Journal of Computers Vol. 31 No. 5, 2020, pp. 1-13
doi:10.3966/199115992020103105001

Dynamic Gesture Recognition Method Based on Conditional Random Field and Weighted Voting Strategy

Yun Wu[1], Dong-Chen Huang[2], Wei-Chang Du[3], Meng-Ke Wu[1*], Xin Hu[1]

[1] Department of Computer Science, Northeast Electric Power University, Jilin 132012, China
{838558160, 124277466, 1099475750}@qq.com

[2] Shenzhen Controlled Corporation of NARI Technology Development Co., Ltd, Shenzhen 518055, China
479482134@qq.com

[3] Department of Information Engineering, I-Shou University, Kaohsiung 84001, Taiwan
wcdu@isu.edu.tw

Received 10 March 2019; Revised 16 July 2019; Accepted 28 August 2019

Abstract. Aiming at the diversity and spatio-temporal difference of dynamic gestures, the over-fitting and feature interference caused by the increase of feature dimensions, this paper proposes a dynamic gesture recognition method based on conditional random field and weighted voting method. Firstly, the two kinds of features of hand shape and trajectory are extracted. Then, the conditional random field model is used to model different feature subsets to obtain the difference base classifier. Finally, using the weighted voting method to fuse the base classifier to judge the gesture. The experimental results show that the method can accurately identify the hand shape and position with rich dynamic changes.

Keywords: conditional random field, gesture recognition, weighted voting strategy

1 Introduction

As a key component of human-computer interaction technology, gesture recognition has broad application prospects. It has been initially applied in the fields of virtual reality/augmented reality, game entertainment, robot control, and substation simulation training. Gestures can be divided into static gestures and dynamic gestures based on the motion characteristics of gestures. There, static gesture recognition is only related to the hand shape. Dynamic gestures are time series composed of static gestures, and both the shape and position change over time. Although dynamic gestures can express more information, it is more susceptible to spatio-temporal position and rate of movement than the static gestures, which makes the related research work more challenging.

In terms of feature extraction, most dynamic gesture recognition tasks extract 2D features. However, the dynamic gesture recognition algorithm based on 2D features is not sufficient to deal with the different orientation of gestures in motion. In recent years, with the emergence of sensors such as Leap Motion and Intel RealSense (RealSense), the aforementioned difficulties have been alleviated. Both of the sensors provide information such as the position and velocity of the hand joints. It is convenient to extract 3D features, making the relevant gesture recognition algorithm can adapt to the difference of gestures orientation. W. Lu et al. [1] made use of the hand information provided by Leap Motion to extract 3D hand features, which solved the problem of different gesture orientations. However, this algorithm lacks the extraction of trajectory features, which makes it impossible to fully describe dynamic gestures.

In terms of dynamic gesture recognition algorithms, Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [2] are classical algorithms. There are many types of HMM models, and they have their own unique advantages in describing time-dependent stochastic processes. But they also have some

* Corresponding Author

shortcomings. Due to strict independence assumptions, it is difficult for the HMM model to combine context information. Relatively speaking, without strict independence assumptions, the CRF model directly calculates the conditional probability of the output node under a given input node, and it has strong ability to utilize context information. However, with increasing the number of features, the interference problems and over-fitting problems between features will be highlighted, which could result in a decrease in classification performance.

In view of the above problems, using RealSense as the input device, this paper proposes a dynamic gesture recognition method based on conditional random fields and weighted voting. Firstly, extracting the dynamic hand gesture and trajectory feature. Then, modeling of different feature subsets by CRF Model, and a plurality of different base classifiers are obtained. Finally, the weighted voting integration strategy is used to fuse the classification results of each base classifier to realize dynamic gesture recognition. The experimental results show that the proposed system can accurately recognize dynamic gestures with different viewing angles. It has an overall recognition rate of 94.9%, indicating its eminent applicability in real-life applications.

2 Feature Extraction

RealSense provides some data about hand, including joint position, silhouette, speed, etc. The feature extraction work in this paper mainly utilizes the joint position data therein. Fig. 1. depicts the traceable hand joints and corresponding numbers.

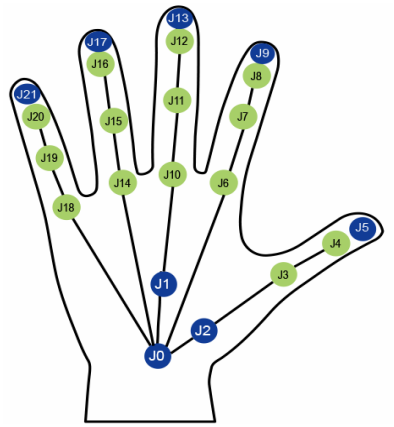


Fig. 1. Labels of the 22 hand joints

There are 22 hand joints positions provided by the SDK, and this paper defines the coordinates of these 22 joint points as $J_i, i = 0, 1, ..., 21$ (shown in Fig. 1.)

Among these 22 joints, J21, J17, J13, J9 and J5 are fingertips, J19, J15, J11, J7 and J4 are proximal interphalangeal joints, J18, J14, J10, J6 and J3 are metacarpophalangeal joints. J1 is the palm, and J0 is the wrist. In this paper, 24 structural vectors based on these 17 joints are constructed. The concrete definitions of these vectors are shown in Table 1.

