# TSCNN: A 3D Convolutional Activity Recognition Network Based on RFID RSSI

Weiqing Huang[1,2,3], Yi Liu[2,3], Shaoyi Zhu[2,3], Siye Wang[1,2,3], Yanfang Zhang[2,3]
[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]Institute of Information Engineering Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{*huangweiqing, liuyi, zhushaoyi, wangsiye, zhangyanfang*}*@iie.ac.cn*

***Abstract*—Human activity recognition has a wide range of applications, especially for the care of elderly people living alone and the monitoring of abnormal behaviors of key personnel. Although conventional video surveillance technology has made many research advances in this field, this technology destroys people's privacy. Activity recognition technology based on RFID avoids damage to people's privacy, and is being widely studied and applied. This paper uses RFID Received Signal Strength Indicator (RSSI) to identify and classify human behaviors. Predecessors employed CNN and LSTM for human activity identification, but there were still some shortcomings: 1) The 2D convolution loses the temporal information of continuous actions and reduces the classification accuracy. 2) LSTM network has a series of training difficulties. 3) No available public dataset for the current mission.**

**To solve these problems, this paper proposes a convolutional neural network called temporal spatial convolutional neural network (TSCNN). Taking the continuous frame sequence as input, the network is designed using 3D convolution to realize real-time activities recognition. The average classification accuracy of our network is 94.6%, 15.6% higher than the state-of-the-art——Tagfree. Our lowest accuracy is 81.8%, and Tagfree is 35.4%. Besides, the ablation experiment proves the necessity of the design in the TSCNN network. Furthermore, we collect more than 60000 RFID signal data and transform them into corresponding pixel maps to form a new dataset. We present and expose the dataset called RF-men.**

***Index Terms*—human activity recognition, RFID, 3D convolution, ablation experiment, dataset**

## I. Introduction

Currently, activity recognition is an essential task and has wide applications. The elderly living alone need to monitor, classify and identify their unsafe behaviors at home through activity identification. In important meeting rooms of commercial companies, event activity recognition is also applied to human behavior control. The main task of this paper is to classify human behaviors based on Radio Frequency Identification (RFID) RSSI. RFID has the advantages of low cost, small size, maintenance-free, etc., and is widely used in many public places, such as identity cards, ETC, access control cards, bus cards, bank cards and the field of mobile applications, including human-object interaction detection [2], human-object tracking [3] and more complex activity identification [6]. Since the camera has problems with high line of sight (LOS) requirements, and unfriendly privacy protection, RFID technology can also be used as an alternative to cameras for detection tracking [11] and indoor positioning [5] and other scenarios. Therefore, we hope to come up with a method to use RFID to classify human behaviors and achieve the purpose of monitoring activity recognition.

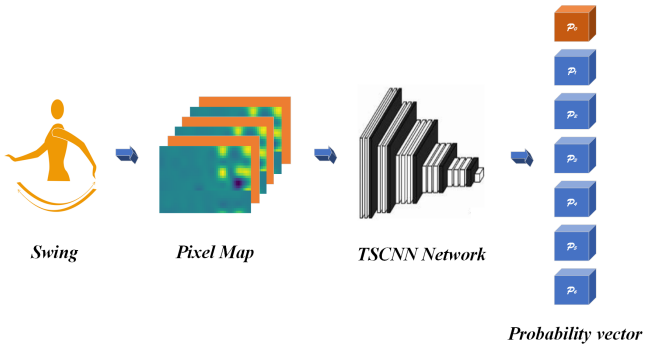


Fig. 1. Overall flow chart of experiment. The first picture from the left is the movement captured by the experimenter. The second picture from the left is a pixel map after pre-processing and visualizing the collected RFID RSSI. The third image from the left is a simple thumbnail of TSCNN. The last graph shows that the output of the network is a probability vector. The orange block represents the maximum value of the probability vector. Its subscript is the output classification of TSCNN.

Researchers in this field have done a lot of work and made breakthroughs to solve this problem, but there are still some shortcomings. Early activity recognition algorithms [33] were based on manual feature extraction methods. The ability to represent behaviors was limited by the extracted features, and it was difficult to reflect behaviour features effectively. With the development of deep learning, predecessors used two dimensional (2D) convolution to design classification networks [17] [18]. However, due to the limitations of 2D convolution, it is impossible to process time domain information, but the relative position relationship between human limbs and torso over a period of time defines the current person's movement. To accurately judge the behavior of others, neural network models are required to extract spatiotemporal features from the input data, so the recent study mixes 2D convolution and long-short-term-memory (LSTM) [20]. LSTM is memorable, so it is often used in the task of dealing with the nonlinear features of sequences, but LSTM itself has problems such as difficulty in training, large footprint, slow convergence, and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

easy to overfit. At the same time, the dataset used in the previous training network, such as weightlifting and falling, are not frequently used in daily life. So we came up with a dataset that is more consistent with everyday human behavior. Inspired by the idea of 'video', different from previous work, we transform an RSSI to a corresponding pixel map, treat it as a frame and stack adjacent frames together as input to the network. The graph method is more suitable for training with convolutional neural network (CNN) than RFID RSSI.

In our work, we propose an end-to-end trainable deep neural network called temporal spatial convolutional neural network (TSCNN). This network has the ability to categorize real-time data. Since we are dealing with 'video' data, we use three dimensional (3D) convolution instead of 2D convolution in previous work to extract feature map. 3D convolution can extract time domain information and space domain feature, so 3D convolutional neural network has better temporal information modeling ability than 2D convolutional neural network, and the classification of action is more accurate. We collected and exposed a dataset that is more consistent with everyday human behaviors.

