Human Recognition Using Joint Coordinate Images (JCIs) with Convolutional Neural Network

Deyu Kong¹, Xuejun Zhang²*, Yini Wei¹, Xianfu Xu¹, Hongjie Zeng¹ and Bin Li¹

¹School of Computer and Electronic Information, Guangxi University, Nanning, China
²Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning, China

*Corresponding author email: xjzhang@gxu.edu.cn

Abstract. Human recognition with skeletal data has the advantage in detecting people without their face characteristic on image. However, the accuracy of recognition by this method is always low because it relies deeply on manual feature selection. We propose a novel human recognition method called Joint Coordinate Images (JCIs) with Convolutional Neural Network (CNN) based on the image generated from skeletal information tracked by KinectV1. In order to represent human physical skeletal characteristic, the coordinate values XYZ of human joint tracked by KinectV1 are firstly created in color image called Joint Coordinate Images (JCIs), in which the relative position of the pixels represents the skeletal structure characteristics of participants with shape in “大” structure. Secondly, a new convolution neural network classifier Lenet-5 model, which always performed well in image classification, was modified to be able to input our JCIs for human recognition. The experimental results show that human recognition using joint coordinate image and Lenet-5 network can reach the highest recognition accuracy of 90.00% on the G3D dataset, which demonstrates the feasibility to transform the skeletal coordinate information into color image for human recognition task and could be used as a complementary method to the well-known application of face recognition.

Keywords: Human recognition; Joint Coordinate Images (JCIs); KinectV1; The “大” structure; Convolution neural network.

1. Introduction

With the COVID-19 swept the world, people inevitably wear masks in daily life to avoid the spread of the epidemic, which brings unprecedented challenges to face recognition technology. People wear masks to resist virus leads to a 60% reduction of the face feature and a decrease of the accuracy of face recognition. Therefore, the researchers try to solve the problem by using other biological features. Gait recognition is a method of remote human recognition, which identifies people by unique walking pattern[1]. Gait recognition can generally be categorized into appearance-based approach and model-based approaches. Appearance-based approaches are highly affected by the environment because it need to separate people contour from the scene[2]. Model-based approaches require the design of a unified model with the characteristics of unique gait for different people. The current popular model-based method is skeleton-based gait recognition which studies the space-time characteristics of skeletal movement during walking for model building. Skeleton features have the advantages of non-intrusiveness, difficulty in copying and stealing, and therefore are widely used in remote authentication. As shown in Figure 1, the red box selects the face, and the orange box selects the human skeleton. It is obvious that it is difficult to obtain all the features of human face when people...
wear masks, but it is easy to capture the skeleton of human when the device(such as Kinect) supports skeleton tracking.

![Figure 1](image.png)

**Figure 1.** The facial and skeleton feature with people wears masks on their faces.

Human recognition technology based on skeletal data belongs to model-based recognition method. The most intuitive feature of skeleton is different people have different anthropometric skeletal features, such as skeletal length, skeletal ratio, etc. Athiwat [3] extracted the length, height, head tilt angle and other features of body parts for human recognition. Li [4] extracts the Euclidean distance between joints, calculates height for human classification. Hu [5] calculated the distance between the joint point and the center of mass for human classification. However, these methods rely deeply on manual feature selection can only obtain a few features, which is only suitable for small-scale human recognition and cannot achieve large-scale identity authentication.

Human recognition belongs to the field of pattern recognition and needs to be input into the classifier for classification after extracting the characteristics of human skeleton. The traditional classifiers used in skeleton-based human recognition are support vector machine classifier [6], nearest neighbor classifier [7], and multilayer perceptron classifier [8]. The most commonly used neural network for human recognition are convolutional neural network [9] and recurrent neural network [10]. Among them, convolutional neural network is widely used in image classification.

In this work, we propose a human recognition method using skeletal data traced by KinectV1 which can not only solve the low accuracy of traditional skeleton-based human recognition, but also assist face recognition to achieve authentication in epidemic situation conveniently. The contribution of our research is as follow: 1) a coordinate representation method of RGB channel is introduced. In RGB channel, the raw skeleton data is normalized and its relative relationship is protected. 2) a method of image representation of human skeleton structure is proposed. We use a “大” structure to preserve the characteristics of the human skeleton composed of arms, legs and trunk, which makes the human representation ability of the model more stronger.