Table 1. Definitions of hand structural vectors $\overrightarrow{J_{19}J_{21}}$

Related joints	Vectors	Definitions
Proximal interphalangeal joints, Fingertips	$\overrightarrow{J_{19}J_{21}}$	The vector pointing from J19 to J21
	$\overrightarrow{J_{15}J_{17}}$	The vector pointing from J15 to J17
	$\overrightarrow{J_{11}J_{13}}$	The vector pointing from J11 to J13
	$\overrightarrow{J_{7}J_{9}}$	The vector pointing from J7 to J9
	$\overrightarrow{J_{4}J_{5}}$	The vector pointing from J4 to J5

Table 1. Definitions of hand structural vectors $\overrightarrow{J_{19}J_{21}}$ (continue)

Related joints	Vectors	Definitions
Proximal interphalangeal joints, Metacarpophal-angeal joints	$\overrightarrow{J_{19}J_{18}}$	The vector pointing from J19 to J18
	$\overrightarrow{J_{15}J_{14}}$	The vector pointing from J15 to J14
	$\overrightarrow{J_{11}J_{10}}$	The vector pointing from J11 to J10
	$\overrightarrow{J_{7}J_{6}}$	The vector pointing from J7 to J6
	$\overrightarrow{J_{4}J_{3}}$	The vector pointing from J4 to J3
Metacarpophal-angeal joints, Fingertips	$\overrightarrow{J_{18}J_{21}}$	The vector pointing from J18 to J21
	$\overrightarrow{J_{14}J_{17}}$	The vector pointing from J14 to J17
	$\overrightarrow{J_{10}J_{13}}$	The vector pointing from J10 to J13
	$\overrightarrow{J_{6}J_{9}}$	The vector pointing from J6 to J9
	$\overrightarrow{J_{3}J_{5}}$	The vector pointing from J3 to J5
Wrist, Metacarpophal-angeal joints	$\overrightarrow{J_{0}J_{18}}$	The vector pointing from J0 to J18
	$\overrightarrow{J_{0}J_{14}}$	The vector pointing from J0 to J14
	$\overrightarrow{J_{0}J_{10}}$	The vector pointing from J0 to J10
	$\overrightarrow{J_{0}J_{6}}$	The vector pointing from J0 to J6
	$\overrightarrow{J_{0}J_{3}}$	The vector pointing from J0 to J3
Palm, Fingertips	$\overrightarrow{J_{1}J_{21}}$	The vector pointing from J1 to J21
	$\overrightarrow{J_{1}J_{17}}$	The vector pointing from J1 to J17
	$\overrightarrow{J_{1}J_{13}}$	The vector pointing from J1 to J13
	$\overrightarrow{J_{1}J_{9}}$	The vector pointing from J1 to J9
	$\overrightarrow{J_{1}J_{5}}$	The vector pointing from J1 to J5
Fingertips	$\overrightarrow{J_{21}J_{17}}$	The vector pointing from J21 to J17
	$\overrightarrow{J_{17}J_{13}}$	The vector pointing from J17 to J13
	$\overrightarrow{J_{13}J_{9}}$	The vector pointing from J13 to J9
	$\overrightarrow{J_{9}J_{5}}$	The vector pointing from J9 to J5

Based on these 24 vectors, the characteristics of hand gestures and trajectories are changed for dynamic gesture recognition. This paper extracts two characteristics of hand shape and trajectory. The two types of features are introduced separately as following.

2.1 Hand Shape Feature

Taking the difference in the size of different human hands into account, for the extraction of distance features, it is necessary to introduce a scaling factor. It is observed that the coordinates of J0 and J1 are relatively stable, so that the scaling factor corresponding to the distance feature is defined as:

$$d = \left\| J_1 - J_0 \right\| . \tag{1}$$

Fingertip distance. The fingertip distance represents the distance from the fingertip to the palm [3], which characterizes the degree of bending of each finger and the degree of hand engagement. The definition is shown in Table 2.

Table 2. Definition of fingertip distance

Fingertip distance	Definition
$D_1 = \left\|\overrightarrow{J_1J_{21}}\right\| / d$	Distance from little finger to the palm
$D_2 = \left\|\overrightarrow{J_1J_{17}}\right\| / d$	Distance from ring finger to the palm
$D_3 = \left\|\overrightarrow{J_1J_{13}}\right\| / d$	Distance from middle finger to the palm
$D_4 = \left\|\overrightarrow{J_1J_9}\right\| / d$	Distance from index finger to the palm
$D_5 = \left\|\overrightarrow{J_1J_5}\right\| / d$	Distance from thumb to the palm

Adjacent fingertips distance. The adjacent fingertip distance refers to the distance between each fingertip and its adjacent fingertip [4], which characterizes the degree of separation between adjacent fingers. The definitions of adjacent fingertips distance are shown in Table 3.