To prove the superiority of TSCNN network, we compare with the state-of-the-art models. All the training and testing is based on our dataset. The average classification accuracy of TSCNN network is 94.6%, which is 22.1%, 16.9%, and 15.6% higher than the average classification accuracy of CSI-DFLAR, RF-finger, and Tagfree networks. We also performed ablation experiments to show that the hyperparameter we selected were the best.

Our contributions are as follows:

- We collected more than 60,000 pieces of data, produced a new dataset called RF-men, and made it public.
- Based on the idea of 'video', a new deep learning network is constructed by using 3D convolution to realize RFID-based activity recognition.
- Our network can distinguish the classification of current actions in real-time, and our experiment proves the validity of the model.
- Ablation experiments shows that the hyperparameter we selected were the best.

## II. RELATED WORK

As we all know, RFID is a promising and practical technology. With the increasing maturity of sensors, RFID has been widely used in mobile applications [3], human-object interaction [2] and more complex activity recognition [6].

### A. Experimental Environment Deployment

Compared with indoor positioning [4] [5] [7] [8], activity recognition [10] needs to capture fine-grained body movements. There are generally two ways to identify human behaviors through RFID: i) Device-based method employs RFID tags to attach to the human body. Reference [7] proposed a method to carry out on-site free-weight activity recognition and assessment of tags by using the doppler frequency shift of RFID signals when RFID tags are pasted on the dumbbell. The downside is that attaching labels to the body can sometimes be inconvenient and can be considered intrusive. ii) Device-free method apply multiple tags to fix in the environment as fixed references. Reference [2] fixed the RFID tag array to a wall and identified the behaviors by referring to the influence of human activities on the signal strength of the tag.

### B. Conventional Classification Treatment

In [1], the RSSI from RFID tag array is analyzed, and attitude classification is performed by using support vector machines and linear kernel functions. Han et al. proposed a method to use of the doppler shift to combine activity sensing, recognition and counting [7]. These classification methods have complicated processes and high computational cost. When new actions are added, the whole layered activity recognition framework needs to be redesigned with poor scalability. Shang guan *et al*. proposed spatiotemporal phase analysis (STPP) [12]. By analyzing the temporal and spatial dynamics of phase distribution, STPP can calculate the spatial order between tags and use template matching to classify human behaviors. Some researches also use hidden markov model, enhancement algorithm, bayesian network and other classification algorithms to infer daily activities from the trajectories used by objects [14], [15]. Reference [16] used RFID radio patterns to extract spatial and temporal features, which in turn were used to describe activities.

### C. Neural Network Classification Processing

With the great improvement of computational power, deep learning has become a very active research field of general activity understanding, and has achieved excellent results. Reference [17] employed CNN to recognize image-based multi-touch gestures. In [18], the wireless image processing method is proposed to extract wireless location and activity recognition without devices from wireless image features through deep learning.

Reference [19] proposed a neural network which combined CNN and GRU to process the velocity profiles of gestures and realize gesture recognition. In [20], a device-free activity identification system based on RFID is proposed. The classification network combining CNN and LSTM was adopted. However, LSTM network is not only very difficult to train, but also has a series of defects such as large space occupied, slow convergence speed and easy to overfit.

Reference [31] proposed the skip-connection mechanism, so that when the network going deeper, the gradient will not disappear. In [32], it enabled neural networks to extract semantic information and increase receptive field while minimizing loss of information. With these two mechanisms, 3D convolution has a lot of applications, such as human action recognition in videos [21]–[23], action detection [24], video caption [25], hand gesture detection [26], video learning [27], video super-resolution [9], [13]. This paper extends the application of 3D convolution and designs a 3D convolutional neural network based on RFID RSSI to classify human behavior.

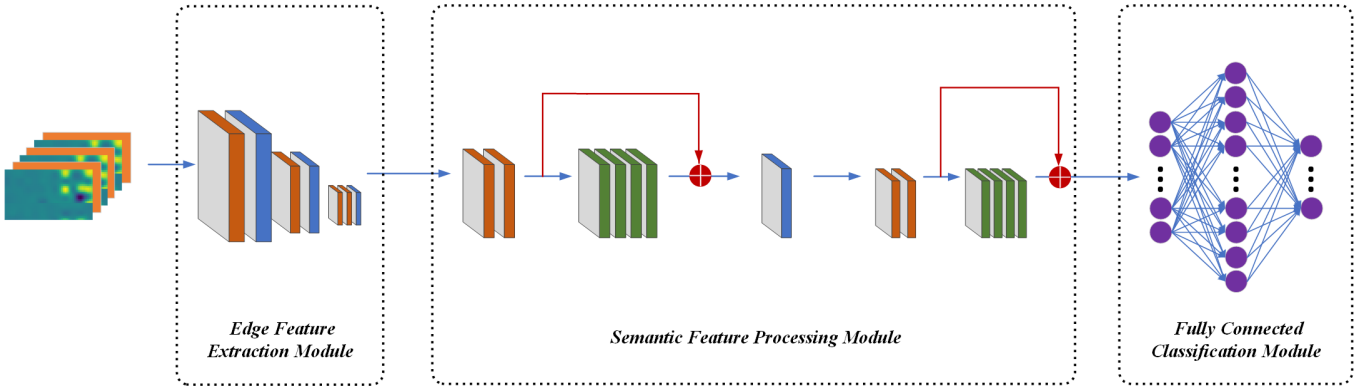


Fig. 2. Overview of our TSCNN model. TSCNN model has 3 modules, edge feature extraction module, semantic feature processing module and fully connected classification module. The orange cuboid represents convolutional layer with BN and ReLU. The blue cuboid represents maxpooling Layer. The green cuboid represents convolutional layer in resnet. $\bigoplus$ represents element-wise add operation. Fully connected classification module consists of two fully connected layers.

