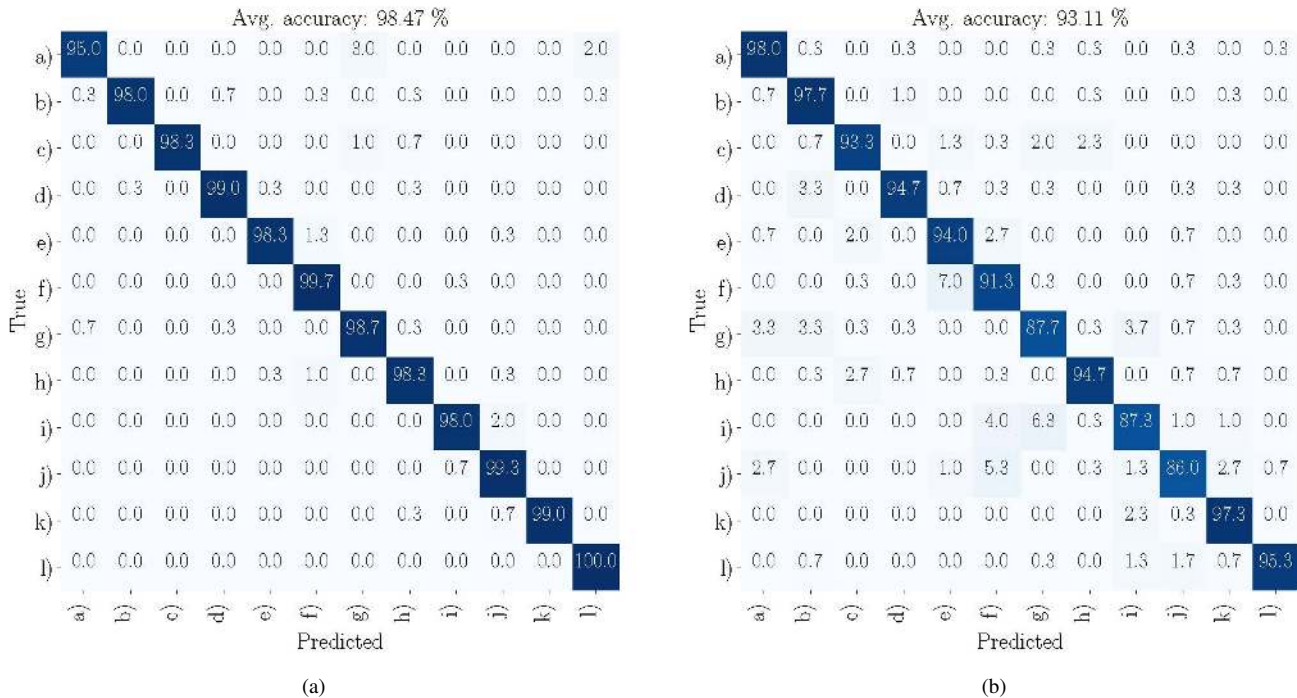| True	a	b	c	d	e	f	g	h	i	j	k	l
a)	98.0	0.3	0.0	0.3	0.0	0.0	0.3	0.3	0.0	0.3	0.0	0.3
b)	0.7	97.7	0.0	1.0	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.0
c)	0.0	0.7	93.3	0.0	1.3	0.3	2.0	2.3	0.0	0.0	0.0	0.0
d)	0.0	3.3	0.0	94.7	0.7	0.3	0.3	0.0	0.0	0.3	0.3	0.0
e)	0.7	0.0	2.0	0.0	94.0	2.7	0.0	0.0	0.0	0.7	0.0	0.0
f)	0.0	0.0	0.3	0.0	7.0	91.3	0.3	0.0	0.0	0.7	0.3	0.0
g)	3.3	3.3	0.3	0.3	0.0	0.0	87.7	0.3	3.7	0.7	0.3	0.0
h)	0.0	0.3	2.7	0.7	0.0	0.3	0.0	94.7	0.0	0.7	0.7	0.0
i)	0.0	0.0	0.0	0.0	0.0	4.0	6.3	0.3	87.3	1.0	1.0	0.0
j)	2.7	0.0	0.0	1.0	5.3	0.0	0.3	1.3	86.0	2.7	0.7	
k)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.3	0.3	97.3	0.0
l)	0.0	0.7	0.0	0.0	0.0	0.0	0.3	0.0	1.3	1.7	0.7	95.3

Predicted

(b)

Fig. 11. Confusion matrices obtained by the proposed multi-feature encoder + CNN based on LOOCV. (a) 10 taught subjects. (b) 10 untaught subjects.

