Research on long-distance hand recognition based on depth information

Yuyang Fu, Lanfang Miao, Zhifei Li
College of Mathematics and Computer Science, Zhejiang Normal University, 668 Yingbin Avenue, Jinhua, Zhejiang, 321004, China
197361794@qq.com

Abstract. This paper proposes a long distance gesture recognition algorithm based on Kinect. First, we use Kinect to capture human skeleton and depth information, track and extract hand information. For the characteristics of the depth image determine that, the experimental results will be not affected by the background, light, skin color and clothing. Then the initial obtained data of hand shape is denoised and smoothed, and the contour and skeleton of hand shape are extracted. When the hand is at long distance, the accuracy of Kinect is not sufficient to get the detail hand shape-information. So we combine the hand depth information with the color information to get hand shape. Finally, we use the Hu moment of the hand shape contour binary image and the hand skeleton binary image as data feature, and utilize SVM to train and identify hand gesture. The experimental results show that the Hu moment of the hand skeleton binary image is more advantageous than the Hu moment of the hand contour binary image, and the proposed long-distance hand recognition algorithm also has the recognition accuracy similar to the close-distance.

1. Introduction
Gesture recognition has always been an important topic in the field of computer vision. Due to the flexibility of the hand, it is more important for human to express their inner thoughts and interaction with the external environment.

Lin et al.[1] proposed that Kinect offers a variety of opportunities for both new and old applications, and it can be used as a reference in gesture recognition, human activity recognition, human biological measurement assessment, 3D surface reconstruction, and healthcare applications. Thanh et al.[2] proposed an approach to extract the discriminative patterns for efficient human action recognition. And they consider each action is consisted of a sequence of unit actions, each of which is represented by a pattern. In addition, they first automatically extract the key-frames from a skeleton sequence and categorize them into different patterns.

Zhao et al.[3] presented an improved strategy for hand segmentation using the randomized decision forest framework based on depth images. In the proposed method, a new depth feature derived from the central point of hand structure is induced to strengthen the ability of generalization of depth feature as well as reduce the requirement for training dataset, while not sacrifice the accuracy of hand segmentation. Compared to traditional images, depth images can also avoid self-occlusion of the hand based on depth information[4]. With the development of deep learning, a hand segmentation method based on complete convolution network has been proposed[5].

Gesture recognition requires the extracting features of the hand data. It is usually divided into static gesture recognition and dynamic gesture recognition. Dynamic gesture recognition is mainly based on
the direction and speed of hand movement. Static gesture recognition is mainly based on hand shape judgment and it will be studied in this paper. The methods of classification recognition mainly include template matching, support vector machine (SVM)[6], neural network and so on.

In this paper, we will use Kinect to extract depth and skeleton information of hand. The Hu moment is selected as the feature of the static gesture, and the SVM classifier is used for gesture recognition. And the recognition effects between the hand contour Hu moment and the hand skeleton Hu moment of the same static gesture at close and long distances are compared by experiment.

The rest of the paper is organized as follows. In Section II, the close distance and long distance hand extraction algorithms are described in detail. In Section III, we introduce the features and tools used in gesture recognition. Section IV shows the experimental results. Concluding remarks are drawn in Section V.

2. Hand Preprocess

In this section, we will introduce the process and related algorithms for extracting hand from close distance and long distance. The flow chart of our method is shown in Figure 1. As shown in Figure 1, our method is divided into two parts: (1) Extract the hand from a close distance (within 2m); (2) Extract the hand from a long distance (beyond 2m).

2.1. Extract hand shape at a close distance (within 2m)

With the Kinect device, we can get the depth and color data streams. The pixel information in the depth image represents the calibrated distance information in the scene. In order to get the human body from the image, we can use the foreground segmentation function provided by Kinect. The main idea of this function is to use thresholds to filter the depth of the background.

In this paper, we use Kinect to extract the foreground extraction of depth image, that is, obtain the corresponding mask depth indirectly from the index number which is randomly assigned for each object by Kinect, as shown in Figure 2(b). Kinect can also track 20 joint points of the human body, so we can get the depth map and color map of the hand according to the hand joint. Since Kinect uses the principle of laser speckle to obtain hand depth image, we can find the edge of the image is rough, especially in the target edge there are many holes and noise points. If the rough image is processed directly, some errors may be caused when extracting the features later. We use the filtering method to repair defects on the image. Although this method cannot eliminate all the noise, they have considerable results[7].

Figure 1. Flowchart of the proposed method
Figure 2. (a) Depth image from Kinect; (b) Human body depth image; (c) Binary image of human hand; (d) Color image of human hand

2.2. Extract hand shape at a long distance (beyond 2m)

When the human body is far away from the Kinect camera, the hand area becomes a group in the depth image, and we can hardly distinguish the movement of the finger, so the hand recognition cannot be performed, as shown in Figure 3(a). In order to recognize the finger at a long distance, we propose a new hand extraction method based on the color image.

First, because the color camera and depth camera of Kinect are not in the same position, we need to use its conversion function to align the color image with the depth image. Then, based on the bone tracking function, we can get the range of hand areas in depth and color images. Although the hand area in the depth image cannot obtain detailed finger information, it can roughly obtain the contour of the hand. We can get the intersection of the depth image and the hand area of the color image to get Figure 3(b). According to the depth map of the hand, we can get the contour of the hand, and the centre point is the point on the hand. Therefore we use the pixel value of the centre point as a reference to recalculate the pixels in the contour to obtain Figure 3(c).