## III. Method

In this section, we will introduce our model called temporal spatial convolutional neural network (TSCNN), which is an end-to-end deep neural network based on 3D convolution, to deal with the problem of behavior classification based on RFID RSSI. We defined $f$ as the TSCNN network, ($x_0$, $x_1$,..., $x_{15}$) as the network input, and $C$ as the real classification of the behavior sequence $\chi$. We formulated the network objectives as follows

$$\arg\max(f(x_0, x_1, ..., x_{15})) = C \tag{1}$$

### A. 3D Convolution

The 3D convolutional network is suitable for the study of spatiotemporal features. 2D convolution is used in the fusion model [30], and most networks will lose the input time signal after the first convolutional layer. In [29], although the time-flow network uses multiple frames as input, the time information completely collapses after the first convolutional layer because of the 2D convolution. In contrast, 3D convolution network can extract spatiotemporal features across frames, and it leads to a better ability to acquire time domain information than 2D convolution network. The time relationship between our RFID RSSI is the reason why we employ 3D convolution to our network.

### B. Model Framework

Our model is divided into three parts: edge feature extraction module, semantic feature processing module and fully connected classification module.

The function of the edge feature extraction module is to extract the shallow features of successive frames and enlarge the receptive field. This module consists of four convolution blocks and three max-pooling layers. Thereinto, convolution block is composed of a convolution layer, a batch normalization layer and a ReLU layer. The function of convolution block is to extract the spatiotemporal feature and improve the nonlinearity of the network. The aim of batch normalization is to assist network training and make the training process more stable. As for the ReLU layer, its purpose is to add the nonlinearity of the network. Max-pooling layer can enlarge the receptive field and reduce training costs. The equation of the module as follows.

$$Fea_{st} = EFEM(x_0, x_1, ..., x_{15}) \tag{2}$$

where ($x_0$, $x_1$,..., $x_{15}$) are 16 frames that are stacked together and are the input of the network. $EFEM$ represents edge feature extraction module and $Fea_{st}$ represents the spatiotemporal feature.

After edge feature extraction module, we gain enough spatiotemporal features. Through semantic feature processing module, semantic information of multiple frames can be extracted from the spatiotemporal feature. The structure of semantic feature processing module is kind of like edge feature extraction module. The difference between them is that semantic feature processing module references the skip-connection mechanism. The skip-connection mechanism gives semantic feature processing module ability to add depth of the network and avoid gradient disappearance problem at the same time.

$$F = SFPM(Fea_{st}) \tag{3}$$

where $SFPM$ represents semantic feature processing module and $F$ represent the output of semantic feature processing module.

Fully connected classification modules integrates feature representations together and outputs a value, which greatly reduces the impact of feature positions on classification. Through fully connected layer, the probability of all classifications is generated.

$$P = FC(F) \tag{4}$$

$$C = \arg\max_i P_i \tag{5}$$

where $FC$ is on behalf of Fully connected classification module and $P$ is the output of our network. $C$ is a visual representation of $P$.

### C. Loss Function

We use the Binary-Cross-Entropy (BCE) function as a loss function to limit our model training. The objective variable of the classification problem is discrete, while the objective variable of the regression problem is continuous. The cross entropy describes the distance between two probability distributions. The smaller the cross entropy is, the closer the two are to each other.

$$L = BCE(P, label) \tag{6}$$

where 'label' is the label information corresponding to the input sample, that is the true category. L is the loss function.

TSCNN network training process is shown in algorithm 1. Its objective function is shown in equation 7.

$$f(\theta) = L(TSCNN(x_0, x_1, ..., x_{15}), label) \tag{7}$$

**Algorithm 1** Training Process of TSCNN.

**Require:**

$f(\theta)$: TSCNN model objective function with parameters $\theta$;
learning rate: $\alpha$;
Exponential decay rate for the moment estimates: $\beta_1$, $\beta_2$;
$\theta_0$: Initial parameter which is initialized by kaiming_normal;
$m_0 \leftarrow 0$: Initialize $1^{st}$ moment vector;
$v_0 \leftarrow 0$: Initialize $2^{nd}$ moment vector;
$t \leftarrow 0$: Initialize timestep;
$max\_t$: Initialize max timestep;
**procedure** UPDATE TSCNN
  **while** $t \leq max\_t$ **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta \ f_t(\theta_{t-1})$
    $m_t \leftarrow \beta_1 \cdot \mathrm{m}_{t-1} + (1 - \beta_1) \cdot \mathrm{g}_t$
    $v_t \leftarrow \beta_2 \cdot \mathrm{v}_{t-1} + (1 - \beta_2) \cdot \mathrm{g}_t^2$
    $\widehat{m_t} \leftarrow \mathrm{m}_t / (1 - \beta_1)$
    $\widehat{v_t} \leftarrow \mathrm{v}_t / (1 - \beta_2)$
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m_t} / (\sqrt{\widehat{v_t}} + \epsilon)$
  **end while**
  Return $\theta_t$ (Resulting parameters)
**end procedure**

## IV. EXPERIMENT

### A. Dataset

In our investigation, since the current task does not have a public available dataset for experimentation, we produced our own training dataset and testing dataset and published them called RF-men [1].