Table 3. Definition of distance between adjacent fingertips

Adjacent fingertips distance	Definition
$L_1 = \left\|\overrightarrow{J_{21}J_{17}}\right\| / d$	Distance from little fingers' fingertip to ring fingers' fingertip
$L_2 = \left\|\overrightarrow{J_{17}J_{13}}\right\| / d$	Distance from ring fingers' fingertip to middle fingers' fingertip
$L_3 = \left\|\overrightarrow{J_{13}J_9}\right\| / d$	Distance from middle fingers' fingertip to index fingers' fingertip
$L_4 = \left\|\overrightarrow{J_9J_5}\right\| / d$	Distance from index fingers' fingertip to thumbs' fingertip

Joint Angle. The joint angle represents the angle corresponding to each joint point of the hand, which characterizes the degree of bending of each finger. The definitions of joint angle are shown in Table 4.

Table 4. Definition of joint angle

Joint angle	Definition
$\alpha_1 = < \overrightarrow{J_{19}J_{21}}, \overrightarrow{J_{19}J_{18}} >$	The angle of J19
$\alpha_2 = < \overrightarrow{J_{15}J_{17}}, \overrightarrow{J_{15}J_{14}} >$	The angle of J15
$\alpha_3 = < \overrightarrow{J_{11}J_{13}}, \overrightarrow{J_{11}J_{10}} >$	The angle of J11
$\alpha_4 = < \overrightarrow{J_7J_9}, \overrightarrow{J_7J_6} >$	The angle of J7
$\alpha_5 = < \overrightarrow{J_4J_5}, \overrightarrow{J_4J_3} >$	The angle of J4
$\alpha_6 = < \overrightarrow{J_0J_{18}}, \overrightarrow{J_{18}J_{21}} >$	The angle of J18
$\alpha_7 = < \overrightarrow{J_0J_{14}}, \overrightarrow{J_{14}J_{17}} >$	The angle of J14
$\alpha_8 = < \overrightarrow{J_0J_{10}}, \overrightarrow{J_{10}J_{13}} >$	The angle of J10
$\alpha_9 = < \overrightarrow{J_0J_6}, \overrightarrow{J_6J_9} >$	The angle of J6
$\alpha_{10} = < \overrightarrow{J_0J_3}, \overrightarrow{J_3J_5} >$	The angle of J3

Adjacent fingers angle. The adjacent fingers angle represents the angle between adjacent fingers, which characterizes the degree of separation between adjacent fingers. The definitions of adjacent fingers angle are shown in Table 5.

Table 5. Definition of adjacent finger angle

Adjacent fingers angle	Definition
$\beta_1 = < \overrightarrow{J_{18}J_{21}}, \overrightarrow{J_{14}J_{17}} >$	The angle between little finger and ring finger
$\beta_2 = < \overrightarrow{J_{14}J_{17}}, \overrightarrow{J_{10}J_{13}} >$	The angle between ring finger and middle finger
$\beta_3 = < \overrightarrow{J_{10}J_{13}}, \overrightarrow{J_6J_9} >$	The angle between middle finger and index finger
$\beta_4 = < \overrightarrow{J_6J_9}, \overrightarrow{J_3J_5} >$	The angle between index finger and thumb

2.2 Trajectory Feature

Hand location. In this paper, the 3D trajectory point of the palm is selected as the position feature of the hand [5]. In order to ensure the position and scale invariance, the maximum and minimum normalization operations are performed on this feature [6-7].

For all trajectories at hand, the starting position of different trajectories is different, which will affect the recognition accuracy to a certain extent. To reduce the error, the current position coordinate data and the initial position coordinate data are subtracted, the obtained new displacement information is used as the final position characteristic, and the error caused by the starting position is effectively offset. The specific procedure is defined as:

$$\Delta V = (x_t', y_t', z_t') = (x_t - x_1, y_t - y_1, z_t - z_1 \mid t \in [1, N]) \cdot \quad \textbf{(2)}$$

Where, N is the number of points in a trajectory.

Spherical coordinate. Referring to the algorithm used in [8], this paper extracts spherical coordinate feature, which depicts the kinematic connectivity of the gesture and incorporates both the spatial and temporal relation simultaneously between the hand and the wrist.

To begin with, the trajectory of the wrist is defined as a root trajectory, and the palm trajectory is defined as a child trajectory. As shown in Fig. 2. Thus, the concept of relative trajectories could be derived, namely the relative motion characteristics between the root trajectory and the child trajectory.

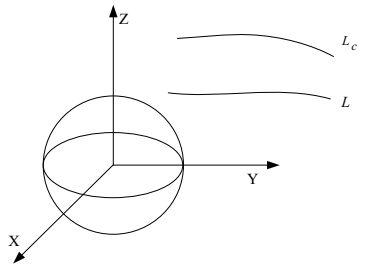


Fig. 2. Root trajectory and sub trajectory

Here, the root trajectory could be defined as $L(t) = \{x_t', y_t', z_t' \mid t \in [1, N]\}$, Similarly, the child trajectory could be defined as $L_c(t) = \{x_t'', y_t'', z_t'' \mid t \in [1, N]\}$. The relative trajectory between the root trajectory and the child trajectory is:

$$\Delta L = L - L_c = \{\Delta x_t, \Delta y_t, \Delta z_t \mid t \in [1, N]\} \cdot \quad \textbf{(3)}$$

In order to eliminate the influence of the restrictive conditions mentioned in the literature [5], the spherical coordinate system is used to describe the changes in the direction and distance of the child trajectory and the root trajectory. The relative trajectory in a spherical coordinate system could be expressed as:

$$\Delta L = \{\theta_t, \phi_t, r_t \mid t \in [1, N]\} \cdot \quad \textbf{(4)}$$

Where, θ_t is inclination from the Y axis, ϕ_t is azimuth from the Z axis in the XZ plane, r_t is the radius, and N is the number of frames.

