A Dynamic Hand Gesture Recognition Algorithm Based on CSI and YOLOv3

Qiang Zhang*, Yong Zhang and Zhiguo Liu
Hefei University of Technology, Hefei 230001, China

*1378082472@qq.com

Abstract. Object detection algorithms based on convolutional neural networks are generally suitable for static gesture recognition. For actual hand gesture scenes, dynamic gestures are also widely used. A dynamic hand gesture recognition algorithm based on Channel State Information (CSI) and You Only Look Once: Version 3 (YOLOv3) is proposed for continuous dynamic hand gesture recognition. The data acquisition adopts a CSI-based radio frequency method. The adaptive weighted fusion, Kalman filtering, threshold segmentation and data conversion are used to generate gray value images. Finally, the YOLOv3 object detection algorithm is used to train and identify the grayscale image which include the information of continuous dynamic hand gestures. The effectiveness of the proposed method is verified by the recognition confusion matrix. And the proposed method has an average recognition accuracy of 94% for four custom dynamic hand gestures.

1. Introduction

With the continuous development and advancement of computer technology, the way of human-computer interaction (HCI) has also been diversified, and dynamic hand gestures have become an important way of HCI. In recent years, with the continuous development and advancement of WiFi technology, the way of dynamic hand gesture data acquisition has also undergone fundamental changes, and non-contact HCI is no longer limited to depth sensors, infrared cameras, etc. [1-2]. The channel state information (CSI) can represent the amplitude and phase information of each subcarrier corresponding to the space in the frequency domain, and fully utilizes the idea of Orthogonal Frequency Division Multiplex (OFDM) to make the wireless signal enables efficient communication between the transmitter and the receiver. Even if the object only performs a relatively small movement, CSI can be sensitively perceived and displayed on each subcarrier[3]. Therefore, some researchers use CSI to do human behavior recognition [4] and indoor intrusion detection [5].

How to classify the collected dynamic hand gesture data is also a research hotspot of dynamic hand gesture recognition technology. At present, the object detection algorithm based on CNN can effectively identify static hand gestures. The YOLOv3, a kind of object detection algorithm proposed by Redmon et al. [6] is one of the best object detection algorithm at present. Currently, YOLOv3 has been successfully applied to pedestrians detection [7], vehicle detection [8] and other fields[9-10]. However, the input of the YOLOv3 algorithm receives a static image, and the original CSI data cannot be used as an input.

In order to solve this problem, this paper adopts adaptive weighted fusion to extract a set of data from multiple links, adopts Kalman filter to filter the noise, and adopts threshold segmentation to extract the CSI valid data of the hand gesture action area. The process of pre-processes the raw CSI
data and data conversion to convert the CSI data into gray value images, and finally adopts YOLOv3 for training and recognition.

2. Proposed method

The algorithm of this paper mainly includes CSI data acquisition, CSI data preprocessing, data conversion, YOLOv3 training and recognition. The algorithm flowchart is shown in Figure 1.

![Flowchart of the proposed method.](image)

2.1. CSI data acquisition

The Intel 5300 NIC is used as the receiving AP, and the TL-WR845N dual-antenna router is used as the transmitting DP. Multi-input Multi-output (MIMO) model is adopted to save bandwidth when actually collecting CSI data.

In order to collect more hand gesture CSI data, the AP uses three antennas to transmit signals, and the DP uses two antennas to receive signals to form a MIMO array. The data collectors made four custom hand gestures (waving to the left, waving to the right, waving forward, and waving backwards) between the DP and the AP. The experimental scene was located at the B1908 Laboratory of Science and Technology Building, Hefei University of Technology. The data collection scenario is shown in Figure 2.

![Schematic diagram of collecting dynamic hand gesture](image)

Each original CSI data packet is divided into 6 data streams, or 6 links, and each link has 30 subcarriers. Therefore, the obtained CSI data form can be expressed as:

\[
\text{CSI} = \begin{bmatrix}
H_{1,1} & H_{1,2} & \cdots & H_{1,30} \\
H_{2,1} & H_{2,2} & \cdots & H_{2,30} \\
\vdots & \vdots & \ddots & \vdots \\
H_{6,1} & H_{6,2} & \cdots & H_{6,30}
\end{bmatrix}
\] (1)
Where $H_{i,j}$ represents the CSI value of the signal state information of the $i$ th link and the $j$ th subcarrier.

2.2. CSI data preprocessing

The original CSI data obtained in this way is co-dimensional data, and the too high dimension means that the complexity is too large, and the data needs to be dimensionally reduced. First, the signals of 30 subcarriers on each link are averaged to represent the data values of the link, and then adaptive weighted fusion [11] is applied to the six links to obtain a value. Kalman filtering is performed on the merged CSI data to remove the noise. Finally, the threshold segmentation is performed to obtain CSI data of the hand gesture action region.