The experimental environment for collecting the RFID RSSI is shown in Figure 3. To ensure the universality and authenticity of the collected data, the sexes of our volunteers are male and female and volunteers' height ranges are 163cm∼185cm and weight ranges are 50kg∼85kg. The volunteer stands between the RFID tag wall and the four antennas array in any position. There are 8 rows and 12 columns of tags on the wall. The antenna array communicates with the RFID tag wall. When the volunteer actions between the four antennas array and the RFID tag wall, the RSSI will be sharply decreased due to the occlusion effect of the human body. Four antennas collect the RSSI values of all tags on the RFID tag wall every turn to form a matrix, that is, the original data. In this work, we formulate our problem as follows.

Let $\mathcal{O} \subset \mathbb{R}^{r \times c}$ ( $r \times c$ is the number of tags) be the domain of observable RSSI $\mathbf{o}$. Suppose we have n rounds RSSI $\{\mathbf{o}_i \in \mathcal{O}, i = 1,..., n\}$. When $n$ = 1, the processed data is real-time data. The raw data can be represented as:

$$\mathbf{O} = [\mathbf{o}_1, ..., \mathbf{o}_n] \in \mathbb{R}^{r \times c \times n} \tag{8}$$

The original data is transformed into the corresponding pixel map $x$ by radio tomographic imaging (RTI) [28]. We obtain the weight matrix $\mathbf{W}$ by setting the width of weighting ellipse to 0.007 and set $\alpha$ to 4.5.

$$\Pi = (\mathbf{W}^T \mathbf{W} + \alpha \mathbf{I})^{-1} \mathbf{W}^T \tag{9}$$

$$x = \Pi \mathbf{O} \tag{10}$$

where $\mathbf{I}$ is the identity matrix.

The details of our dataset are as follows. Our data sets are all half-body data, not whole-body data. In real life, there may be obstacles to block. We can not guarantee that we collect the whole body data. The difference between RFID and image recognition is that it can only use the data pre-processing method without the need for an additional neural network to separate the upper and lower limbs, so we choose half-body data instead of whole-body data.

*1) Training data:* There are seven classes in our training data: swing, wave, still, sit, bow, stand and walk. Among them, swing, wave, still and bow are upper limb movements, while sit, stand and walk are lower limb movements. Our various data quantities are shown in Table I. The action classification diagram is shown in Figure 4.

TABLE I
THE AMOUNT OF TRAINING DATA

| Classes | swing | wave | still | sit | bow | stand | walk |
|---|---|---|---|---|---|---|---|
| Amount | 3688 | 3751 | 8312 | 8085 | 3930 | 7974 | 8265 |

*2) Testing data:* There are eight classes in our training data: sit-bow, sit-wave, sit-still, stand-bow, stand-wave, stand-still, walk-swing and walk-still. Each class corresponds to its respective upper and lower limb movements, and the specific data amount is shown in Table II.

The reason why Testing data and Training data are not completely unified is that the movements of the same upper limb can correspond to the movements of different lower limbs. Due to the difference in overall movements, the data collected for the same upper limb movements will also differ.

[1]Connection: https://pan.baidu.com/s/1mfRf7dAdUA57jaHleXHfFw Extraction code: jldk

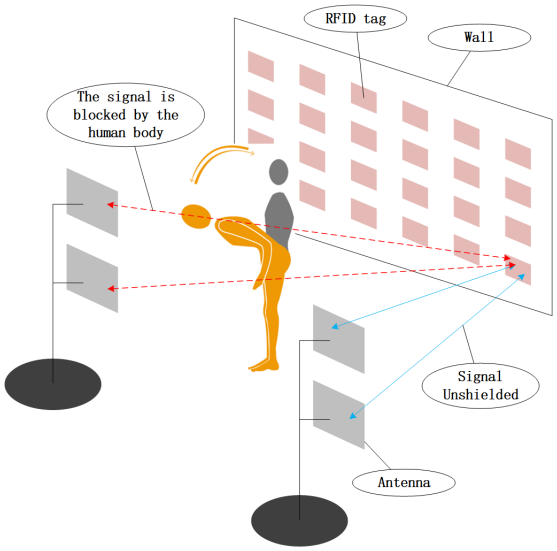


Fig. 3. RSSI data acquisition experiment scenario. The wall in the upper right corner of the picture is the ordinary solid white wall of the experimental site. The small powders on the wall are RFID tags. In practical applications, the RFID tag is a small square piece of transparent plastic, which is almost invisible to the naked eye. The four antennas in the lower-left corner of the figure are fixed directly in front of the RFID tag wall through two metal racks. They are 1.8m from the wall and are used to collect RFID signals. The experimenter acted at any point between the wall and the antenna.

In training data, we combine all the data that belong to the same category, but slightly different due to different upper or lower limbs, as one class. However, in practical applications, the whole-body behavior is classified, so testing data has practical significance.

### B. Training Details and Parameter

*1) Training Details:* We preprocessed the pixel maps $x$ as follows to form the input suitable for the TSCNN network. First, we interpolate the 18 $\times$ 47 pixel map to 48 $\times$ 48 by bicubic interpolation, making the new pixel map $x$. Then, we use a similar sliding window mechanism to generate real-time data, with the window sliding back one frame at a time and the window size is 16. Stack the pixel map $x$ in the sliding window together to form a continuous pixel map flow ($x_0$, $x_1$,..., $x_{15}$), which we view as a 'video clip.' In this way, our network can handle real-time data. Finally, we take this continuous pixel map flow ($x_0$, $x_1$,..., $x_{15}$) as the input to our TSCNN network.