As shown in Fig. 3. Thus, the spatial relative trajectory could be uniquely represented by a tilt angle, a direction angle and a distance.

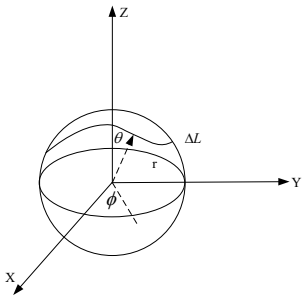


Fig. 3. Spherical coordinates of the trajectory

3 Classification Algorithm Design

Dynamic gesture recognition is essentially a classification problem. The key to the classification problem lies in obtaining a classifier with a good generalization effect. Ensemble learning provides a solution to the problem. The core idea is to first train some base classifiers, and then adopt appropriate strategies to integrate the classification results to obtain integrated classification results. Theoretical and experimental studies show that ensemble learning could improve the generalization ability of the base classifier. An integrated system that is integrated by complementary and accurate classifiers is superior to the best performing base classifier [9]. Therefore, the task of this section is to construct complementary and accurate base classifiers and adopt appropriate strategies to integrate the classification results of the base classifiers. The construction of the base classifier and the integration strategy are described below.

3.1 Constructing Base Classifiers

Due to the characteristics of being able to utilize multiple features and context information, this paper uses CRF model to construct the base classifier. Given a set of observation sequences $X = \{x_1, x_2, ..., x_N\}$, the corresponding state sequence is $Y = \{y_1, y_2, \ldots, y_N\}$. The posterior probability $p(Y|X)$ could be directly modeled by using conditional random fields. Given an undirected graph $G(V, E)$, V is the vertex in the graph, and the edge in the graph is E. Under conditions given by the observation X, the conditional distribution of Y satisfies the Markov nature:

$$p\left(y_i \middle| X, y_j, j \neq i, j \in V\right) = p\left(x_i \middle| y, x_j, j \in \eta_i\right). \tag{5}$$

Where η_i is the neighborhood of a node, namely the set of nodes connected to the edge of the node in the graph.

Given a set of input nodes $X = \{x_1, x_2, ..., x_N\}$, the probability of a corresponding set $Y = \{y_1, y_2, ..., y_N\}$ of output nodes is:

$$p\left(Y \middle| X, \theta\right) = \frac{1}{Z\left(X, \theta\right)} \exp\left\{\sum_{c \in C} \phi_c\left(Y_c, X, \theta\right)\right\}. \tag{6}$$

Where $Z(X, \theta)$ is and normalized function, $Z(X,\theta) = \sum_x \exp\left\{\sum_{c \in C} \phi_c\left(Y_c, X, \theta\right)\right\}$, $\phi_c\left(Y_c, X, \theta\right)$ is the potential functions with parameters θ defined on the group c.

When considering only CRF models for single position and pairwise potential functions, Equation (7) could be written as:

$$p\left(Y\left|X,\theta\right.\right)=\frac{1}{Z\left(X,\theta\right)}\exp\left\{\sum_{i=1}^{N}\phi_i\left(y_i,X,w\right)+\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\phi_{ij}\left(y_i,y_j,X,v\right)\right\}. \tag{7}$$

Potential functions need to be defined according to different applications. The commonly used potential functions include Multinomial Logistic Regression (MLR) [10-13], Boost [14-15], polynomial function [16], and so on. MLR could be adapted to more observation data without observing the normal distribution and other observational data. Therefore, the MLR is selected as a single position potential function:

$$\phi_i\left(y_i,X,w\right)=\sum_{l=1}^{L}\delta\left(x_i=l\right)\log\ p\left(x_i=l\left|X,w\right.\right). \tag{8}$$

Where

$$p\left(y_i=l\left|X,w\right.\right)=\begin{cases}\dfrac{exp\left(w_l^T x_i\right)}{1+\sum_{k=1}^{L-1}exp\left(w_k^T x_i\right)}, if\ \ l<L\\ \dfrac{1}{1+\sum_{k=1}^{L-1}exp\left(w_k^T x_i\right)}, if\ \ l=L\end{cases}. \tag{9}$$

Where, L is the total number of gestures, is the parameter vector of k the class gesture, D is the dimension of x_i observation. w is a $D\times(L-1)$ dimensional vector composed of $w_k, k=1,2,...,L-1$.

The pairwise potential function could express the context information of the gesture sequence. In this paper, the generalized Ising/Potts model as shown in equation (10) is selected as the pairwise potential function model [17]:

$$\phi_{ij}\left(x_i,x_j,X,v\right)=\sum_{k,l\in\{1,2,\cdots,L\}}v_{kl}^{\mathrm{T}}\cdot g_{ij}\left(X\right)\delta\left(y_i=k\right)\delta\left(y_j=l\right). \tag{10}$$

Where $g_{ij}(X)$ is a position vector pair (i, j) extracted from the entire observation data and a B dimensional feature vector, v_{kl} is a parameter vector, v is an dimensional parameter vector composed of $v_{kl}, k, l=1,2,...,L$.

