A Trajectory-based Gesture Recognition in Smart Homes based on Ultra-Wideband System

Anna Li, Eliane Bodanese, Tianwei Hou, Kaishun Wu, Fei Luo

***Abstract*—In this paper, we propose a cost-effective ultra-wideband (UWB) system for gesture recognition in a smart home environment, where the interference and the device selection issues can be beneficially solved. In the proposed UWB system, the gesture trajectories obtained by positioning are employed for recognizing human gestures instead of directly using wireless signals. To this end, we first collect the datasets of four different fine-grained gesture activities. Then, we integrate the squeeze-and-excitation block (SE) into the Convolutional Neural Network (CNN) seamlessly for gesture recognition, for convenience, namely the SE-Conv1D model. We compare the accuracy of various classifiers, which are support vector machine (SVM), K-nearest neighbor (KNN) and random forest (RF). All of these activities are correctly recognized with an accuracy of over 95%, among which our proposed SE-Conv1D model achieves the best accuracy of 99.48%. Finally, We implement our system to perform a case study, demonstrating that our proposed UWB-based system achieves higher recognition accuracy in real-time in a smart home environment compared to the previous contributions. We have also publicly archived our UWB gesture datasets, which may have a number of important implications for future practice.**

***Index Terms*—Gesture Recognition, Smart home, squeeze-and-excitation, Trajectory, Ultra-Wideband**

I. Introduction

Gesture recognition is one of the most critical sub-topics in human activity recognition (HAR), and plays a key role in the development of multiple applications, including smart home, health care, and virtual reality [1]. Gesture recognition is able to remote control the devices without physical contact, which is convenient and efficient for users. Recent developments in gesture recognition have heightened the need for smart homes, e.g., SeleCon [2], which has attracted great attention in using gestures to control smart devices.

A. Limitation of State of the Art

Previous attempts for gesture recognition utilize sensing modalities containing cameras, audio-based approaches, Wi-Fi technique, radio-frequency identifications (RFID), and Bluetooth techniques [3]–[5]. They suffer from inherent drawbacks, including privacy leakage, inconvenience, as well as limited sensing range and interference. For example, vision-based approaches have to deal with well-known environmental challenges, where the line-of-sight (LoS) is strictly required between the camera and users [6]. In addition, the feasibility of the vision-based recognition is impacted by the variability in brightness, contrast, and exposure. It is worth noting that both vision-based and speech-based recognition approaches violate users' privacy because the recorded video and audio data contentiously release to the remote cloud servers. To overcome privacy issues, Wi-Fi, RFID, and Bluetooth techniques are more applicable. However, the Wi-Fi technique is limited by the low spatial resolution, signal strength, multi-path reflections, as well as electromagnetic interference [7]. Bluetooth and RFID are greatly restricted to the short-range sensing capability. Recently, research on radar-based gesture recognition [8]–[11] has been considered as an alternative for overcoming the problems as mentioned above. Radar-based recognition has no privacy and LoS issues, which stands as a potential solution for fine-grained gesture recognition. However, radar-based recognition also suffers from interferences from the other devices and the multi-path effects, which dramatically decrease the signal-to-interference-plus-noise-ratio of the wireless signals [12]. Most of the solutions utilize digital signal processing techniques for mitigating co-channel interference. However, it is still hard to distinguish the interference for gesture recognition in practice.

Another challenge is previous studies of gesture recognition have not perfectly tackle the device selection problem, especially for smart home applications [13], [14]. Only few existing gesture recognition solutions can select a device and control it without increasing the tags for each device [2]. In most of studies, different gestures are assigned to different devices in smart home. However, assigning semantic tags for each device, such as 'light 1' or 'light 2' could burden the users. With the increasing number of devices in smart homes, this process becomes cumbersome. Therefore, it is natural to ask whether we can simultaneously control and select a specific device without defining massive gestures.

B. Motivations

To solve mentioned problems, we need a practical solution that can be applied to smart homes. We proposed a low-cost system based on Ultra-Wideband (UWB) technology, which is the radio with a wide bandwidth ($\geq$500 MHz). Although some researchers have investigated it [15]–[17], they put emphasis on UWB's high-frequency pulse signals, which can be expressed in the following formats: time-amplitude, range-amplitude, the time-range, time-Doppler, range-Doppler frequency/time-Doppler speed, and time-frequency. However, these existing studies by using UWB radars do not satisfactorily address interference problems well. In addition, if gesture

Anna Li and Eliane Bodanese are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: anna.li@qmul.ac.uk, eliane.bodanese@qmul.ac.uk).

Fei Luo and Kaishun Wu are with the School of Computer Science, Shenzhen University, Shenzhen 518000, China. (email: luofei2018@outlook.com, wu@szu.edu.cn).

Tianwei Hou is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China. (email: twhou@bjtu.edu.cn).

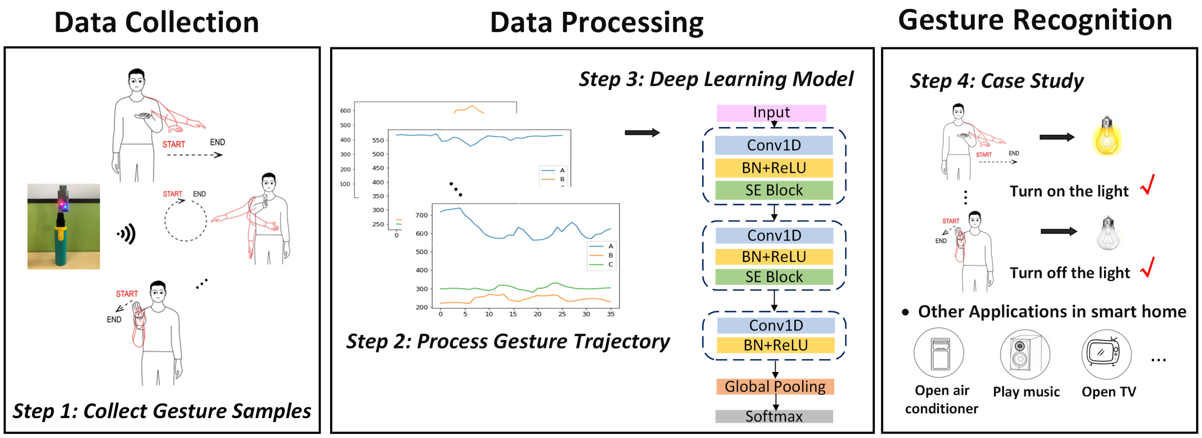


Fig. 1: Architecture of the proposed UWB system.

verification is performed at a distance or direction that is not used for training, the accuracy may be reduced, thereby limiting the real-world applications.