*2) Parameter:* During the training, we set the batch size to 16 and trained a total of 200 epochs. For optimization, we use Adam with $\beta$ = 0.9, learning rate = 0.0001. The detailed network hyperparameters are shown in Table III.

### C. Baseline

We used our dataset to train CSI-DFLAR [18], RF-finger [17], Tagfree [20] networks. Their batches are 16, 32, 32. They are all trained in 200 epochs, and learning rate is 0.0001. Their performance are shown in Table IV. The first column corresponds to the whole body classification, the second column corresponds to the upper and lower limb classification, and the remaining four columns correspond to the classification accuracy of CSI-DFLAR, RF-finger, Tagfree, and TSCNN (ours), respectively.

### D. Performance of Network

It can be intuitively seen from the confusion matrix in Figure 6 that the TSCNN network has achieved a good behavior classification effect. The abscissa is the predicted classification result of TSCNN network. The ordinate is the true classification. The confusion matrix is diagonally distributed.

We compared it with several state-of-the-art networks. Result shows in Table IV and Figure 5. We can easily find out our network has the highest average accuracy rate. In detail, compared with the CSI-DFLAR network, the average accuracy rate of our TSCNN network is 22.1% higher than it. Especially in the three types of whole-body classification of stand-bow, stand-still, and stand-wave, whether it is upper limb classification or lower limb classification, the accuracy rate of our TSCNN network has been greatly improved, and even for the classification of stand-still——stand (lower limb), our accuracy rate is improved by up to 70.7%.

Compared with the CSI-DFLAR network, the average accuracy rate of RF-finger network is 5.2% higher than it. It seems that the RF-finger network has made some progress, but still 16.9% lower than our TSCNN network. For the three categories of stand-bow, stand-still, and stand-wave, the CSI-DFLAR network performs poorly, and the RF-finger network does not perform better. In these three categories, the best accuracy rate obtained by the RF-finger network is 74.8%, while the worst accuracy rate of our TSCNN network is 82.4%.

Due to the introduction of LSTM network, the experimental effect of Tagfree network is slightly improved, with an average classification accuracy of 79.0%. However, It is still 15.6% lower than our TSCNN network. According to the experiment, the lowest classification accuracy of Tagfree is 35.4%. Therefore, Tagfree network does not well solve this task.

### E. Analyses

There is essentially no difference between the RF-finger network and the CSI-DFLAR network. The difference between those two networks is that the RF-finger network is deeper than the CSI-DFLAR network. In most classifications, RF-finger is slightly better than CSI-DFLAR networks, but in sit-bow and walk-swing classifications, it is lower than CSI-DFLAR networks. This phenomenon directly proves that the deeper the number of layers, the effect of the network on all classifications may not be better.

Through Table IV, we can easily find that when the action to be detected is a dynamic action, the network (Tagfree, TSCNN) recognition effect with multiple frames as the model input is better than the network (CSI-DFLAR, COTS) with single frame as the model input. Compared with single frame activity recognition, frame series have additional temporal information. The Tagfree network introduces LSTM in order to

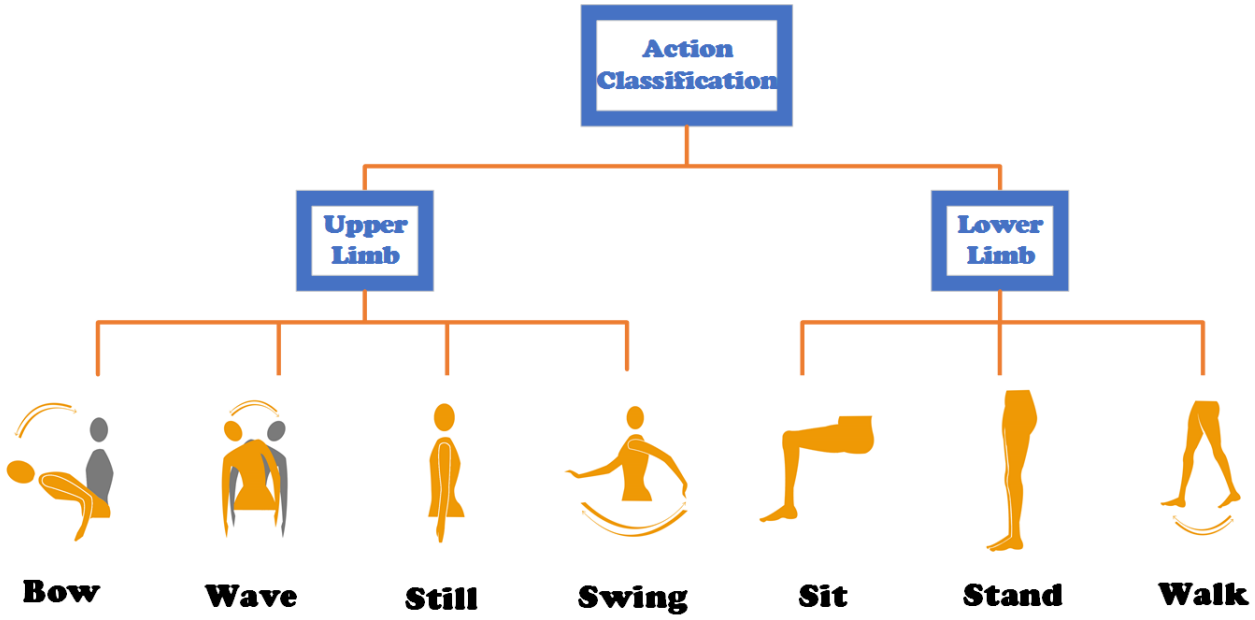


Fig. 4. Action classification diagram. The dataset is classified into two main categories, upper limb behaviors and lower limb behaviors. There are four types of upper limb behaviors: bow, wave, still, and swing, and three types of lower limb behaviors: sit, stand, and walk.