2.2.1. Adaptive weighted fusion. Taking the hand gesture waving to left for example, the data of the 30 subcarrier signals on the six links after taking the tie value is shown in Figure 3. The abscissa represents time and the ordinate represents amplitude (in dB).

![CSI data waveform diagram of different links](image)

**Figure 3.** CSI data waveform diagram of different links

Next, adaptive weighted fusion of the CSI data obtained by the six links is performed to eliminate the possible redundancy between multiple data links, and reduce complexity and uncertainty.
Assuming the number of links is \( n \), the corresponding \( n \) groups of CSI data is \( X_1, X_2, \ldots, X_n \) respectively, adaptive weighted fusion is to use an adaptive method to obtain an optimal weighting factor matching the set of data, so that the total variance of the set of data is the smallest. \( \hat{X} \) represents the value after adaptive weighted fusion. The adaptive weighted fusion algorithm model is shown in Figure 4.

\[
\begin{align*}
\hat{X} &= \sum_{i=1}^{n} W_i X_i \\
\sum_{i=1}^{n} W_i &= 1
\end{align*}
\]

The total variance is:

\[
\sigma^2 = E[(X - \hat{X})^2] = E[\sum_{i=1}^{n} W_i X_i - \sum_{i=1}^{n} W_i X_i]^2] 
\]

After finishing:

\[
\sigma^2 = E[\sum_{i=1}^{n} W_i^2 (X - X_i)^2] = \sum_{i=1}^{n} W_i^2 \sigma_i^2
\]

Where \( \sigma_i^2 \) represents the variance of the \( i \)th group CSI data. The minimum value of the \( \sigma^2 \) is \( \sigma^2_{\text{min}} \):

\[
\begin{align*}
\sigma^2_{\text{min}} &= \min \left( \sum_{i=1}^{n} W_i^2 \sigma_i^2 \right) \\
\sum_{i=1}^{n} W_i &= 1
\end{align*}
\]

where \( \min(\bullet) \) represents the minimum operation. The optimal weighting factor is obtained when the variance is minimum:

\[
W_i^* = \frac{1}{\sqrt{\sigma_i^2 \sum_{j=1}^{n} \frac{1}{\sigma_j^2}}} \quad (i = 1, 2, \ldots, n)
\]

The corresponding total variance is the minimum at this time:

\[
\sigma^2_{\text{min}} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}
\]

The fused value of 6 links of CSI data obtained by adaptive weighted fusion is shown in Figure 5.
2.2.2. Kalman filter. Although the general low-pass filter can remove the noise signal, it also filters out some useful signals, and the filtering effect is not good when the gesture is slow. Kalman Filtering is a state equation of a linear system. The basic idea is to use the input and output of the system to observe the data, and then to estimate the state of the entire linear system [12]. The calculation principle is as follows:

\[ Z_t = F_t Z_{t-1} + B_t U_t + Q_t \]  

(8)

Where \( Z_t \) denotes the state vector at the time of \( t \), \( F_t \) denotes the state transition matrix, \( B_t \) denotes the control matrix, \( U_t \) denotes the control vector, \( Q_t \) denotes the noise matrix of the \( n_1 \) dimension. \( S_t \) is adopted to represents the observation vector of system at the time of \( t \) and the equation can be expressed as:

\[ S_t = P_t Z_t + V_t \]  

(9)

where \( P_t \) is the \( n_1 \times n_2 \) dimensional observation matrix, \( V_t \) is the \( n_2 \) dimensional observation noise matrix. Both \( V_t \) and \( Q_t \) are Gaussian white noise, which have have the following properties:

\[ E[Q_t] = 0, E[V_t] = 0 \]  

(10)

\[ E[Q_t Q_t^T] = C, E[V_t V_t^T] = D \]  

(11)

where \( C \) represents the prediction noise covariance matrix, and \( D \) represents the measurement noise covariance matrix.

Kalman filter is used to estimate the state quantity of the time system. The state prediction value \( Z_t \) at the system time of \( t \) can be expressed as:

\[ Z_t' = F_t Z_{t-1} + B_t U_t \]  

(12)

The covariance matrix prediction value \( G_t \) at the system time of \( t \) can be expressed as:

\[ G_t' = F_t G_{t-1} F_t^T + C_t \]  

(13)

Where \( G_{t-1} \) represents the covariance matrix value calculated at the system time of \( t-1 \). The Kalman gain matrix \( K_t \) can be expressed as:

\[ K_t = \frac{G_t P_{t}^T}{P_t G_t P_t^T + D_t} \]  

(14)

The optimal estimate of the state quantity of the system time of \( t \) is:

\[ \hat{Z}_t = Z_t + K_t (S_t - P_t Z_t) \]  

(15)
Update the covariance matrix:

$$G_t = (I - K_t P_t) G_t$$

(16)

where $I$ represents the unit matrix.