Different from the previous study, in this paper, we focused on the high localization accuracy of UWB, which is regarded as one of the most accurate and promising technologies. We proposed a new solution that only uses gesture trajectories instead of other spectrums to recognize human gestures using UWB technology. The trajectory of human gestures collected by UWB is a sequence of coordinates, which contains both spatial and temporal information. We developed a deep learning model to recognize human gestures and use different machine learning algorithms to verify our hypothesis. In this way, we are able to solve the problem of interference and improve the performance of our proposed system. In addition, with our proposed system, we can solve device selection for smart homes.

C. Contributions

A trajectory-based solution for gesture recognition using the UWB system, which aims to overcome the constraints of the existing works and achieves nearly real-time recognition, was proposed for the first time in this paper. We propose a deep learning model for recognizing dynamic gestures and use different machine learning algorithms to verify our hypothesis. It's worth noting that this system is not just a radar sensor chip or a new signal processing algorithm. Instead, it is a complete end-to-end sensing system specifically designed for tracking and recognizing fine-grained gestures, as illustrated in Fig 1. The main contributions can be summarized as follows:

- We first proposed a novel concept of recognizing human activities only based on trajectories of activities using our proposed cost-effective UWB system. We collected datasets of four different fine-grained activities in a laboratory that simulates the smart home, e.g., turning on/off the light. In addition, we designed our experiment that use the same gesture to control different lights. Finally, trajectories of pre-defined gestures from different directions and distances were produced.
- We proposed a novel framework for gesture recognition, for convenience, namely the SE-Conv1D model, which achieves excellent gesture recognition performance of 99.48% overall accuracy. We compared our results to other different machine learning algorithms, which are support vector machine (SVM), K-nearest neighbor (KNN), and random forest (RF). All of these activities are correctly classified with an accuracy of over 95% by three learning models, in which the SVM algorithm is shown to yield a high 98.12% classification accuracy.
- We prototype our system to interact with appliances in practical smart homes. It proves that our proposed system is a complete end-to-end sensing system specifically designed for tracking and recognizing fine-grained gestures. The superiority of our proposed solution is both interference-free and works robustly against changes in distance or direction, which means it is more reliable in real-world applications. In addition, our proposed system provides a practical method of IoT device selection for smart homes only using gestures.

D. Organization

The rest of this paper is organized as follows. Section II presents the related work, in which we will review UWB-based gesture recognition, learning algorithms, and trajectory-based solutions. Section III gives a detailed review of the UWB technology, double-sided two-way ranging, and our learning models. Section IV shows the experimental setup. The experimental results are illustrated in section V. Section VI discusses the future work and concludes this paper.

II. Related Work

This section presents the related work concerning UWB-based gesture recognition, learning algorithms, and trajectory-based recognition with different sensors.

A. Radar-based Gesture Recognition

Many studies have investigated radar-based systems for gesture recognition [10], [18], [19]. For instance, Sakamoto *et* *al.* [18] utilized a 2.4-GHz continuous wave (CW) Doppler radar for the automatic recognition of human gestures and achieved the overall accuracy exceeding 90%. However, a drawback of an unmodulated CW radar is that it is unable to measure distance. To solve this problem, the frequency modulated continuous wave (FMCW) radar, of which the signals contain both range and Doppler information, was proposed for gesture recognition. In [19], Peng *et al.* used a 24-GHz FMCW radar to recognize human gestures at different ranges. However, in real-world applications, purely range-based gesture recognition using radars may suffer from the interferences of the Doppler frequencies of nearby moving.

Recently, there has been renewed interest in UWB technology, which is regarded as one of the most accurate and promising technologies to provide high localization accuracy, high immunity against the multi-path problem, and low output power [20]. An early study for gesture recognition using UWB radar was in 2016, Park *et al.* [21] recognized human gestures by using impulse radio UWB radar. With the help of machine learning, they extracted features from the received signals and then achieved a total gesture recognition accuracy of nearly 100%. Khan [22] collected gestures based on the human hand and finger motions with the similar methodology in [21], which can be used to control different electronic devices inside a car. It can be taken as progress in real-world applications by using UWB radars. In [2], Alanwarwe *et al.* proposed a pointing approach to interact with different devices in smart homes, which uses a UWB equipped smartwatch. Their results demonstrate that it achieves 97% overall accuracy for gesture recognition. However, radar-based gesture recognition also suffers from interferences from the other devices and the multi-path effects. To date, there are few radar-based studies that have investigated the problem of equipment selection and interference simultaneously.

B. Trajectory-based Recognition

Trajectory-based methods provide a means of solving interferences [23]. Much of the previous research on trajectory-based HAR using cameras. Liang *et al.* [24] proposed a Long Short-Term Memory (LSTM) model to recognize trajectory and activity from videos. In [25], Bashir *et al.* presented novel classification algorithms for recognizing object activity using object motion trajectory. However, such approaches have failed to address the problem of small visual scope because the trajectories extracted from a video depend heavily on the azimuth and inclination of the camera. In addition, the privacy issue is still a much-debated question. Martin *et al.* [26] used Global Positioning System (GPS) trajectories to recognize the human mobility behavior. However, this solution cannot be used for fine-grained activities. To be noted that, there is little research in trajectory-based gesture recognition using a UWB-based system.

C. Learning Algorithms

In the last decade, machine learning techniques have been widely used in Gesture Recognition. In [27], Camgöz *et al.* proposed RF-based model for spotting and recognizing continuous human gestures. In [28], a weighted KNN was utilized to achieve real-time gesture recognition. In [29], a system that aimed to reconstruct gestures by measuring a user's tendon movements was proposed.

TABLE I: The Parameter settings of the proposed UWB system.

Parameters	Values
Centre Frequency	3 GHz
Bandwidth	500 MHz
Bit Rate	110 kbps
Range	0-9 m

Since it was reported in 2018, squeeze and excitation (SE) blocks have been attracting a lot of interest in different fields [30]–[32]. In [30], Rundo *et al.* incorporated SE blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. Their research reveals that SE blocks provide excellent intra-dataset generalization in multi-institutional scenarios. In [31], an end-to-end intelligent recognition of epileptic electro-encephalogram (EEG) seizure detection framework was proposed by using a novel channel-embedding spectral-temporal SE network (CE-stSENet). In [32], a new network architecture based on the faster region-based convolutional neural network (R-CNN) was proposed to further improve the detection performance by using SE mechanisms, which were used for ship detection in Synthetic aperture radar (SAR) images. As the existing works using SE blocks have achieved some signs of success, we incorporate the proposed SE blocks in our 1 dimensional CNN to improve the accuracy of gesture recognition for smart home applications.