TABLE II
THE AMOUNT OF TESTING DATA

| Classes | sit-bow | | sit-wave | | sit-still | | stand-bow | | stand-still | | stand-wave | | walk-swing | | walk-still | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fine | bow | sit | wave | sit | still | sit | bow | stand | still | stand | wave | stand | swing | walk | still | walk |
| Amounts | 832 | 832 | 789 | 789 | 1017 | 1017 | 845 | 845 | 969 | 969 | 813 | 813 | 1574 | 1574 | 1563 | 1563 |

TABLE III
THE DETAILED NETWORK HYPERPARAMETERS

| Layer | Fliter Num |
|---|---|
| Convolutional Layer with BN, ReLU | $32 \times 3 \times 3$ |
| Maxpooling Layer | $2 \times 2 \times 2$ |
| Convolutional Layer with BN, ReLU | $64 \times 3 \times 3$ |
| Maxpooling Layer | $2 \times 2 \times 2$ |
| Convolutional Layer with BN, ReLU | $128 \times 3 \times 3$ |
| Convolutional Layer with BN, ReLU | $128 \times 3 \times 3$ |
| Maxpooling Layer | $2 \times 2 \times 2$ |
| Convolutional Layer with BN, ReLU | $256 \times 3 \times 3$ |
| Convolutional Layer with BN, ReLU | $256 \times 3 \times 3$ |
| Resnet (Convolutional Layer $\times$ 4) | $256 \times 3 \times 3$ |
| Maxpooling Layer | $2 \times 2 \times 2$ |
| Convolutional Layer with BN, ReLU | $256 \times 3 \times 3$ |
| Convolutional Layer with BN, ReLU | $256 \times 3 \times 3$ |
| Resnet (Convolutional Layer $\times$ 4) | $256 \times 3 \times 3$ |
| Maxpooling Layer | $2 \times 2 \times 2$ |
| Fully Connected Layer with ReLU | 2048 |
| Fully Connected Layer with Softmax | 7 |

TABLE IV
EFFECT COMPARISON WITH STATE-OF-ART

| Classes | Detail | CSI-DFLAR | RF-finger | Tagfree | TSCNN |
|---|---|---|---|---|---|
| sit-bow | bow | 92.8% | 84.0% | **93.7%** | 88.4% |
| | sit | 93.4% | 99.8% | 96.2% | **100.0%** |
| sit-wave | wave | 84.9% | 88.0% | 92.4% | **100.0%** |
| | sit | 99.7% | 99.6% | 99.9% | **100.0%** |
| sit-still | still | 86.7% | **99.3%** | 84.2% | 95.3% |
| | sit | **100.0%** | 99.5% | 99.0% | **100.0%** |
| stand-bow | bow | 40.4% | 49.6% | 43.3% | **82.4%** |
| | stand | 45.7% | 52.0% | 35.4% | **92.5%** |
| stand-still | still | 53.7% | 74.8% | 82.5% | **93.5%** |
| | stand | 26.2% | 35.0% | 53.0% | **96.9%** |
| stand-wave | wave | 60.6% | 74.7% | 86.7% | **100.0%** |
| | stand | 49.2% | 59.8% | 46.2% | **100.0%** |
| walk-swing | swing | 66.0% | 57.9% | 82.0% | **82.7%** |
| | walk | 97.1% | 97.6% | 99.6% | **100.0%** |
| walk-still | still | 65.3% | 73.8% | 71.3% | **81.8%** |
| | walk | 98.5% | 98.3% | 99.1% | **100.0%** |

record the temporal information. Compared with CSI-DFLAR and COST networks, the accuracy of classifying dynamic actions has been significantly improved, but it is still not satisfactory. This is due to the shortcomings of LSTM itself. LSTM's training process is slow and unstable. The loss of the LSTM is hard to constringe. In contrast, 3D convolution does not have such a problem. The experimental results prove that TSCNN has significantly improved the classification effect of dynamic classes than Tagfree.

### *F. Ablation Experiment*

Due to the uncertainty of the actions, the completion time of each action, the start and end time of each action cannot be

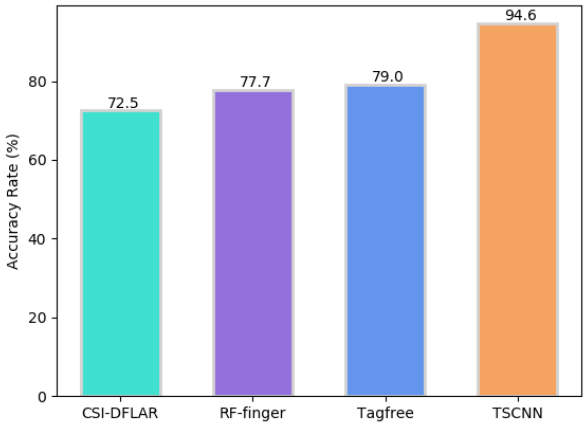


Fig. 5. The average classification accuracy of CSI-DFLAR, RF-finger, Tagfree, TSCNN.