The result of performing the above Kalman filtering on the CSI data obtained by adaptive weighting is shown in Figure 6.

![Figure 6. Kalman filtered CSI data waveform](image)

2.2.3. Threshold segmentation. The CSI data obtained by the Kalman filter obtained in the previous section is a hand gesture of moving to left motion of 0 to 4 seconds. Although the noise is filtered, there are many meaningless data. Threshold segmentation separates the CSI data of the hand gesture action occurrence area from the whole data. In the process of CSI data analysis, it is found that the state change of the phase difference variance between CSI antennas is closely related to the hand gesture motion detection. Therefore, the threshold difference-based sliding window method is used to detect the phase difference of the CSI signal [13]. Specific steps are as follows:

1. Collect CSI data in a quiet environment without any action, and calculate the average value $\mu$ and normalized standard deviation $\sigma$ of the signals in the sliding window.

2. Determine the threshold $\delta$ based on the $\mu$ and $\sigma$ calculated in the previous step, and the relationship between the three is satisfied:

$$\mu + 6 \sigma \leq \delta$$

(17)

3. Acquire the signal flow in the sliding window, compare the average value of the signal in the sliding window with the threshold, and if the average value is greater, determine that the signal is a signal in the hand gesture action area. In the actual experiment process, since the dynamic hand gesture is a continuous process, the signal in the hand gesture action area should be continuous. There is a start point and an end point, and after determining the start point and the end point of the hand gesture action, the CSI data is segmented.

After threshold segmentation, the CSI data of the hand gesture action area obtained is as shown in Figure 7.

![Figure 7. Schematic diagram of CSI data after threshold segmentation](image)
2.3. Data conversion

After the threshold segmentation, the CSI data of the hand gesture action area is obtained. In order to facilitate the YOLOv3 training and recognition, the CSI data is converted into an image as an input of neural network for training and recognition. The CSI data is sampled to obtain 100 sets of data, and then the 100 sets of data are normalized and converted into gray value data, and finally mapped to the grayscale image.

2.3.1. Normalized operation. The CSI data of the hand gesture action area obtained in previous section is equally spaced to obtain 100 sets of CSI data \( A = \{a_1, a_2, \cdots, a_{100}\} \). Find the maximum and minimum values, and then normalize the 100 sets of data. The new data is \( L = \{l_1, l_2, \cdots, l_{100}\} \):

\[
l_i = \frac{a_i - \min}{\max - \min} \quad (i = 1, 2, \cdots, 100)
\]

Then the new data obtained is converted into gray value data \( O = \{o_1, o_2, \cdots, o_{100}\} \):

\[
o_i = l_i \times 255 \quad (i = 1, 2, \cdots, 100)
\]

The final result is 100 sets of gray value data between 0 and 255. The 100 sets of data contain all the information of the continuous dynamic hand gesture of waving to left in the sample.

2.3.2 Grayscale image mapping. The 100 sets of gray value data obtained in the previous section are arranged in a \(10 \times 10\) matrix:

\[
\begin{bmatrix}
o_1 & o_2 & \cdots & o_{10} \\
o_{11} & o_{12} & \cdots & o_{20} \\
\vdots & \vdots & \ddots & \vdots \\
o_{91} & o_{92} & \cdots & o_{100}
\end{bmatrix}
\]  

The \(10 \times 10\) matrix data is mapped into the image of the \(10 \times 10\) grids so that it corresponds to the gray value in the image grids. Figure 8 shows the grayscale value image obtained after the normalized operation of the CSI data of the hand gesture action area and the mapping of the grayscale image.

![Figure 8. CSI data mapping grayscale image](image)

2.4. YOLOv3 training and recognition

Through the above operations, the CSI data of the original continuous dynamic hand gesture action is converted into image data, and the images are assigned a training set and a test set in a ratio of 3:1. Next, the training set image is trained using the object detection algorithm YOLOv3 based on CNN, and the test set image is tested and identified.

The basic feature extractor of the YOLOv3 network uses the Darknet-53 model, the training set and the test set adopt the CSI hand gesture data to map the gray value image. The specific process is as follows:
(1) Copy each gray value image in the training set and adjust their resolutions to 300 × 225, 400 × 300, 500 × 375, 600 × 450, respectively. In this way, the number of images are expanded four times.

(2) Each of the gray value images obtained in the previous step is manually labeled in accordance with the format of the VOC data set, and the category information and the position information of the hand gestures are annotated to generate a hand gesture tag file.