means that TP + FN = 60. As presented in Table IV, the avg. precision and recall over 12 types of gestures using two taught subjects as test set are 93.90% and 94.44%, respectively, while those using two untaught subjects as test set are 91.20% and 86.11%. It needs to be mentioned that, the off-line avg. accuracies in Fig. 11, namely 98.47 % and 93.11%, can also be regarded as the recall in taught and untaught cases. After comparing with the recall in the off-line case, we could observe an approx. 4% and 7% degradation in recall in the real-time case considering both the taught and untaught subjects. The reason is that, in the off-line performance evaluation, we assumed that each gesture is detected perfectly. However, in the real-time case, the recall reduction is caused by the facts that our HAD performance miss-detected some gestures or incorrectly triggered the classifier even when the gesture was not completely finished. For example, due to the small movement of the hand, the HAD sometimes failed to detect the gesture "Pinch index". Similarly, the recall of the gesture "Cross" is also impaired, since the gesture "Cross" has a turning point, which leads to a short pause. In some cases where the subject performs the gesture "Cross" with low-velocity, the HAD would incorrectly consider the turning point as the end of "Cross", resulting in a wrong classification. Overall, in both taught and untaught cases, the F_1-score of our radar-based gesture recognition system reaches 94.17% and 88.58%, respectively.

2) Detection Matrix: We summarized the gesture detection results of our real-time system. Since we did not aim to evaluate the classification performance here, we depicted the detection results in Table V considering all four test subjects. Our system correctly detected 1388 true positive gestures, and provoked 25 false alarms among the total of 1864 test samples in which there are 1440 true gestures and 424 true negative RMs, respectively. Furthermore, we define two different types of miss-detections (MDs), in which the MDs from HAD means that our HAD miss-detects a gesture, while the MDs from the classifier means that, the HAD detects the gesture, but this gesture is incorrectly rejected by our classifier as a RM. The false alarm rate (FAR) and miss-detection rate (MDR) of our system are 5.90% and 3.61%, respectively.

3) Runtime: As depicted in Table VI, in the HIL context, we also noted the avg. runtime of the multi-feature encoder, HAD and CNN based on all the 1838 classifications, which include 1388 true positives, 399 true negatives, 25 false alarms and 26 MDs from the classifier. The multi-feature encoder includes the 2-D FFT, 25 points selection, RD and AoA estimation. It needs to be mentioned that the multi-feature encoder and the HAD were executed in the CPU using unoptimized C/C++ code, while the CNN ran in the GPU based on TensorRT. The multi-feature encoder and HAD took only approx. 7.12 ms and 0.38 ms without using any FFT acceleration engine, while the CNN took only 25.84 ms on average. The overall runtime of our proposed radar-based gesture recognition system is only approx. 33 ms.

VIII. CONCLUSION

We developed a real-time radar-based gesture recognition system built in an edge-computing platform. The proposed multi-feature encoder could effectively encode the gesture profile, i.e., range, Doppler, azimuth, elevation, temporal information as a feature cube, which is then fed into a shallow CNN for gesture classification. Furthermore, to reduce the latency caused by the fixed number of required measurement-cycles in our system, we proposed the STA/LTA-based gesture

TABLE IV
PRECISION, RECALL AND F_1-SCORE IN % OF THE REAL-TIME RADAR-BASED GESTURE RECOGNITION SYSTEM IN BOTH TAUGHT AND UNTAUGHT CASES

		avg.	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	F_1-score
Taught	precision	93.90	85.94	100.0	85.51	96.77	98.33	84.51	93.44	100.0	90.32	100.0	98.28	93.75	94.17
	recall	94.44	91.67	73.33	98.33	100.0	98.33	100.0	95.0	98.33	93.33	90.0	95.0	100.0	
Untaught	precision	91.20	88.24	88.33	95.83	85.07	98.18	87.23	73.53	98.08	96.55	83.33	100.0	100.0	88.58
	recall	86.11	100.0	88.33	76.67	95.0	90.0	68.33	83.33	85.0	93.33	83.33	96.67	73.33	

TABLE V
GESTURE DETECTION MATRIX BASED ON FOUR TEST SUBJECTS

Detected Positive	Detected Negative		Overall
1388	26	26	1440
True Positives	MDs from HAD	MDs from classifier	True Gestures
25	399		424
False Alarms	True Negatives		Negative Samples