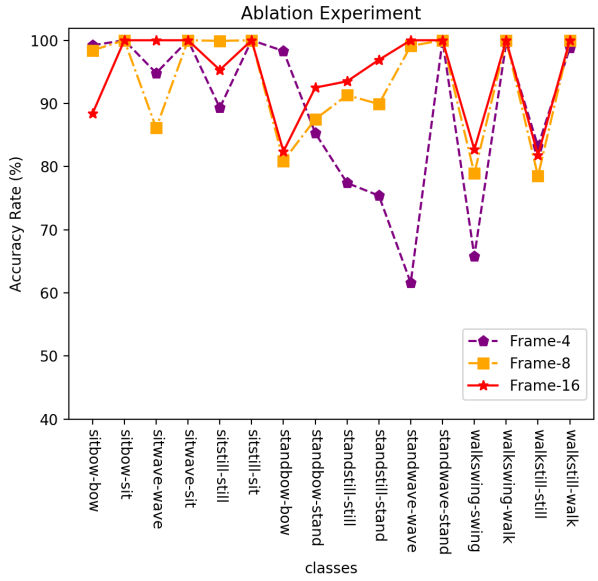


Fig. 7. Line chart of ablation experiment results. Among them, the x-axis is the behavior classification, and the y-axis is the correct rate. Line chart of ablation experiment results. The purple line is the performance of the TSCNN network on the test data when the input is 4 frames. The orange line is the performance of the TSCNN network on the test data when the input is 8 frames. The red line is the performance of the TSCNN network on the test data when the input is 16 frames. It can be seen that when the input is 16 frames, the TSCNN network performs best and most stable.

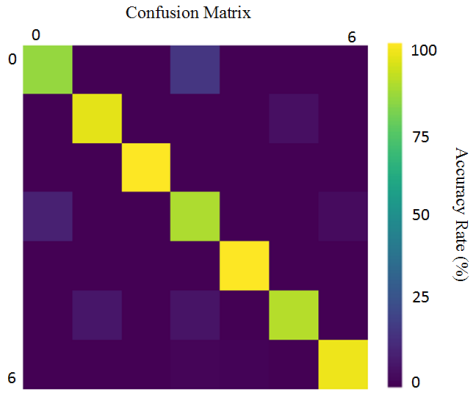


Fig. 6. Confusion matrix visualization. The abscissa is the predicted classification result of TSCNN network. The ordinate is the true classification. Legend color from dark to light indicates accuracy from low to high.

TABLE V
TEST CALCULATION PROCESSING TIME COMPARISON

| Classes | CSI-DFLAR | RF-finger | Tagfree | TSCNN |
|---|---|---|---|---|
| Times(s) | 0.0109 | 0.0221 | 0.0425 | 0.0268 |

determined. Unlike computer vision task, RFID gathers one frame to require about 1 second. Although the smaller the number of frames input by the model each time, the faster the model starts, but the model detection effect is unstable. In the case of high accuracy, we need to ensure that the correct recognition rate of all actions can be guaranteed. As can be seen from Figure 7, with 16 frames as model input, the minimum recognition rate of motion is 81.8%. Therefore, we choose 16 frames as the model input.

Table V compares the computational processing time for each frame tested. CSI-DFLAR takes the shortest time due to the simplest network. Tagfree uses the LSTM network for the longest time.

## V. FUTURE WORK

Our model has achieved excellent results on the current task, but there are still some shortcomings. The minimum accuracy rate of TSCNN network recognition is 81.8%, which already meet the basic requirements. However, in the case of monitoring the behavior of key personnel and monitoring important meeting rooms of commercial companies, such accuracy is insufficient. We hope to propose a new network and increase the accuracy rate of all classification to more than 95%.

On the basis of behavior classification, RFID-based behavior detection and abnormal behavior detection are further implemented. We will supervise whether there is any interesting action in the monitoring area, locate the time of action, trim the action type according to the time of action, and define the abnormal, so as to achieve the detection of abnormal behavior in a specific place, and report to the police in time.

Due to the difficulty of RFID data collection and preprocessing, for example, it takes too long to collect data, specific experimental sites need to be set up, and the method of data preprocessing directly affects the success of the experiment. Therefore, it takes a lot of time and energy to make the dataset required for the experiment. In the future, we hope to expand the dataset and improve the robustness and accuracy of the model by generating adversarial samples.

## VI. CONCLUSION

In this paper, we design the TSCNN network which successfully extracted the spatiotemporal features of RFID RSSI, and identify and classify 8 common human behaviors. TSCNN is suitable for situations where LOS requirements cannot reach the installation of video surveillance equipment or privacy protection is required. The TSCNN network uses 3D convolution to extract the temporal and spatial features in consecutive frames. At the same time, we also collected and disclosed the dataset called RF-men used in our experiments. Through experiments, we have proved the effectiveness of the TSCNN network. Its average classification accuracy rate is 94.6%, which is far more better than the state-of-art solutions. We also demonstrate through ablation experiments that the hyperparameters we selected are satisfactory and can meet the daily needs of the high privacy protection place.

## VII. ACKNOWLEDGEMENT

This research was supported by the National Key Research and Development Project "High precision, low delay electromagnetic spectrum monitoring and comprehensive situation analysis system" (grant No. 2018YFF0301202).