2. **Proposed Method**

A novel human recognition method based on skeletal information traced by KinectV1 is proposed. Figure 2 shows the flow chart of the research which mainly involves three processes: data pre-processing, feature extraction and classification. In the data pre-processing, the coordinate values XYZ of joints are transformed into RGB space in the range of 0 to 255 to realize the mapping from spatial coordinates to RGB channel. Secondly, the shape in “大” structure is adopted for the layout of pixels to store the relative positions of joints as much as possible to obtain the joint coordinate images. In feature extraction and classification, the modified lenet-5 network is employed to feature extraction and human recognition of joint coordinate images.
Skeleton sequences

Data preprocessing

"大" structure

XYZ convert to RGB

Figure 2. Process overview of the proposed human recognition technique.

2.1. Skeletal Information Represents by RGB Channels

Yong [11] proposed an action recognition method which converts skeletal information into images and achieves 100% recognition of human action on Berkeley MHAD dataset. Inspired by the above article, our study proposes a method of carrying skeletal data with color images for human recognition. This method uses different pixels to store the joint coordinates, which realized the conversion from the discrete coordinate to the joint pixels. The implementation of our study is as follows.

![Diagram](image)

As show in figure 3, a skeleton frame tracked by KinectV1 contains 20 joint points. Therefore, the coordinate values XYZ of joints can be stored in the channels RGB of color image one by one and each pixel of the image represents the spatial coordinates of joint. Assuming that the pixels of the image is $p$, the transformation from skeletal information to channels RGB can be realized after the application of formula (1).
\[ p = \text{floor}(255 \times \frac{c - c_{\text{min}}}{c_{\text{max}} - c_{\text{min}}}) \]  

(1)

Where \( p \) represents the transformed pixel value, which contains the coordinate information of joints and also called joint coordinate pixel. \( c_{\text{max}} \) represents the maximum coordinate value of the skeletal joints in a frame, \( c_{\text{max}} = \max(x_i, y_i, z_i), i = 0, 1, 2, 3, \ldots, 19 \). \( c_{\text{min}} \) represents the minimum coordinate value of the skeletal joints in a frame, \( c_{\text{min}} = \min(x_i, y_i, z_i), i = 0, 1, 2, 3, \ldots, 19 \). \( \text{floor} \) means rounding down.

2.2. Joint Coordinate Image with Shape in “大” Structure

The most intuitive feature of skeleton-based human recognition is different people have different skeletal anthropometric feature, such as skeletal length, skeletal ratio, etc. In order to preserve the relative position and natural connection between joints and the most primitive human skeleton structure in color image, this paper designs a “大” structure to solve the pixel layout problem of joint coordinate pixels. After the transformation using formula (1), joint coordinate pixel will be written into the color image with the size of 9*9*3 and the shape in "大" structure in turn.

![Figure 4. Character's joint coordinate image.](image)

A joint coordinate image (JCI) with the size of 9*9 is showed in figure 4 which generated from skeletal frame. As shown in the figure, the black arrow indicates the pixel position of head joint, left knee joint and right shoulder joint. KinectV1 requires people to stand within the range of 1 to 4 meters from it when tracking human skeleton. Therefore, the values of the blue channels corresponding to the values of the coordinate Z and the distance from camera are much larger than that of other two channels. So, the overall tone of the joint coordinate image is biased towards blue.

2.3. Feature Extraction and Classification Based on the Modified Lenet-5 Network

In our study, Lenet-5\(^{[12]} \) model is selected as the network for feature extraction and classification. Because the image set of our experiment is color image with the size of 9*9, which is different from the data set of handwritten numeral recognition, our study modified the lenet-5 model to make it more suitable for human classification based on joint coordinate image.

![Figure 5. Structure of the modified Lenet-5 network.](image)
As shown in the figure 5, the feature extraction stage of the CNN training model in this experiment has three convolution layers and two down-pooling layers. The adaptive filters in the convolution layer are 4*4 size filters and the down-pooling layer based on the maximum pooling principle ensured the weight of important joint features in classification is larger. The last two layers use a fully connected neural network for classification. The first fully connected layer contains 120 neurons and the last layer is just equal to the number of persons that intend to classify. In our work, the network is designed for 10 classifications. All the activation functions in the modified model use ReLu function, so the loss function $L(x)$ in this experiment is,

$$L(x) = - \sum_{m=0}^{M-1} \ln \sum_{k=0}^{C-1} \delta(k-r) p(C_k | x_m)$$

(2)

In the above formula, $M$ represents the total number of images in the training set of joint coordinate images, and $C$ represents the total number of categories of character classification. $\delta(.)$ represents the Cronek function, $p(C_k | x_m)$ means the probability that the sample $x_m$ belongs to the person $C_k$.