Before using the CRF model for gesture recognition, the potential function parameter θ in the model need to be learned from the training set. In this paper, the maximum likelihood estimation method is used to obtain the CRF model parameters. Assuming that the training data contains N samples, the corresponding log-likelihood function could be expressed as:

$$L\left(\theta\right)=\sum_{i=1}^{N}\log p\left(Y_i\left|X_i,\theta\right.\right). \tag{11}$$

Where, X_i is a sequence of observations in the training set; Y_i is the corresponding label for the sequence of observations. In order to avoid over-fitting problems, penalty functions are usually introduced and penalty terms $\frac{1}{2\sigma^2}\|\theta\|$ are used. The log-likelihood function is expressed as:

$$L\left(\theta\right)=\sum_{i=1}^{N}\log p\left(Y_i\left|X_i,\theta\right.\right)-\frac{1}{2\sigma^2}\left\|\theta\right\|. \tag{12}$$

For formula (12), the gradient descent method could be used to look for the optimal parameter θ.

Afterwards, the set of solution Y^* that maximize the conditional probability $p_\theta(Y|X)$ could be solved according to CRF model parameter θ and the observation sequence. The problem is essentially equivalent to the inference problem on the probability map model, which can be solved by the LBP algorithm [18-19].

In the part of feature extraction, hand shape and trajectory features are extracted. In order to build complementary and accurate base classifiers, four different subsets of the origin feature set is selected as input for training the CRF model to obtain four different base classifiers. The composition of these four sets of feature subsets are: fingertip distance + trajectory, adjacent fingertips distance + trajectory, joint angle + trajectory and adjacent fingers angle + trajectory.

3.2 Combination Strategy

In this paper, the weighted voting method is performed to combine the classification results of each base classifier. Assume that the base classifier is h_i, the set of class labels is $\{C_1, C_2, ..., C_N\}$, and the output of h_i on the sample is represented as an N-dimensional vector $(h_i^1(x);h_i^2(x);...;h_i^N(x))$, where $h_i^j(x)$ is the output of h_i on the class label C_j. The core idea of the weighted voting method is to complete the final prediction of the integration of the classifier by assigning a larger weight to the better base classifier. There are two types of $h_i^j(x)$ in real-world tasks: class tags and class probabilities. In this paper, the type of $h_i^j(x)$ is a class probability, and $h_i^j(x) \in [0,1]$. The output class label by using the weighted voting method is:

$$Class(x) = \arg \max_{j=1,2,...,N} \sum_{i=1}^{T} w_i h_i^j(x) . \tag{13}$$

Where, w_i the weight of h_i, the value range is $w_i \in [0,1]$. T is the number of base classifiers, in this paper, $T = 4$.

The weight of each classifier was set proportional to its classification accuracy, namely:

$$w_i \propto \log \frac{\alpha_i}{1-\alpha_i} . \tag{14}$$

Where, α_i is the accuracy of each base classifier.

To more clearly describe the process of dynamic gesture recognition, Fig. 4. shows the corresponding flow diagram.

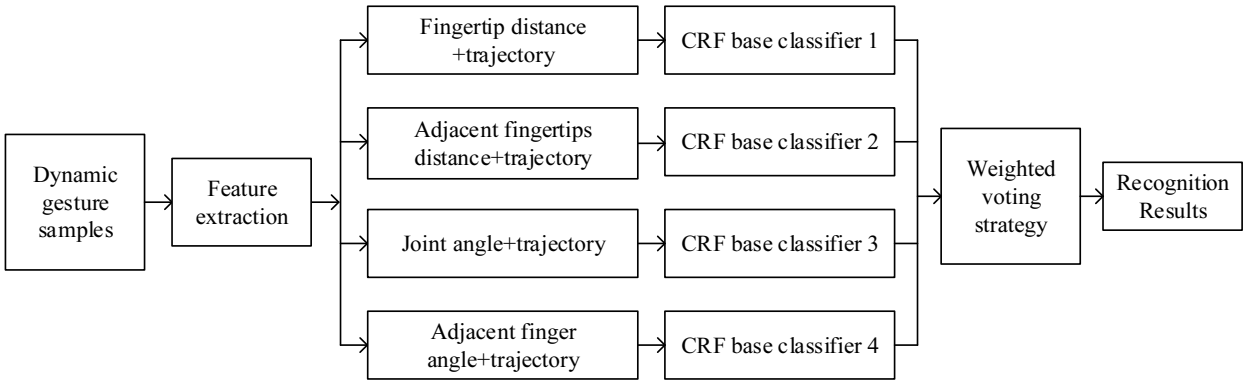


Fig. 4. Flow diagram of dynamic gesture recognition

4 Experimental Results and Analysis

4.1 Dataset and Implemental Details

In order to verify the method of this paper, 16 American dynamic gesture data sets according to the common operations of virtual reality/augmented reality are constructed. As shown in Fig. 5, twenty volunteers have joined the dynamic gesture data collection, each person uses three different viewpoints in Fig. 6 to perform the gesture several times. This dataset consists of a total of 20,000 sample gesture

sequences, each of which has a frame count of 20 to 50 frames. A random sequence of 10,000 gesture actions is used for training, and the remaining 10,000 gesture action sequences are used for testing.