## REFERENCES

[1] Yao, Lina , et al. "RF-Care: Device-Free Posture Recognition for Elderly People Using A Passive RFID Tag Array." Eai Endorsed Transactions 2015. in press.

[2] Wang, Jue , D. Vasisht , and D. Katabi . "RF-IDraw: virtual touch screen in the air using RF signals." Acm Conference on Sigcomm ACM, 2014. in press.

[3] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. "Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices." ACM, 2014. In press.

[4] Dong Cui, Qiang Zhang. "The RFID data clustering algorithm for improving indoor network positioning based on LANDMARC technology." Cluster Computing 5(2017):1-8.

[5] He Xu , Manxing Wu, Peng Li, Feng Zhu, Ruchuan Wang. "An RFID indoor positioning algorithm based on support vector regression." Sensors 18.5(2018):1504-.

[6] Ding, Han , et al. "FEMO: A Platform for Free-weight Exercise Monitoring with RFIDs." Acm Conference on Embedded Networked Sensor Systems 0. in press.

[7] Ding, Han , et al. "A platform for free-weight exercise monitoring with passive tags." IEEE Transactions on Mobile Computing (2017):1-1.

[8] He Xu, Ye Ding, Peng Li, Ruchuan Wang, Yizhu Li. "An RFID Indoor Positioning Algorithm Based on Bayesian Probability and K-Nearest Neighbor." Sensors 17.8(2017):1806-.

[9] Yan Huang, Wei Wang, and Liang Wang. "Video Super-Resolution via Bidirectional Recurrent Convolutional Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence PP.99(2017):1-1.

[10] Takanobu Nakahara, and Katsutoshi Yada. "Analyzing consumers' shopping behaviour using RFID data and pattern mining." Advances in Data Analysis & Classification 6.4(2012):355-365.

[11] W. Huang, S. Zhu, S. Wang, J. Xie and F Zhang, "Sparse Representation for Device-Free Human Detection and Localization with COTS RFID," 2019 International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Melbourne, 2019.

[12] Longfei Shangguan, Zheng Yang, Alex X. Liu, Zimu zhou, Yunhao Liu. "Relative localization of RFID tags using spatial-temporal phase profiling." Usenix Conference on Networked Systems Design & Implementation USENIX Association, 2015. in press.

[13] Li, Sheng , et al. "Fast Spatio-Temporal Residual Network for Video Super-Resolution." CVPR, 2019. in press.

[14] Maja Stikic, Tam Huynh, Kristof Van Laerhoven, Bernt Schiele. "ADL recognition based on the combination of RFID and accelerometer sensing." Pervasive Computing Technologies for Healthcare, 2008. in press.

[15] Michael Buettner, Richa Prasad, Richa Prasad, Matthai Philipose, David Wetherall. "Recognizing daily activities with RFID-based sensors." UbiComp 2009: Ubiquitous Computing, 11th International Conference, UbiComp 2009, Orlando, Florida, USA, September 30 - October 3, 2009, Proceedings ACM, 2009. in press.

[16] L. Wang, T. Gu, H. Xie, X. Tao, J. Lu, and Y. Huang. "A Wearable RFID System for Real-Time Activity Recognition Using Radio Patterns." International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services Springer, Cham, 2013. in press.

[17] Chuyu Wang, et al. "Multi-Touch in the air: Device-free finger tracking and gesture recognition via COTS RFID." IEEE INFOCOM, 2018. in press.

[18] Qinghua Gao, Jie Wang, Xiaorui Ma, Feng Xueyan, Hongyu Wang. "CSI-based device-free wireless localization and activity recognition using radio image features." IEEE Transactions on Vehicular Technology (2017):1-1.

[19] Yue Zheng, et al. "Zero-effort cross-domain gesture recognition with Wi-Fi." Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, 2019. in press.

[20] Xiaoyi Fan, Wei Gong, Jiangchuan Liu. "TagFree activity identification with RFIDs" Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018.

[21] Du Tran, et al. "A closer look at spatiotemporal convolutions for action recognition." CVPR, 2018. in press

[22] Moez Baccouche, Franck Mamalet, Franck Mamalet, Christian Wolf, Atilla Baskurt, Christophe Garcia, Atilla Baskurt. "Sequential Deep Learning for Human Action Recognition." Human Behavior Unterstanding - Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011.

[23] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu. "3D Convolutional Neural Networks for Human Action Recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, Jan 2013.

[24] Shou Zheng, Dongang Wang, Shih-Fu Chang. "Temporal action localization in untrimmed videos via multi-stage cnns." CVPR, 2016. in press

[25] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, Yong Rui. "Jointly modeling embedding and translation to bridge video and language." CVPR, 2016. in press.

[26] Molchanov, Pavlo , et al. "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks." CVPR, 2016. in press

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." 2015 IEEE International Conference on Computer Vision (ICCV) IEEE, 2015. in press.

[28] Joey Wilson, Neal Patwari. "Radio tomographic imaging with wireless networks." IEEE Transactions on Mobile Computing, 2010, 9(5):621-632.

[29] Simonyan Karen, Zisserman Andrew. "Two-stream convolutional networks for action recognition in videos." NIPS, 2014. in press.

[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale video classification with convolutional neural networks." CVPR, 2014. in press.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Identity mappings in deep residual networks." ECCV, 2016. in press.

[32] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. "Densely connected convolutional networks." CVPR, 2017. in press.

[33] Laptev Ivan. "On space-time interest points." International Journal of Computer Vision 64.2-3(2005):107-123.