3. Result and Discussion

Although our method takes gait recognition as the theoretical basis, it focuses more on the represent ability of static skeleton features rather than gait features. Therefore, G3D dataset with greater difficulty is selected as the data source of human recognition. G3D dataset\(^{[13]}\) is an action recognition dataset collected by KinectV1. The dataset contains 20 actions of 10 people, such as walking, running, jumping, etc. In the experiment, our study mainly selected the skeletal dataset of 10 people who bowling as the source data. In bowling dataset, Kinect collects skeletal data at the speed of 30 frames per second and each person has 3000 frames approximately.

Our research method will be described as JCI-CNN algorithm in the following. Firstly, the skeleton data of bowling dataset is transformed into RGB channel by formula (1), and then the joint coordinate image(JCI) is generated by the method of part 2.2. Then, all JCIs are divided into training set and test set, and the label of the image is the name of participants. In the model training stage, the learning rate of the modified lenet-5 network is set to 0.01, the training batch size is 128, the optimizer is Adam optimizer, and then the training set is input into the model for classifier training. Several evaluation criterias commonly used in machine learning are introduced in our experiment: accuracy, precision, recall and F1score. This study tested the generalizatation ability of the classifier under different test set sizes, in which the test set images were 40, 80 and 100 respectively. The experimental data are shown in Table 1.

| Test set Size | Accuracy | Precision | Recall  | F1score |
|--------------|----------|-----------|---------|---------|
| 40           | 90.00%   | 89.99%    | 84.97%  | 87.40%  |
| 80           | 90.00%   | 89.98%    | 84.99%  | 87.42%  |
| 100          | 90.00%   | 89.99%    | 84.98%  | 87.41%  |

It can be seen that the accuracy of the classifier is stable at 90%, the accuracy is stable at 89.98% \(\pm\) 0.01%, the error of recall rate is stable at 84.97% \(\pm\)0.01%, and the F1score is 87.41% \(\pm\)0.01%, which shows that the generalization performance of the trained model is stable. The reason is that JCIs stores the relative original data of skeleton and the relative positions of joints. With the assistance of CNN classifier, these two feature types can recognize up to 90% of people, and the stability ratio of the classifier is high.

In order to explore more classifiers, this study further uses SVM for human recognition. SVM is a commonly used classifier in machine learning. It can transform samples which are difficult to be linearly segmented into high-dimensional space to achieve linear separability. In high-dimensional feature space, based on the theory of structural risk minimization, the optimal segmentation hyperplane is constructed in order to make the classifier get global optimization and realize the classification of samples. Since the input sample of SVM is a one-dimensional vector, two different
methods are used to process JCI images. 1) All JCI pixels are flatten to get a vector JCI (243) with dimension 243. 2) Only 20 joint pixels containing joint points in JCI are selected, and then the pixels are flattened to get a vector JCI (60) with dimension of 60. In this experiment, the kernel function of SVM is radial basis function.

**Table 2.** The experimental results are compared with SVM.

| Methods      | Accuracy |
|--------------|----------|
| JCI(243)-SVM | 60.00%   |
| JCI(60)-SVM  | 64.44%   |
| **JCI-CNN**  | **90.00%** |

The above experimental results for 10 people are shown in Table 2. It can be seen that JCI (243)-SVM has the lowest recognition rate with 60%, followed by JCI (60)-SVM with 64.44%. The reason is that only 20 pixels in JCI contain joint information, and the rest pixels are redundant information, so the feature vector JCI (243) contains a lot of useless information, so the accuracy of human recognition is lower than that of JCI (60)-SVM. However, JCI-CNN algorithm using convolution layer not only to extract the position relationship between the joint pixels, but also to calculate the number relationship between the joint points, so JCI-CNN methods obtained the highest accuracy of human recognition.