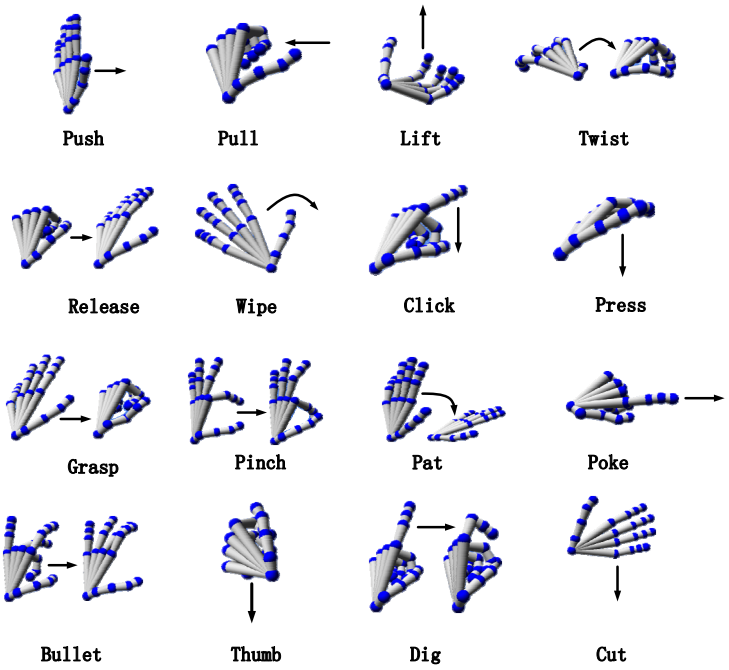


Fig. 5. Dynamic gesture dataset in the proposed paper

(a) Frontal view (b) Side view 1 (c) Side view 2

Fig. 6. Gestures of three different views

In this experiment, each base classifier is used to train and test the samples, and then four overall classification accuracies were obtained, and then four classification accuracies were normalized to get the final weight of the base classifier, the values are 0.254, 0.241, 0.258 and 0.247 respectively.

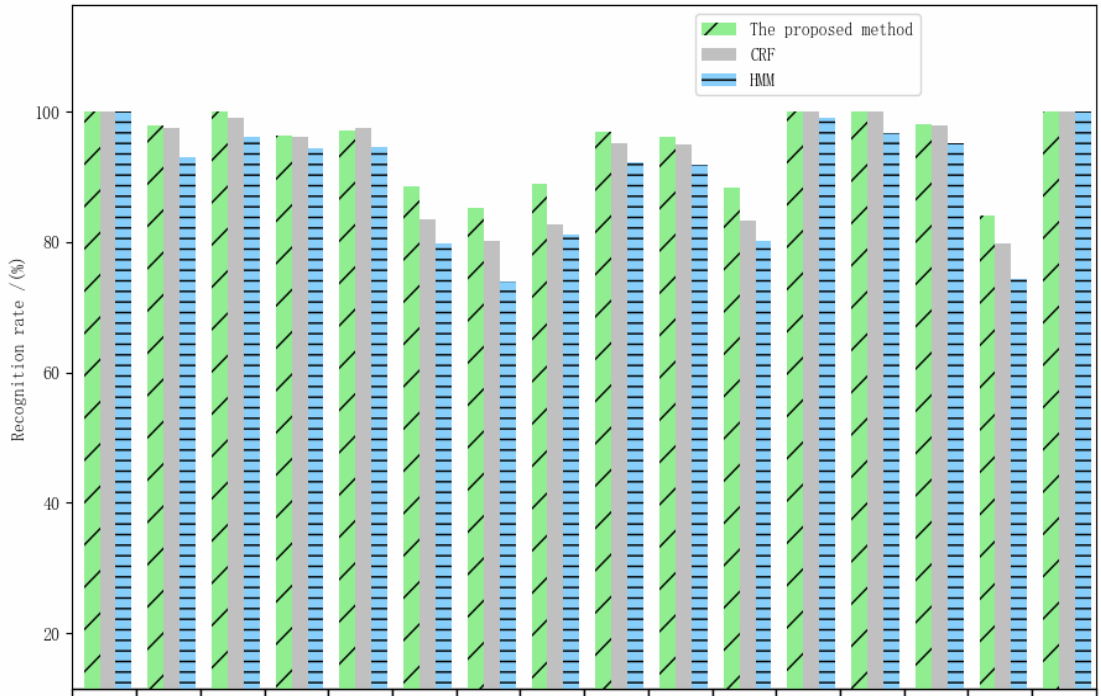
4.2 Comparison of the Proposed Method with CRF and HMM

Analysis of experimental results. Table 6 shows the average recognition rate for different recognition methods using the same feature data as input. As could be seen from Table 6, the recognition rates of the proposed method and the method based on a single CRF classifier are higher than that based on HMM, which indicates that the CRF model structure could better relate context information contained in dynamic gestures.