TABLE VI
AVERAGE RUNTIME OF THE GESTURE RECOGNITION SYSTEM

	Multi-feature encoder (CPU)	HAD (CPU)	CNN (GPU)	Overall
avg. runtime	7.12 ms	0.38 ms	25.84 ms	33.15 ms

detector, which detects the tail of a gesture. In the off-line case, based on LOOCV, our proposed gesture recognition approach achieves 98.47% and 93.11% avg. accuracy using gestures from taught and untaught subjects, respectively. In addition, the trained shallow CNN has a small model size and requires few GFLOPs. In the HIL context, our approach achieves 94.17% and 88.58% F_1-scores based on two taught and two untaught subjects as test sets, respectively. Finally, our system could be built in the edge-computing platform, and requires only approx. 33 ms to recognize a gesture. Thanks to the promising recognition performance and low computational complexity, our proposed radar-based gesture recognition system has the potential to be utilized for numerous applications, such as mobile and wearable devices. In future works, different gesture datasets with large diversity need to be constructed according to specific use cases. What's more, in some use cases where the radar is not stationary to the user, the classification accuracy of the proposed system might decrease and accordingly algorithms, such as ego motion compensation, could be considered.

ACKNOWLEDGMENT

REFERENCES

[1] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: Phenomenon, model, and simulation study," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 1, pp. 2–21, 2006.

[2] J. Hasch *et al.*, "Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 3, pp. 845–860, 2012.

[3] N. Pohl, T. Jaeschke, and K. Aufinger, "An ultra-wideband 80 GHz FMCW radar system using a SiGe bipolar transceiver chip stabilized by a fractional-N PLL synthesizer," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 3, pp. 757–765, 2012.

[4] 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, 2019.

[5] 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, 2019.

[6] C. Gu, J. Wang, and J. Lien, "Motion sensing using radar: Gesture interaction and beyond," *IEEE Microw. Mag.*, vol. 20, no. 8, pp. 44–57, 2019.

[7] D. Bohn, *Google Pixel 4 and 4 XL hands-on: This time, it's not about the camera*, 2019. [Online]. Available: https://www.theverge.com/2019/10/15/20908071/google-pixel-4-xl-photos-video-hands-on-camera-screen-specs-price

[8] G. Kampf, D. Todt, S. Pfaender, and E. Steinmann, "Persistence of coronaviruses on inanimate surfaces and its inactivation with biocidal agents," *J. Hospital Infection*, 2020.

[9] Y. Sun, T. Fei, F. Schliep, and N. Pohl, "Gesture classification with handcrafted micro-Doppler features using a FMCW radar," in *Proc. IEEE MTT-S Int. Conf. Microw. for Intell. Mobility*, 2018, pp. 1–4.

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

[11] M. G. Amin, Z. Zeng, and T. Shan, "Hand gesture recognition based on radar micro-Doppler signature envelopes," in *IEEE Radar Conf.*, 2019.

[12] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[13] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 2, pp. 655–665, 2017.

[14] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with Soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. the 29th Annu. Symp. User Interface Softw. Technol.* ACM, 2016, pp. 851–860.

[15] 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. 22 902–22 913, 2019.

[16] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33 610–33 618, 2019.

[17] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125 623–125 633, 2019.

[18] 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, 2019.

[19] 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.

[20] Z. Zhang, Z. Tian, Y. Zhang, M. Zhou, and B. Wang, "u-DeepHand: FMCW radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6811–6821, 2019.

[21] 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, 2018.

[22] Z. Zhang, Z. Tian, M. Zhou, W. Nie, and Z. Li, "Riddle: Real-time interacting with hand description via millimeter-wave sensor," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.
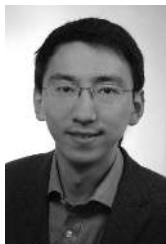
[23] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2012.

[24] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[25] Y. Vaezi and M. Van der Baan, "Comparison of the STA/LTA and power spectral density methods for microseismic event detection," *Geophysical J. Int.*, vol. 203, no. 3, pp. 1896–1908, 2015.

[26] M. Kronauge and H. Rohling, "New chirp sequence radar waveform," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 4, pp. 2870–2877, 2014.

[27] Y. Sun, T. Fei, and N. Pohl, "A high-resolution framework for range-Doppler frequency estimation in automotive radar systems," *IEEE Sensors J.*, vol. 19, no. 23, pp. 11 346–11 358, 2019.

[28] Y. Sun, T. Fei, and N. Pohl, "Two-dimensional subspace-based model order selection methods for FMCW automotive radar systems," in *Proc. IEEE Asia-Pacific Microw. Conf.*, 2018, pp. 1247–1249.

[29] M. I. Skolnik, *Radar handbook*, 2nd ed., New York: McGraw-Hill, 1990.

[30] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[31] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv: 1505.00853*, 2015.

[32] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symp. Operating Syst. Des. and Implementation*, 2016, pp. 265–283.