**Figure 6.** CMC curve of the JCI-CNN algorithm and SVM with 10 ranks, 10 people in G3D data set. Cumulative Match Characteristic (CMC) curve is an evaluation index of pattern recognition, which is often used to evaluate the performance of biometric recognition. As shown in figure 6, it is the CMC curve of JCI (60)-SVM, JCI(243)-SVM and JCI-CNN. We can see that JCI-CNN can achieve 90% accuracy in Rank1 and 100% accuracy in Rank4. The accuracy of JCI(60)-SVM is only 64.44% in Rank1 and 100% in Rank5. The CMC curve of JCI(243)-SVM achieves 100% accuracy only when rank6. It can be seen that CNN is more suitable than SVM as a classifier of JCIs, and JCI-CNN algorithm can achieve 100% accuracy after 4 times of human recognition according to the prediction probability.

Table 3 compares two studies using skeletal data for human recognition, Li[4] calculated the Euclidean distance and height between joints, and only achieved 82.00% accuracy. Hu[5] manually extracted the distance between the joint point and the skeletal centroid, but only got 70.00% .

**Table 3.** Comparison of recognition accuracy of different skeletal features.

| Methods | Accuracy |
|---------|----------|
| Li[4]   | 82.00%   |
| Hu[5]   | 70.00%   |
| **Ours** | **90.00%** |
While in our method, the recognition rate is up to 90.00% on the G3D data set based on the joint coordinate image, and obviously the recognition accuracy is much higher than that of the method based on manual feature extraction. Therefore, it can be concluded that the method based on the joint coordinate image is feasible and effective.

4. Conclusion
In view of the current situation of COVID-19, the accuracy rate of face recognition is low as people wear masks on their faces. And the traditional skeletal-based human recognition which can used as a complementary method to face recognition deeply relies on manual feature selection led to low accuracy. Hence, a novel human recognition method based on skeleton called Joint Coordinate Images (JCIs) with Convolutional Neural Network (CNN) is introduced in our study. our study constructs joint coordinate images to represent the skeleton coordinates for human identification, and proposes a modified lenet-5 network for image classification. This method avoids the complex process of manually feature extraction and reach the highest recognition accuracy of 90.00% on the G3D dataset for human recognition. In future work, the research will focus on the layout structure of JCI image and extracting feature from consecutive frames.

Acknowledgments
This paper was supported by the National Natural Science Foundation of China (Grant No.81760324), Guangxi Key projects of science and technology (Grants No. AA21077007), Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi, China and School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, China.

References
[1] Choi S, Kim J, Kim W, et al. 2019 Skeleton-based Gait Recognition via Robust Frame-level Matching(IEEE Transactions on Information Forensics and Security), p 1
[2] Chai T, Mei X, Li A et al 2021 Silhouette-Based View-Embeddings for Gait Recognition Under Multiple Views(2021 IEEE International Conference on Image Processing ), p 2319
[3] Athiwat C, Nirattaya K, Cholwich N 2013 View independent human identification by gait analysis using skeletal data and dynamic time warping(in Proc. 14th Internat. Symposium Advance. Intellegence System)
[4] Li H 2013 Skeletal recognition system based on Kinect skeleton tracking function(Xi’an University of Electronic Science and Technology). (in chinese)
[5] Hu Y F 2018 The research and implementation of gait recognition based on Kinect multi-feature fusion(Jinan University). (in chinese)
[6] Zhang Y, Wang L, Wu Q et al. 2019 Review of Gait Recognition Methods(2019 IEEE 9th International Conference on Electronics Information and Emergency Communication) p 166
[7] Pratama F I and Budianita A 2020 Optimization of K-Nn Classification In Human Gait Recognition(2020 Fifth International Conference on Informatics and Computing) p 1
[8] Sun J, WangY F, Li J, et al. 2018 View-invariant gait recognition based on kinect skeleton feature vol 77 (Multimedia Tools and Applications) p 24909
[9] Piya L, Nirattaya K, Cholwich N 2020 View-Independent Gait Recognition Using Joint Replacement Coordinates (JRCs) and Convolutioanal Neural Network vol 15 (IEEE Transaction on Information Forensic and Security) p 3430
[10] Jun K, Lee D W, Lee K, et al. 2020 Feature Extraction Using an RNN Autoencoder for Skeleton-based Abnormal Gait Recognition vol 8 (IEEE Acces) p 19196
[11] Yong D, Yun F, Ling W 2015 Skeleton based action recognition with convolutional neural network(2015 3rd IAPR Asian Conference on Pattern Recognition)
[12] Lecun Y, Bottou L 1998 Gradient-based learning applied to document recognition vol 86 (Proceedings of the IEEE) p 2278
[13] G3D dataset. http://dipersec.king.ac.uk/G3D/G3D.html. Last accessed 4 Apr. 2021