III. Methodology

This section gives an overall review of the UWB technology-based on DWM1000[1], the theory of double-sided two-way ranging (DS-TWR) and SE block, as well as our proposed models.

A. The UWB Technology based on DWM1000

According to the U.S. Federal Communications Commission, UWB technology, which is the radio with a wide bandwidth ($\geq$500 MHz). UWB technology is one of the most potent choices for critical positioning applications that require highly accurate results [33]. In this paper, we utilized one type of sensor (DWM1000 module) to implement our method, which is compliant with the IEEE 802.15.4-2011 UWB standard. The module size is 54 mm × 20 mm × 2.9 mm, and it integrates antennas, all RF circuits, power management, and clock circuitry in one module. As shown in Table I, the center frequency is 3 GHz, bandwidth is 500 MHz, and the bit rate is 110 kbps.

[1] The DWM1000 module is based on DecaWave's DW1000 Ultra-Wideband (UWB) transceiver IC. For further information on this, please refer to www.decawave.com.

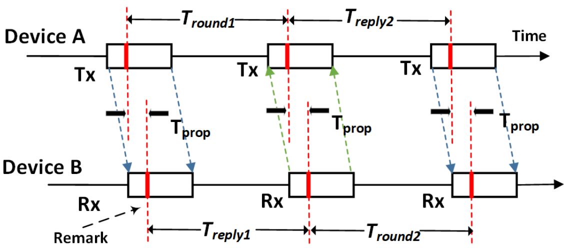


Fig. 2: The theorem of double-sided two-way ranging system.

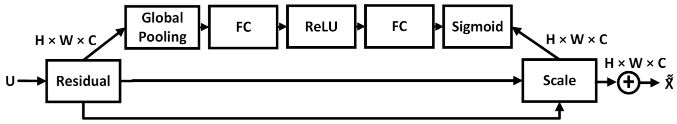


Fig. 3: The schema of SE module.

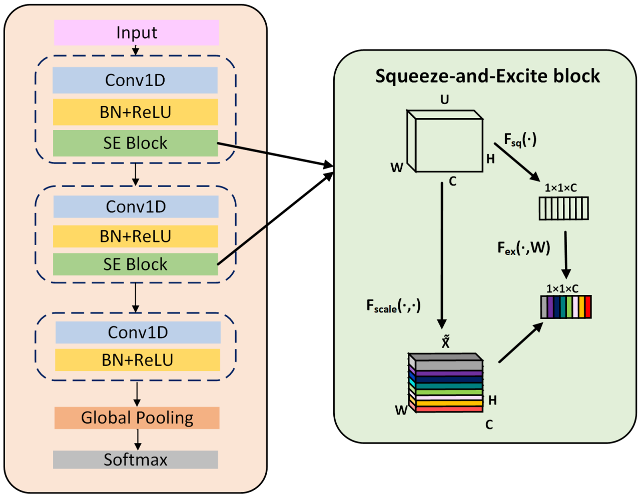


Fig. 4: The proposed SE-Conv1D model.

B. Double-sided Two-way Ranging (DS-TWR)

UWB ranging is suitable for real-time locating systems (RTLS) [34]. In addition, the wide frequency bandwidth allows for high-resolution channel impulse response estimation, along with accurate time-of-flight (ToF) measurements in a dense multi-path environment [35]. The system used in this paper is based on the double-sided two-way ranging (DS-TWR) [36], in which two round trip time measurements are used and combined to give a ToF result which has a reduced error even for quite long response delays.

In this paper, three DWM1000 modules were configured as anchors (receivers) while another was configured as a tag (transmitter). All the devices were placed at the same height for localization. The trilateration solver gives two solutions equidistant from each side of the plane of the anchors, which is assumed to be all horizontal. The core of a DS-TWR exchange is shown in Fig. 2, where device A initiates the first round trip measurement to which device B responds, after which device B initiates the second round trip measurement to which device A responds, completing the full DS-TWR exchange. Each device precisely timestamps the transmission and reception times of the messages. The remark is the part of the frame that is assumed to be time-stamped at the device antennas. The resultant ToF estimate, $\hat{T}_{prop}$, which is the propagation time of the message between tag and anchors, is calculated:

$$\hat{T}_{prop} = \frac{(T_{round1} \times T_{round2} - T_{\text{reply}1} \times T_{\text{reply}2})}{(T_{round1} + T_{round2} + T_{\text{reply}1} + T_{\text{reply}2})}. \qquad (1)$$

Finally, if we assume the speed of the radio waves through the air is equal to the speed of light c, then the distance between the anchor and tag can be expressed by:

$$Distance = c \times TOF. \qquad (2)$$

C. Squeeze-and-Excitation Block

Recently, *squeeze-and-excitation* (SE) blocks [37] have become an integral part of models, which are used to rescale the input feature map to highlight useful channels. Hence, these blocks able to be lightweight to increase the model complexity and computation time, and ease training of the network by improving gradient flow. The general schema of SE module is shown in Fig. 3. The input feature map $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_c}]$ is considered as a combination of channels $\mathbf{u}_i \in {}^{\mathrm{W}\times H}$, in which $\mathbf{u}_{_c}$ can be calculated as follows:

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_{_c}^{s} * \mathbf{X}^s, \qquad (3)$$

where $\mathbf{v}_{_c}^{s}$ refers to 2D spatial kernel, v_c refers to single channel, $\mathbf{X}$ refers to corresponding channel, and * refers to the convolution operation. With the help of a global average pool, the *squeeze* operation is able to generate channel-wise statistics, $z \in \mathbb{R}^C$, by utilizing the contextual information outside the local receptive field. $\mathbf{z}_{_c}^{s}$ is calculated by computing $\mathbf{F}_{sq}(\mathbf{u}_c)$, which is the channel-wise global average over the spatial dimensions $W \times H$. The c-th element of $\mathbf{z}$ is calculated by:

$$\mathbf{z}_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{\mathrm{W}\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} u_c(i,j). \qquad (4)$$

For temporal sequence data, the channel-wise statistics is generated by shrinking $\mathbf{U}$ through the temporal dimension T, where the c-th element of $\mathbf{z}$ is calculated by:

$$\mathbf{z}_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{T}\sum_{t=1}^{T} u_c(t). \qquad (5)$$

The aggregated information obtained from the *squeeze* operation is followed by the *excitation* operation, which aims to capture the channel-wise dependencies. To meet these criteria, a simple gating mechanism is employed with a sigmoid activation:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1 z)), \qquad (6)$$

where $\mathbf{F}_{ex}$ refers to a neural network, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ are learnable parameters of F_{ex} , σ refers to the Sigmoid activation function, δ refers to the ReLU activation function, and r is the reduction ratio. To limit model complexity and aid generalisation, $\mathbf{W}_1$ and $\mathbf{W}_2$ are utilized.