Xibo Li received the B.Sc. degree in mechanical engineering from Beijing Institute of Technology. Beijing, China. He is currently pursuing the M.Sc. degree in automotive engineering and transport at RWTH Aachen University, Aachen, Germany.

His current research interests include automotive radar signal processing, machine learning and sensor fusion.

Alexander Warnecke received the M.Sc. degree in electrical engineering and the Dr.-Ing. degree in degradation mechanisms NMC based lithium-ion batteries from RWTH Aachen University, Aachen, Germany, in 2012 and 2017, respectively. As a Research Associate at the Institute for Power Electronic and Electrical Drives (ISEA), he was involved in several projects related to ageing of lithiumion batteries at the chair for electrochemical energy conversion and storage systems. From 2015 to 2017, he was an Executive Manager at the Center for Ageing, Reliability, and Lifetime Prediction (CARL).

He is currently the head of the Research Institute for Automotive Electronics (E-LAB), HELLA GmbH & Co. KGaA, Lippstadt, Germany.

Ernst Warsitz received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Paderborn University, Paderborn, Germany, in 2000 and 2008, respectively. He joined the Department of Communications Engineering of the University of Paderborn in 2001 as a Research Staff Member, where he was involved in several projects related to single- and multi-channel speech processing and automated speech recognition. From 2007 he worked as a development engineer at HELLA GmbH & Co. KGaA, Lippstadt, Germany, in the field of signal processing algorithms for automotive radar systems. He is currently the head of the Radar Signal Processing and Signal Validation Department at HELLA GmbH & Co. KGaA, Lippstadt, Germany.

Yuliang Sun (GSM'18) received the B.Eng. degree in mechatronic engineering from Tongji University, Shanghai, China, and Aschaffenburg University of Applied Sciences, Germany, in 2014, and the M.Sc. degree in electrical and computer engineering from Technical University of Munich, Germany, in 2017. He is currently pursuing the Dr.-Ing. degree in electrical engineering with the Department of Integrated Systems, Ruhr University Bochum, Germany and the Research Institute for Automotive Electronics (E-LAB) in collaboration with HELLA GmbH & Co. KGaA, Lippstadt, Germany. His research interests are automotive radar signal processing, radar-based human motion recognition and machine learning.

He received the Best Student Paper Award at IEEE International Radar Conference in 2020.

Nils Pohl (GSM'07-M'11-SM'14) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Ruhr University Bochum, Bochum, Germany, in 2005 and 2010, respectively. From 2006 to 2011, he was a Research Assistant with Ruhr University Bochum, where he was involved in integrated circuits for millimeter-wave (mm-wave) radar applications. In 2011, he became an Assistant Professor with Ruhr University Bochum. In 2013, he became the Head of the Department of mm-wave Radar and High Frequency Sensors with the Fraunhofer Institute for High Frequency Physics and Radar Techniques, Wachtberg, Germany. In 2016, he became a Full Professor for Integrated Systems with Ruhr University Bochum. He has authored or coauthored more than 100 scientific papers and has issued several patents. His current research interests include ultra-wideband mm-wave radar, design, and optimization of mm-wave integrated SiGe circuits and system concepts with frequencies up to 300 GHz and above, as well as frequency synthesis and antennas.

Prof. Pohl is a member of VDE, ITG, EUMA, and URSI. He was a co-recipient of the 2009 EEEfCom Innovation Award, the 2012 EuMIC Prize, and the 2015 Best Demo Award of the IEEE Radio Wireless Week, and a recipient of the Karl-Arnold Award of the North Rhine-Westphalian Academy of Sciences, Humanities and the Arts in 2013 and the IEEE MTT Outstanding Young Engineer Award in 2018.

Tai Fei (S'12-M'19) received the B.Eng. degree in telecommunication engineering from Shanghai Maritime University, Shanghai, China, in 2005 and the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology from Darmstadt University of Technology (TUD), Darmstadt, Germany, in 2009 and 2014, respectively. From 2009 to 2012, he worked as a Research Associate with the Institute of Water-Acoustics, Sonar-Engineering and Signal-Theory at Hochschule Bremen, Bremen, Germany, in collaboration with the Signal Processing Group at TUD, Darmstadt, Germany, where his research interest was the detection and classification of underwater mines in sonar imagery. From 2013 to 2014, he worked as a Research Associate with Center for Marine Environmental Sciences at University of Bremen, Bremen, Germany.

Since 2014, he has been working as a development engineer at HELLA GmbH & Co. KGaA, Lippstadt, Germany, where he is mainly responsible for the development of reliable signal processing algorithms for automotive radar systems.