Table 6. Average recognition rate for different recognition methods using the same feature data

Method	The average recognition rate / (%)
The proposed Method	94.9
CRF	93
HMM	90.1

Fig. 7 shows the comparison of the recognition rate of each method with the CRF-based method and the HMM-based method when the same feature data is input. As shown in Fig. 7, for a few gestures (push, poke, and cut), the recognition rate for the proposed method, CRF-based method, and HMM-based method is comparable. For most of the gestures, the proposed method and the CRF-based method are higher than the HMM-based method. Compare the proposed method with CRF-based method. For the hand gestures and gestures with large differences in trajectory change from other gestures (push, pull, lift, twist, release, poke, bullet, thump and cut), the recognition rate of the proposed method is close to that of the method based on a single CRF classifier. For gestures that are somewhat similar in hand shape and trajectory change (grasp and pinch), the recognition rate of the proposed method is slightly higher. For hand gestures with similar hand and trajectory changes (wipe, press and pat, click and dig), the recognition rate of in this paper is much higher than that based on a single CRF classifier. This shows that after using the weighted voting method to fuse the different classifiers, the problem of feature interference and overfitting are effectively solved.



	Push	Pull	Lift	Twist	Release	Wipe	Click	Press	Grasp	Pinch	Pat	Poke	Bullet	Thump	Dig	Cut
The proposed method	100	97.9	100	96.4	97.1	88.6	85.2	89	97	96.2	88.4	100	100	98.1	84.1	100
CRF	100	97.5	99	96.2	97.5	83.5	80.1	82.7	95.1	94.9	83.3	100	100	97.9	79.8	100
HMM	100	93.1	96.1	94.3	94.6	79.7	73.9	81.1	92.2	91.9	80.1	99	96.8	95.2	74.3	100

Fig. 7. Average recognition rate of different recognition methods

Deeper analysis by confusion matrix. The result of the experiment can be more deeply analyzed by using the confusion matrix. As shown in Fig. 8, the result of this experiment is analyzed with this method. In the confusion matrix, the rows represent a dynamic gesture that is actually recognized. The columns represent the recognized gesture after identification. The values on the diagonal represent the recognition accuracy of each gesture, and the other non-diagonal values represent the probability of a corresponding gesture that predicts the error. In Fig. 8, the value of the diagonal is mostly above 0.90, which shows that most of the gesture recognition is still more accurate. Some gesture recognition rates are below 0.90, such as Wipe and Pat. Their gestures may easily recognize the gesture of the other party, probably because their initial movements are similar. Similarly, the actions of Click and Dig are easy to be confused and easily recognized as the gesture of the other party. The recognition accuracy of Press is 0.89, and their actions are more similar to those of Twist and Pat, so they can easily be recognized as these gestures.