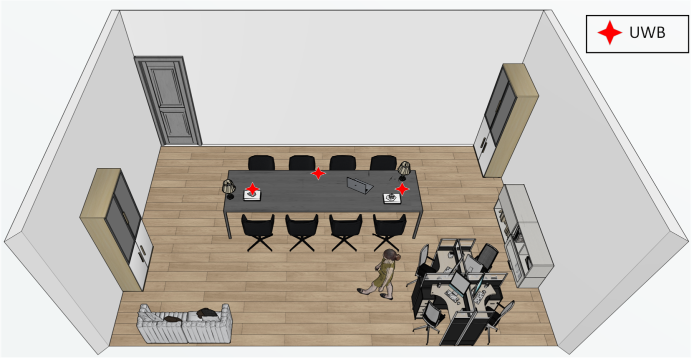


Fig. 5: Experimental scenario. Office has a size of 4.5 m × 4 m with one sofa, two tables, and three bookcases. The UWB devices were configured on the table.

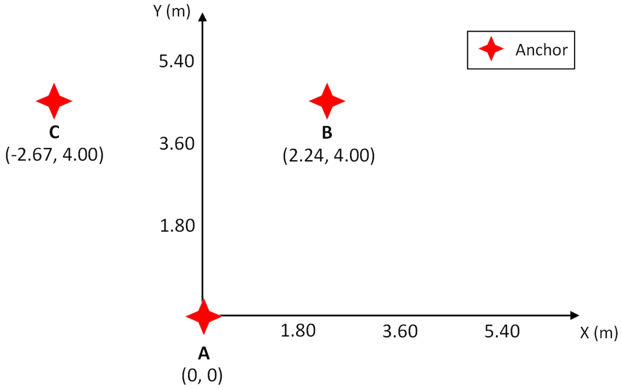


Fig. 6: The location of device configuration.

The final output of the block is achieved by rescaling $\mathbf{U}$ as follows:

$$\tilde{\mathbf{X}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, \mathbf{s}_c) = \mathbf{s}_c \cdot \mathbf{u}_c, \tag{7}$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, ..., \tilde{\mathbf{X}}_c]$ and $\mathbf{F}_{scale}$ is the channel-wise multiplication between the feature map $\mathbf{u}_c \in \mathbb{R}^T$ and the scale s_c.

D. Our Proposed Network Architecture

We integrate the squeeze-and-excitation block (SE) block into the *Convolutional Neural Network* (CNN) models to enhance recognition accuracy. The structure and learning method of the proposed framework for gesture recognition, for convenience, namely the SE-Conv1D model, as illuminated in Fig. 4. The fully convolutional block contains three temporal convolutional blocks, which are used as feature extractors. The SE-Conv1D network is comprised of 2 blocks of (128, 256, 128) filters for all models, with kernel sizes of 16, 3, and 5, respectively. Each convolutional layer is succeeded by the batch normalization layer and the ReLU activation function. A global average pooling layer follows the final temporal convolutional block. During the training phase, All the networks are trained using sparse categorical cross-entropy [38]. In [39], the reduction ratio r was set to a value of 16 for the task of time series classification. To find its optimal setting in our application, we perform experiments with different values of r from 2 to 16. Finally, for all SE blocks, we set r to 16. The batch size is 60 for 25 epochs. We use the Adam optimizer [40], with the final learning rate set to 0.001.

E. Other machine learning classifiers for comparisons

We use three different machine learning classifiers, which are taken as the baselines to make a comparison with our proposed SE-Conv1D network.

1) Support Vector Machine (SVM): Support vector machines (SVM) are learning algorithms, which are able to select the hyperplanes that maximize the distance between the nearest training samples and the hyperplanes [41]. In this paper, we use the radial basis function (RBF) kernel function [42] to calculate the distance by using the equation as follows:

$$K(x_i, y_i) = \exp(-\gamma \parallel x_i - y_i \|^2), \gamma > 0, \tag{8}$$

where γ is the kernel parameter, and cross-validation is used to tune the hyperparameters. In this paper, two hyperparameters (C, γ) in SVM-RBF were specified manually. Given a hyperparameter space C: [0, 20], γ: [0, 1.0E - 5], a different pair of parameters will be selected from the hyperparameter space by using the cross-validation in each training and validation epoch to build the SVM-RBF model. After training, we can obtain the optimal parameters for SVM-RBF model. Finally, in this paper, we find that when C is 10, and γ is 1.0E - 4.3, the performance is best.

2) K-Nearest Neighbor (KNN): K-nearest neighbor (KNN), which follows an assumption that similar things are closed to each other [43]. The Euclidean distance is commonly used for continuous variables to calculate the distance in KNN, which can be calculated by [44]:

$$Dist(S_c, S_i) = \sqrt{\sum_{p=1}^{M} (f_p^c - f_p^i)^2}. \tag{9}$$

We need to calculate the distance between each labeled sample S_c $(1 \leq c \leq N)$ and S_i to find k closest samples to S_i. Cross-validation [45] is used to select a suitable value of k, which can minimize the overall distance between those k nearest labeled and the unlabeled samples. After this step, the unlabeled samples will be classified to the class label based upon a majority vote from the k nearest labeled samples. In this paper, we perform in-car HAR by using kNN with k = 2.

3) Random Forest (RF): Random Forest (RF) is a holistic learning method of classification and regression by constructing a collection of decision trees. Gini impurity is the default metric in a decision tree, which is a method of how often those randomly chosen elements will be labeled incorrectly, if they are randomly labeled according to the distribution of labels in the subset. To calculate Gini impurity with J classes, assume that $i \in \{1, 2, ..., J\}$, and let p_i be the fraction of items labeled with class i in the set:

$$I_{G(P)} = \sum_{i=1}^{J} p_i \sum_{k \neq 1}^{p_k} = 1 - \sum_{i=1}^{J} p_i^2. \tag{10}$$

RF is trained by constructing a set of decision trees with the number of trees specified by a hyperparameter N. Thus, the

TABLE II: The description of pre-defined gestures.