(3) Combine all the gray value images for training obtained in the previous step and the corresponding hand gesture tag files, and train the YOLOv3 network as a training set (VOC format).

(4) Identify and detect the gray value image in the test set with the trained YOLOv3 model.

3. Experimental results and analysis

During the experiment, 1000 sets of CSI data were generated by waving to left, waving to right, waving forward, and waving backwards. After CSI data preprocessing and data conversion process, 4000 gray value images were obtained. For each kind of hand gesture motion, 750 images were extracted as a training set, and 250 images were used as a test set.

In this paper, the recognition confusion matrix is used to evaluate the experimental results. Each row of the matrix represents the probability that a real category is identified as each category, and each column represents the probability that each real category is identified as a particular category.

Assuming $P_{ij}$ represents the value of the cell of row $i$ and column $j$, then the value indicates the probability that the true category of the row $j$ is identified as the predicted category of the column $j$. The recognition confusion matrix of the four categories of CSI hand gesture motion gray value images is shown in Figure 9.

| Moving to left | Moving to right | Moving forward | Moving backward |
|---------------|----------------|---------------|-----------------|
| 0.95          | 0.04           | 0.01          | 0.00            |
| 0.05          | 0.93           | 0.01          | 0.01            |
| 0.02          | 0.01           | 0.94          | 0.03            |
| 0.01          | 0.00           | 0.06          | 0.93            |

Figure 9. Recognition Confusion matrix

It can be seen from the recognition results that the recognition accuracy rates of the four dynamic hand gestures of waving to left, waving to right, waving forward, and waving backward are 95%, 93%, 94%, 93%, respectively, and the average recognition accuracy is about 94%. Although there are only a few types of gestures, in the case of only 1000 sets of data for each type of action, recognition accuracy of 94% can be obtained.

4. Conclusion

This paper proposed a dynamic hand gesture recognition algorithm based on CSI and YOLOv3. The CSI data of the hand gesture action is collected by the network card and the router. The 30 subcarrier signals are averaged, and adaptive weighted fusion is performed, so that a set of data is combined from the CSI data of the six links. Then Kalman filter is used to perform the CSI data. By Filter denoising and threshold segmentation, the CSI data of the hand gesture action area is obtained. Next, the CSI data is converted into gray value image, and the YOLOV3 network is used for training and recognition. Recognition accuracy of 94% is obtained, which shows the feasibility and effectiveness of the proposed method.
References

[1] Wu X, Chu Z, Yang P, Xiang C, Zheng X, Huang W. Tw-see: Human activity recognition through the wall with commodity wi-fi devices [J]. IEEE Transactions on Vehicular Technology, 2019, 68 (1): 306–319.

[2] Man D, Wang W, Wang X, Lv J, Du X, Yu M. Pwig: A phase-based wireless gesture recognition system [C]. 2018 International Conference on Computing, Networking and Communications (ICNC), 2018: 837–842.

[3] Liu X, Cao J, Tang S, Wen J. Wi-sleep: Contactless sleep monitoring via wifi signals, 2014 IEEE Real-Time Systems Symposium, 2014: 346–355.

[4] Mei H F. Research and implementation of human behavior recognition based on CSI [D]. Wuhan University of Technology, 2017.

[5] Zhou J. Research on indoor intrusion detection and behavior recognition based on CSI in wireless sensing network [D]. Nanjing University of Posts and Telecommunications, 2018.

[6] Redmon J, Farhadi A. Yolov3: An incremental improvement [Z/OL]. Computer Vision and Pattern Recognition (CVPR), 2018, pp.126-134.

[7] Wang D W, He Y H, Li D, Liu Y, Xu Z J, Wang, J. Improved pedestrian detection algorithm for YOLOv3 infrared video images [J]. Journal of Xi'an University of Posts and Telecommunications, 2018, 23(04): 48-52+67.

[8] Zhang F K, Yang F, Li C. Fast vehicle detection method based on improved YOLOv3 [J]. Computer Engineering and Applications, 2019, 55(02): 12-20.

[9] Shuo Y, Zhang J X, Bo C J, Wang M, Chen L J. Fast vehicle logo detection in complex scenes [J]. Optics and Laser Technology, 2018.

[10] Liu B, Wang S Z, Zhao J S, Li M F. Ship tracking recognition based on Darknet network and YOLOv3 algorithm [J/OL]. Computer application, pp.1-7.

[11] Hua S Z. Research and application of multi-sensor data fusion algorithm [D]. Northeastern University, 2017.

[12] Liu J, Kong Y C. Fusion Positioning of WiFi Inertial Navigation Using Linear Kalman Filter [J]. Electronic Measurement Technology, 2017, 40(4):1-4.

[13] Liu J L. Research on human motion detection in complex scenes based on CSI [D]. Dalian University of Technology, 2017.