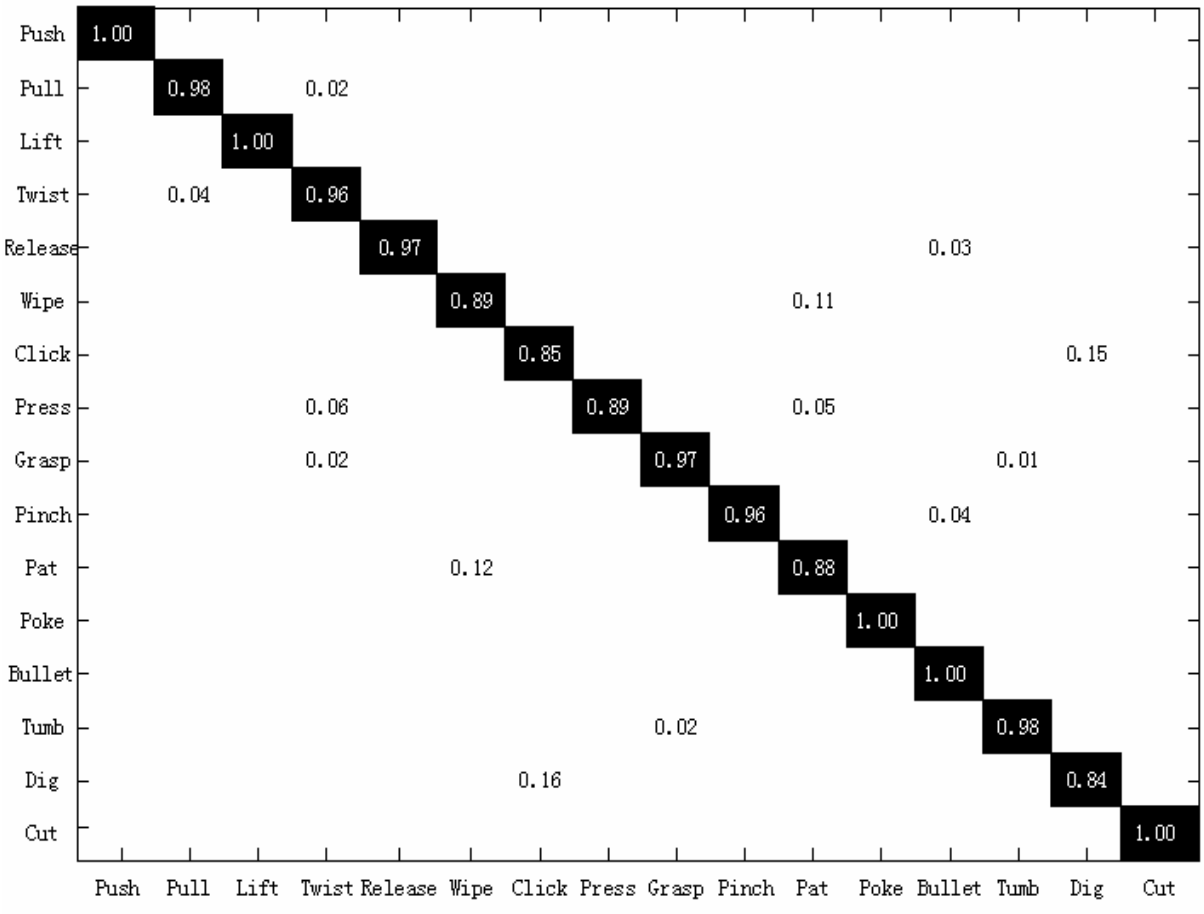


Fig. 8. The resulting confusion matrix of the proposed approach

4.3 Comparison between the Proposed Method and Other Method

In addition, the proposed method of this paper is compared with the method of the literature [1]. Table 7 shows the corresponding recognition rates of the two methods. As could be seen from Table 7, the recognition rate of this method is better than that of the method of the literature [1]. The method of the literature [1] could not fully describe the dynamic gesture due to the lack of trajectory features extraction, which affects the recognition rate to some extent.

Table 7. Comparison of the proposed method and the method of the literature [1]

Method	The average recognition rate/ (%)
The proposed method	94.9
HCNF combine hand shape features	88.5

5 Conclusion

In this paper, a dynamic gesture recognition method based on conditional random fields and weighted voting is proposed. Based on the static hand shape feature, the method adds the trajectory feature, and then combines the conditional random field model and the weighted voting method to identify the gesture. Compared with the traditional method, the proposed method has the characteristics of utilizing multiple features and contacting context information. It can also solve the interference problem and over-fitting problem between features. The experimental results show that the proposed method has excellent recognition performance. However, in the study of gesture recognition, only one hand is considered. In contrast, the dynamic gestures of the hands are more diverse, and there are many different combinations to interact with each other to design a more diverse semantics.

Acknowledgements

This work was supported by the research project of Jilin provincial education department (The research of the key technology of short-term power prediction of photovoltaic power generation) and the project of Jilin provincial science and technology development plan (The research and application of the key technology of mine safety Internet of things).

References

[1] W. Lu, Z. Tong, J.-H. Chu, Dynamic hand gesture recognition with leap motion controller, IEEE Signal Processing Letters 23(9)(2016) 1188-1192.

[2] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. 18th Conference on Machine Learning, 2001.

[3] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with jointly calibrated leap motion and depth sensor, Multimedia Tools and Applications 75(22)(2016) 14991-15015.

[4] J. Liu, J. Huang, D. Han, F. Tian, Template matching algorithm for 3D gesture recognition, Journal of Computer-Aided Design & Computer Graphics 28(8)(2016) 1365-1372.

[5] L. Geng, X. Ma, H. Wang, J. Gu, Y. Li, Chinese sign language recognition with 3D hand motion trajectories and depth images, in: Proc. IEEE Conference on Intelligent Control and Automation, 2014.

[6] M. Yang, X. Huang, X. Su, Study onultra-short term prediction method of photovoltaic power based on ANFIS, Journal of Northeast Electric Power University 38(4)(2018) 14-18.

[7] M. Yang, Q. Zhang, The research of ultra short-term wind power prediction error distribution based on nonparametric estimation, Journal of Northeast Electric Power University 38(1)(2018) 15-20.

[8] Z.-P. Shao, Y.-F. Li, A new descriptor for multiple 3D motion trajectories recognition, in: Proc. IEEE International Conference on Robotics and Automation, 2013.

[9] B. Sun, J.-D. Wang, H.-Y. Chen, Y.-T. Wang, Diversity measures in ensemble learning, Control and Decision 29(3)(2014) 385-395.

[10] P. Zhong, R. Wang, Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery, IEEE Transactions on Neural Networks and Learning Systems 25(7)(2014) 1319-1334.

[11] M. Yang, C.-L. Yang, Uncertainty research of wind power prediction considering wind direction, Journal of Northeast Electric Power University 38(5)(2018) 9-15.

[12] C.-H. Lee, S. -J. Wang, F. Jiao, D. Schuurmans, R. Greiner, Learning to model spatial dependency: semi-supervised discriminative random fields, in: Proc. 20th Conference on Neural Information Processing Systems, 2006.

[13] S. Kumar, M. Hebert, Discriminative random fields, International Journal of Computer Vision 68(2)(2006) 179-201.

[14] W. Jiang, S.-F. Chang, A.-C. Loui, Context-based concept fusion with boosted conditional random fields, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.

[15] Y.-F. Huang, J. Chen, J.-F. Liu, et al., Mechanism analysis and simulation of forward and flyback DC-DC converter, Journal of Northeast Electric Power University 38(4)(2018) 35-39.

[16] T.-M.-T. Do, T. Artières, Polynomial conditional random fields for signal processing, in: Proc. European Conference on Artificial Intelligence, 2006.

[17] S. Kumar, M. Hebert, A hierarchical field framework for unified context-based classification, in: Proc. IEEE International Conference on Computer Vision, 2005.

[18] R. Fabio, F. Dieter, D.-W. Hugh, CRF-matching: conditional random fields for feature-based scan matching, in: Proc. 3th Conference on Robotics: Science and Systems, 2007.

[19] S. Charles, M. Andrew, An introduction to conditional random fields, Foundations and Trends in Machine Learning 4(4)(2012) 313-328.