ID Number	Gesture Activity	Descripition	Case study
0	Swipe right	The gesture of swiping the right hand to the right	Turn on the light
1	Up and down	The gesture of Lifting the right hand from bottom to top and then put it down	Coloring
2	Circle clockwise	The gesture of rotating the right hand for a circle clockwise.	Turn on all lights
3	Push	The gesture of pushing the right hand outwards perpendicular to the ground	Turn off all lights
4	Others	Daily behaviors, e.g., walking and sitting	/

random forest consists of N trees. For a new dataset, each item of the dataset is input to each of the N trees. In this paper, a RF built with 110 decisions trees is utilized.

IV. Evaluation Setup

In this section, experimental setup, including device configuration, the experimental environments, gesture sets, participants, and data processing, are presented.

A. Data Measurement

The UWB data collection was conducted in an office in Shenzhen University, where we simulated the real smart home scenarios, which is shown in Fig. 5. In this experiment, three DWM1000 modules were configured as anchors (receivers) while another was configured as a tag (transmitter). The tag was connected to the human body, and radar data were collected in real-time. As shown in Fig. 6, three anchors were deployed, namely anchors A, B, and C. Anchors B and C were located 89 cm above the ground, and anchor A was located 33 cm above the ground. These two anchors were kept fixed at a separation distance of 4.91 meters from each other.

We predefined a set of gestures to represent different user commands in smart home, including four continuous gestures ‘swipe right’, ‘up and down’, ‘circle clockwise’ and ‘push’. In order to explore how our proposed system will be incorrectly triggered by daily behaviors, e.g., ‘walking’ and ‘sitting’, we asked volunteers to perform the daily behaviors for 10 minutes. We defined them as ‘Others’. The standard for each class is shown in TABLE II. The data were collected from three volunteers: two males and one female. They were unpaid volunteers recruited from different departments of Shenzhen university. The experimenter demonstrated how each gesture should be performed one by one before starting the data collection. It is worth noting that all activities were performed in a naturally different orientation, as would be the case in the real-world scenario. When the instruction ‘start’ was given, a volunteer, who faces different directions at different positions, performs a pre-defined gesture, each of which was performed around 100 times.

B. Data Processing

Our final dataset comprises 19121 samples in total, including 3145 samples of ‘swipe right’, 3291 samples of ‘up and down’, 4737 samples of ‘circle clockwise’, 3948 samples of ‘push forward’, 4000 samples of ‘others’, respectively. Each window contains 18-25 points, which are extracted from gesture trajectories. The choice of window size is essential for accurate gesture recognition. If the size is too small, the signals of a human gesture cannot be entirely captured by the window. On the contrary, if the window size is too large, signals of two or more human gestures can be included. After testing, we then segmented them using a sliding window with an average length of 20 location points.

V. Results

In this section, firstly, the evaluations of the test dataset by using different learning algorithms are presented. In addition, we compare our results with the related work. Finally, the case study is presented.

A. Evaluation Metrics

The metrics used to evaluate recognition performance from different perspectives in this paper are: 1) Overall classification accuracy (OA); 2) Recall; 3) F1-score and 4) Normalized confusion matrix. The metrics used to evaluate performance of our system in this paper are defined as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{12}$$

$$\text{F1} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{13}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. During the training process, we divided the trajectory samples into two segments: 85% for training, 15% for testing.

B. The results of the proposed SE-Conv1D network

The training loss and validation accuracy of the SE-Conv1D network are shown in Fig. 8. By using our proposed SE-Conv1D network, the achieved overall accuracy is 99.48% and loss less than 0.02. A confusion matrix is commonly utilized for analyzing recognition performance. As shown in the normalized confusion matrix of Fig. 9, all gestures are correctly classified with nearly 100%, except the gesture ‘up and down’. 1% of ‘up and down’ samples were misclassified into ‘push’, while the rest of those were misclassified into ‘circle clockwise’.

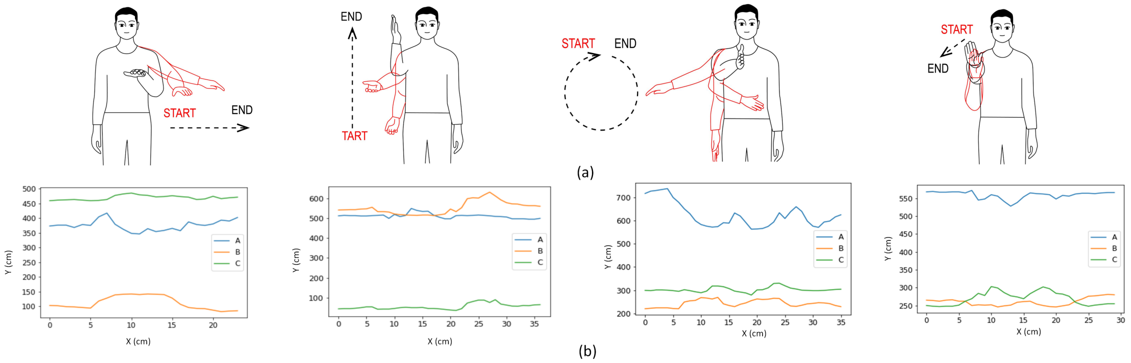


Fig. 7: (a) The design of four potential human gesture activities in the car. From left to right: ‘swipe right’, ‘up and down’, ‘circle clockwise’, ‘push’. (b) Trajectories are randomly chosen from different directions and distances produced by UWB. ‘A’, ‘B’, and ‘C’ refers to station A, B, and C.

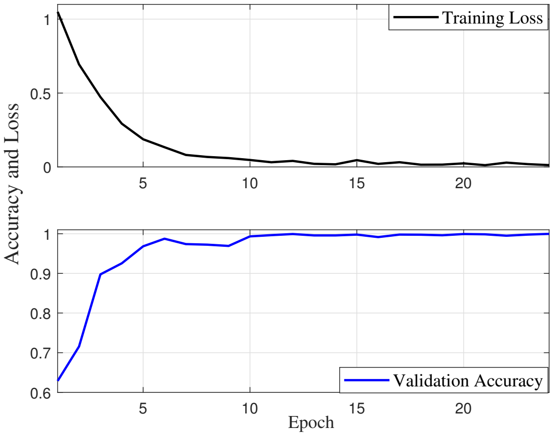


Fig. 8: Validation loss and accuracy of the proposed LSTM and TCN.

TABLE III: Comparison of the classification models.

	OA (%)	Recall (%)	F1 (%)
SVM-RBF	98.12	98.05	98.04
RF	96.51	96.22	96.42
KNN	95.47	95.54	95.48
SE-Conv1D	99.48	99.45	99.49

C. *Results of different Baseline approaches*

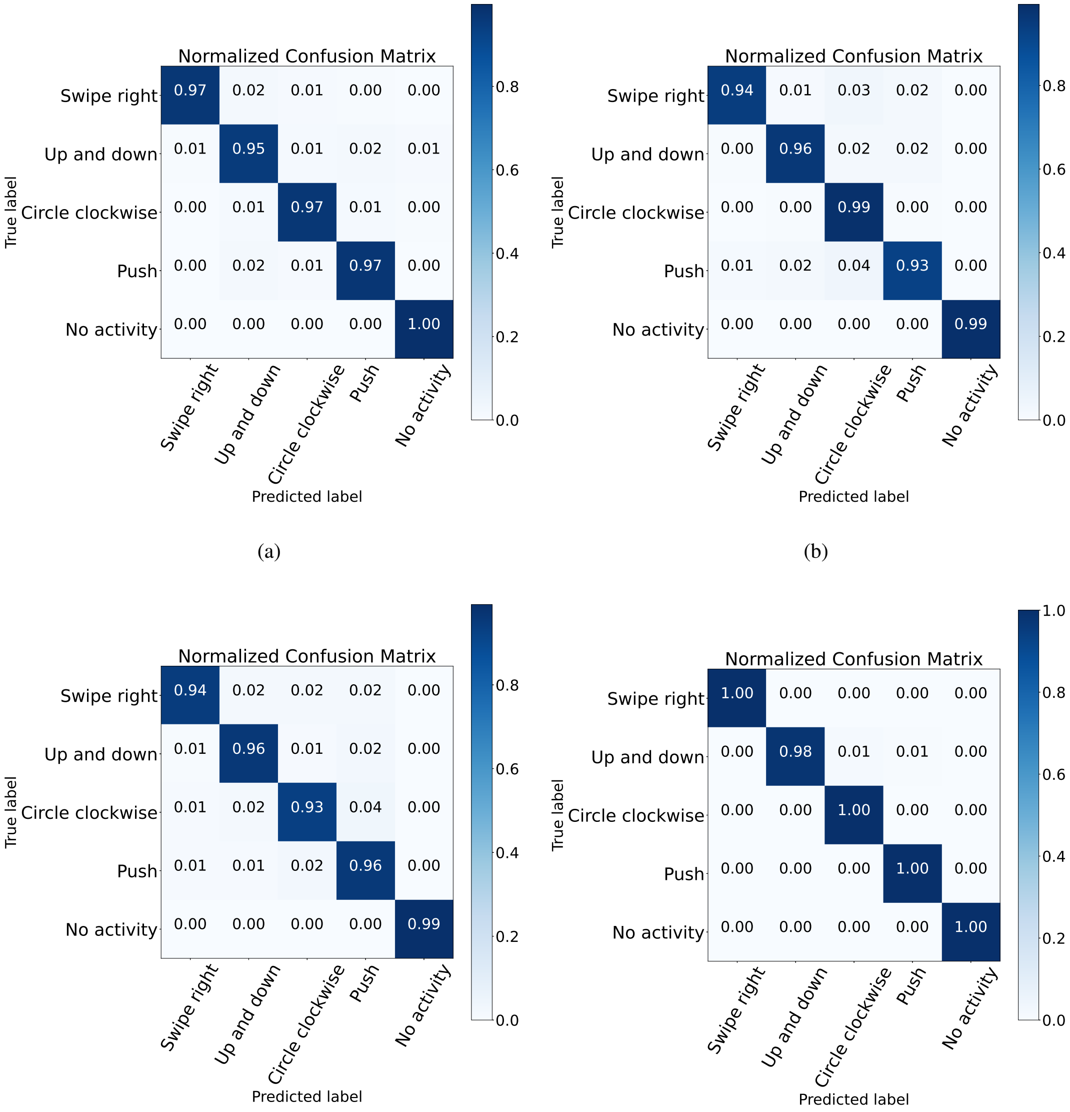
As shown in Table III, the achieved overall accuracies by SVM, RF, kNN are 98.12%, 96.51%, and 95.47%, respectively. SVM performs best among these three algorithms. As shown in Fig. 9, the normalized confusion matrix of the human gesture activities are correctly classified with above 90% by three learning methods. SVM achieves the highest performance in ‘swipe right’ (97%), ‘push’ (97%), and ‘no activity (100%). SVM achieves poor performance inactivity ‘up and down’, while 1% of samples have been misclassified into ‘swipe right’, ‘circle clockwise’ and ‘no activity, respectively, and 2% of samples have been misclassified into ‘push’. This is because compared with other activities, some trajectory samples from these classes have a similar pattern to other activities. RF shows better performance on activity ‘up and down’ and ‘circle clockwise’. Especially for activity ‘circle clockwise’, the OA is nearly 100%. However, for activity ‘push’, the OA is only 93%. In conclusion, our proposed SE-Conv1D network has a huge superiority over than other three baseline approaches.

D. *Analysis About Time Consumption*

The hardware platform is a laptop with an Intel(R) Core(TM) i7-10510U CPU inside, and the CPU clock frequency and the memory size are 2.3 GHz and 16.0 GB, respectively. The software platform is python with the Tensorflow backend, and the operating system is Windows 10. The SE-Conv1D network consumes around 7min30s for training and testing all gestures. In real-world applications, gesture recognition will be asked to be real-time processing. Since the running time for classifying one hand gesture by the proposed SE-Conv1D network is only 0.023 s, it is promising to achieve real-time processing in practical applications.

E. *Comparison with Related Work*

Table III presents the performance comparison of the proposed approach with the existing state-of-the-art methods. Ahmed *et al.* [46] perform gesture recognition within cars by using impulse UWB radar with CNN, which achieves an overall accuracy of approximately 97%. Khan *et al.* [47] proposed a novel gesture recognition algorithm, which is able to achieve the best performance of 97%. Maitre *et al.* recognized activities of daily living from UWB radars by using the deep learning methods. In [48], Wang *et al.* used the Soli radar developed by Google [49] along with deep-learning to classify hand gestures. However, all of these works using UWB technology cannot solve the problem of interference from the ground surrounding objects and multi-path effects, as



(a)

(b)

(c)

(d)

Fig. 9: Normalized confusion matrix. (a) Normalized confusion matrix using SVM-RBF. (b) Normalized confusion matrix using RF. (c) Normalized confusion matrix using kNN. (d) Normalized confusion matrix using SE-Conv1D.

TABLE IV: Comparison with Related Works

Study	Algorithm	Accuracy
This paper	SVM	98.12%
Khan *et al.* [47]	K-means	97%
Ahmed *et al.* [46]	CNN	97%
Maitre *et al.* [15]	Stacked-LSTM	97%
Vu *et al.* [50]	HMM	95%
Proposed	**SE-Conv1D**	**99.48%**

well as the device selection problem. In addition, our proposed SE-Conv1D network gives a better accuracy rate in trajectory-based human gesture recognition than other standard gesture-based learning algorithms. Our trajectory-based activity recognition is a potentially promising method, which significantly impact the real-world application.

F. Case Study

We prototype our system to interact with appliances in practical smart homes. We realize its functions responding to the following pre-defined gestures, which is also shown in TABLE II:

- Swipe right: Turn on the light.
- Up and down: Coloring.
- Circle clockwise: Turn on all lights.
- Push: Turn off all lights.

The detailed operation is formulated in Alg.1, and the dynamic trajectory collection process for each person is shown in Fig. 10. We control the light with the distance between two lights instead of adding more pre-defined gestures, which may bring continence to real-world applications. The detailed demonstration is shown in the link.

Algorithm 1: The proposed Gesture control algorithm

```
Input: Location results; Learning parameter
Output: Turning on/off Bulb1 and/or Bulb2; changing
        Bulb1 and/or Bulb2 to different colors
Initialize the locations of all BSs and bulbs
foreach A volunteer has been detected do
    while Gesture detection: Gesture0 do
        Turn on the nearest Bulb;
    end
    while Gesture detection: Gesture1 do
        Change the color of the nearest Bulb;
    end
    while Gesture detection: Gesture2 do
        Turn on both Bulb1 and Bulb2;
    end
    while Gesture detection: Gesture3 do
        Turn off both Bulb1 and Bulb2;
    end
end
```

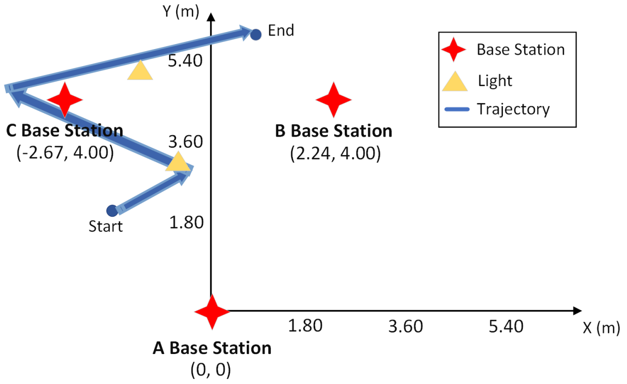


Fig. 10: The tested experiment with dynamic trajectory.

In order to quantify the user experience for the future improvement of our system, We made an extensive user study. For this, we recruit 12 volunteers to try it and quantize their experience using our designed questionnaires. We record the user experiences based on the following aspects: convenience, flexibility, accuracy, robustness, interactivity. For each item, users are asked to grade their perception into the following levels, i.e., ’A’ refers to Excellent, ’B’ refers to Very good, ’C’ refers to good, ’D’ fair, or ’E’ does poorly. We acquired the following options and summarized them. All users find that our proposed UWB-based system is more convenient than the traditional method, e.i., Using smartphones, because it does not ask the users to open the app to choose which device they want to open. More than 80% of users believe that our solution is more difficult to be triggered by mistake in daily usage than the voice-control solution. Hence, compared with the current popular solutions for smart homes, our solution has a more convenient user experience based on similar accuracy. Therefore, our solution has the potential to enable a convenient smart home gesture-based application.

VI. Conclusion and Future work

In this paper, a novel solution to perform gesture recognition by using a low-cost and -power UWB-based system was proposed for the first time. We collected datasets of five different fine-grained gesture activities. Our proposed SE-Conv1D model achieved an excellent result of 99.48% overall accuracy, which is superior to the results achieved by SVM, RF, and KNN. We compared this article to the state-of-the-art approach for activity recognition using UWB. Our proposed networks outperform the state-of-the-art method. Compared with previous research, the superiority of this study is that our solution is interference-free and works robustly against changes in distance or direction, which means it is more reliable in real-world applications. It proves that our proposed novel approach is able to deliver a high recognition accuracy in nearly real-time. While we are very optimistic about the current capabilities of our proposed system, we admit that there are still exist the following limitations in our current system. Our system requires the user to wear equipment with UWB, and it cannot be used by more than one user simultaneously. Hence, in our future work, we consider improving our system to support multiple people recognition.

References

[1] H. Liu, Y. Wang, A. Zhou, H. He, W. Wang, K. Wang, P. Pan, Y. Lu, L. Liu, and H. Ma, “Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–28, 2020.

[2] A. Alanwar, M. Alzantot, B.-J. Ho, P. Martin, and M. Srivastava, “Selecon: Scalable iot device selection and control using hand gestures,” in *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, 2017, pp. 47–58.

[3] M. Wang, Z. Yan, T. Wang, P. Cai, S. Gao, Y. Zeng, C. Wan, H. Wang, L. Pan, J. Yu *et al.*, “Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors,” *Nature Electronics*, vol. 3, no. 9, pp. 563–570, 2020.

[4] P.-G. Jung, G. Lim, S. Kim, and K. Kong, “A wearable gesture recognition device for detecting muscular activities based on air-pressure sensors,” *IEEE Transactions on Industrial Informatics*, vol. 11, no. 2, pp. 485–494, 2015.

[5] H. Kang, Q. Zhang, and Q. Huang, “Context-aware wireless based cross domain gesture recognition,” *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[6] D. W. Hansen and Q. Ji, “In the eye of the beholder: A survey of models for eyes and gaze,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2009.

[7] M. Scherer, M. Magno, J. Erb, P. Mayer, M. Eggimann, and L. Benini, “TinyRadarNN: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars,” *IEEE Internet of Things Journal*, 2021.

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

[9] C. Ding, H. Hong, Y. Zou, H. Chu, X. Zhu, F. Fioranelli, J. Le Kernec, and C. Li, “Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6821–6831, 2019.

[10] B. Dekker, S. Jacobs, A. Kossen, M. Kruithof, A. Huizing, and M. Geurts, “Gesture recognition with a low power FMCW radar and a deep convolutional neural network,” in *2017 European Radar Conference (EURAD)*, 2017, pp. 163–166.

[11] S. K. Leem, F. Khan, and S. H. Cho, “Detecting mid-air gestures for digit writing with radio sensors and a CNN,” *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1066–1081, 2019.

[12] K. A. Hamdi, “On the statistics of signal-to-interference plus noise ratio in wireless communications,” *IEEE transactions on communications*, vol. 57, no. 11, pp. 3199–3204, 2009.

[13] P. Asadzadeh, L. Kulik, and E. Tanin, “Gesture recognition using rfid technology,” *Personal and Ubiquitous Computing*, vol. 16, no. 3, pp. 225–234, 2012.

[14] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, “Whole-home gesture recognition using wireless signals,” in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013, pp. 27–38.

[15] J. Maitre, K. Bouchard, C. Bertuglia, and S. Gaboury, “Recognizing activities of daily living from uwb radars and deep learning,” *Expert Systems with Applications*, vol. 164, p. 113994, 2021.

[16] M. Piriyajitakonkij, P. Warin, P. Lakhan, P. Leelaarporn, N. Kumchaiseemak, S. Suwajanakorn, T. Pianpanit, N. Niparnan, S. C. Mukhopadhyay, and T. Wilaiprasitporn, “Sleepposenet: Multi-view learning for sleep postural transition recognition using UWB,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1305–1314, 2021.

[17] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, “Hand gestures recognition using radar sensors for human-computer-interaction: A review,” *Remote Sensing*, vol. 13, no. 3, 2021.

[18] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke, and V. M. Lubecke, “Radar-based hand gesture recognition using I-Q echo plot and convolutional neural network,” in *2017 IEEE Conference on Antenna Measurements Applications (CAMA)*, 2017, pp. 393–395.

[19] Z. Peng, C. Li, J.-M. Muñoz-Ferreras, and R. Gómez-García, “An FMCW radar sensor for human gesture recognition in the presence of multiple targets,” in *2017 First IEEE MTT-S International Microwave Bio Conference (IMBIOC)*. IEEE, 2017, pp. 1–3.

[20] M. R. Mahfouz, C. Zhang, B. C. Merkl, M. J. Kuhn, and A. E. Fathy, “Investigation of high-accuracy indoor 3-d positioning using uwb technology,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 6, pp. 1316–1330, 2008.

[21] J. Park and S. H. Cho, “Ir-uwb radar sensor for human gesture recognition by using machine learning,” in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2016, pp. 1246–1249.

[22] F. Khan, S. K. Leem, and S. H. Cho, “Hand-based gesture recognition for vehicular applications using ir-uwb radar,” *Sensors*, vol. 17, no. 4, p. 833, 2017.

[23] F. Luo, S. Poslad, and E. Bodanese, “Temporal convolutional networks for multiperson activity recognition using a 2-d lidar,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7432–7442, 2020.

[24] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, “Peeking into the future: Predicting future person activities and locations in videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5725–5734.

[25] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, “Object trajectory-based activity classification and recognition using hidden markov models,” *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[26] H. Martin, D. Bucher, E. Suel, P. Zhao, F. Perez-Cruz, and M. Raubal, “Graph convolutional neural networks for human activity purpose imputation,” in *NIPS spatiotemporal workshop at the 32nd Annual conference on neural information processing systems (NIPS 2018)*, 2018.

[27] N. C. Camgöz, A. A. Kindiroglu, and L. Akarun, “Gesture recognition using template based random forest classifiers,” in *European Conference on Computer Vision*. Springer, 2014, pp. 579–594.

[28] Y. Liu, X. Wang, and K. Yan, “Hand gesture recognition based on concentric circular scan lines and weighted k-nearest neighbor algorithm,” *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 209–223, 2018.

[29] X. Liang, R. Ghannam, and H. Heidari, “Wrist-worn gesture sensing with wearable intelligence,” *IEEE Sensors Journal*, vol. 19, no. 3, pp. 1082–1090, 2019.

[30] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi *et al.*, “USE-Net: Incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets,” *Neurocomputing*, vol. 365, pp. 31–43, 2019.

[31] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, “Epileptic seizure detection in eeg signals using a unified temporal-spectral squeeze-and-excitation network,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 782–794, 2020.

[32] Z. Lin, K. Ji, X. Leng, and G. Kuang, “Squeeze and excitation rank faster R-CNN for ship detection in SAR images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 751–755, 2019.

[33] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[34] D. Zhang, L. T. Yang, M. Chen, S. Zhao, M. Guo, and Y. Zhang, “Real-time locating systems using active RFID for internet of things,” *IEEE Systems Journal*, vol. 10, no. 3, pp. 1226–1235, 2014.

[35] D. Dardari, A. Conti, U. Ferner, A. Giorgetti, and M. Z. Win, “Ranging with ultrawide bandwidth signals in multipath environments,” *Proceedings of the IEEE*, vol. 97, no. 2, pp. 404–426, 2009.

[36] P. Corbalán, G. P. Picco, and S. Palipana, “Chorus: UWB concurrent transmissions for GPS-like passive localization of countless targets,” in *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2019, pp. 133–144.

[37] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[39] F. Karim, S. Majumdar, H. Darabi, and S. Harford, “Multivariate lstm-fcns for time series classification,” *Neural Networks*, vol. 116, pp. 237–245, 2019.

[40] Z. Zhang, “Improved adam optimizer for deep neural networks,” in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–2.

[41] W. S. Noble, “What is a support vector machine?” *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[42] Z. Ramedani, M. Omid, A. Keyhani, S. Shamshirband, and B. Khoshnevisan, “Potential of radial basis function based support vector regression for global solar radiation prediction,” *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 1005–1011, 2014.

[43] X. Yu, K. Q. Pu, and N. Koudas, “Monitoring k-nearest neighbor queries over moving objects,” in *21st International Conference on Data Engineering (ICDE’05)*. IEEE, 2005, pp. 631–642.

[44] N. L. Crookston and A. O. Finley, “yaImpute: an R package for kNN imputation,” *Journal of Statistical Software. 23 (10). 16 p.*, 2008.

[45] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient knn classification with different numbers of nearest neighbors,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1774–1785, 2017.

[46] S. Ahmed, F. Khan, A. Ghaffar, F. Hussain, and S. H. Cho, “Finger-counting-based gesture recognition within cars using impulse radar with convolutional neural network,” *Sensors*, vol. 19, no. 6, 2019.

[47] F. Khan, S. K. Leem, and S. H. Cho, “Hand-based gesture recognition for vehicular applications using IR-UWB radar,” *Sensors*, vol. 17, no. 4, 2017.

[48] 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 *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 851–860.

[49] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, “Soli: Ubiquitous gesture sensing with millimeter wave radar,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[50] T. H. Vu, A. Misra, Q. Roy, K. C. T. Wei, and Y. Lee, “Smartwatch-based early gesture detection 8 trajectory tracking for interactive gesture-driven applications,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018.