After obtaining Figure 3(c), we should calculate the value of threshold to divide the hand part (area A) and the other part (area B). We assume there are 255 values and the sum of the remaining blocks is N, the number of blocks whose depth value is \( i \) is \( n_i \), and the probability of appearance of each depth value is \( p \), then obviously:

\[
N = \sum_{i=0}^{255} n_i, \quad p_i = \frac{n_i}{N} \tag{1}
\]

The appearance probability of the area A and the area B satisfies the following equation

\[
p_A = \sum_{i=0}^{x} p_i, \quad p_B = \sum_{i=x+1}^{255} p_i = 1 - p_A \tag{2}
\]

Where \( x \) is the selected threshold.

The inter-class variance of the area A and the area B satisfies the following equation:

\[
w_A = \sum_{i=0}^{x} i p_i / p_A, \quad w_B = \sum_{i=x+1}^{255} i p_i / p_B \tag{3}
\]

\[
w_0 = p_A w_A + p_B w_B = \sum_{i=0}^{255} i p_i \tag{4}
\]

\[
\sigma^2(x) = p_A (w_A - w_0)^2 + p_B (w_B - w_0)^2 \tag{5}
\]

Where \( w_A \) and \( w_B \) represent the average pixel values of area A and area B respectively, \( w_0 \) represents the average pixel values of the global image and \( \sigma^2(x) \) represents the inter-class variance of the two areas.

The main idea of this process is to choose the value of threshold according to the pixel characteristic so as to maximize the variance between the area A and the area B. Variance is a measure of the pixel distribution uniformity. The greater inner-class variance between area A and area B, indicates that the
greater the difference between these two areas. Therefore, segmentation with the highest variance between classes means that the probability of misclassification is the smallest. The final optimal threshold \( x \) is:

\[
x = \max_{0 \leq x \leq 255} \{\sigma^2(x)\}
\]  

(6)

After calculating optimal threshold of the above equation, we can extract the gesture area as shown in Figure 3(d).

![Figure 3](image3.png)

Figure 3. (a) Original binary image of human hand; (b) Original image intersect with gray image; (c) Difference calculation; (d) Improve binary image of human hand

2.3. Hand Skeleton Extraction

Human skeleton extraction can use image thinning algorithms, which have been used for pattern recognition and image analysis for a long time. The thinning algorithm reduces the binary number pattern to obtain a unit width skeleton that maintains geometric and topological properties.

Zhang et al.[8] proposed a fast parallel thinning algorithm. It consists of two sub-items: one to remove the southeast boundary point and the northwest corner point, and the other to delete the northwest boundary point and the southeast corner point. The connectivity and insensitivity to boundary noise are very good in this method. Tarabek et al.[9] proposed a robust parallel thinning algorithm based on ZS algorithm. It maintains the good performance of the ZS algorithm and overcomes the shortcomings of not producing a unit width skeleton by removing the post-processing steps of redundant pixels by combining the additional conditions for identifying critical patterns. This paper uses this algorithm for human skeleton extraction.

![Figure 4](image4.png)

Figure 4. (a) Improve binary image of human hand; (b) Skeleton binary image of human hand

3. Gesture Recognition

3.1. Feature Extraction

Some people used the hand image Hu moment as the static gesture data feature, while Liu et al.[10] used the hand contour Hu moment as the static gesture data feature, they all used the SVM classifier to recognize the gesture. In order to explore the data characteristics with the maximum recognition rate, this paper chooses the hand contour Hu moment and the hand skeleton Hu moment as the data features of the same static gesture. Then we use the SVM classifier to identify and compare the recognition effects.

The Hu moment mainly represents the geometric features of the image region, and has the characteristics of rotation, translation, and scale invariance. The Hu moment is characterized by seven moment invariants, which are composed of a linear combination of the second and third order central moments of the image[11].

3.2. Classification and Identification tools

This paper uses SVM as a tool for gesture recognition. Support Vector Machine (SVM) is a machine learning method developed by statistics. The main idea is to map the sample space into a high-dimensional feature space through a nonlinear mapping, so that the problem of nonlinear separability in the original sample space is transformed into a linearly separable problem in the feature space.
The experiment uses the opencv-2.4.3 library to calculate the Hu moment of the image, and the SVM module inside it used to train the model.

4. Experiment Results

4.1. Gesture Recognition Experiment

The purpose of the experiment is to classify and recognize the five static gestures shown in Figure 5. The experimental development environment is Visual Studio 2013 and opencv-2.4.3 library, the programming language uses C++, and the hardware device uses Kinect for Windows 1.8.

In order to complete the experiment, the programmed program performs the following functions in order:

- Different people use five static gestures within 2m and beyond 2m respectively. We use Kinect to record depth gesture map and color gesture map separately, and process them to get corresponding hand contour map and hand skeleton image.
- After remove some of the blurred images in the sample due to frame skipping, we select 20 images of each hand contour and hand skeleton image within 2m as training samples. And the other image are test samples. We use the OpenCV to calculated Hu moment and use SVM to train model to get the rate of hand recognition.
- Using the same method, we do experiment with the images beyond 2m. Because Kinect has been unable to extract hand images beyond 2m, we use our hand extraction method to compare hand contour and hand skeleton recognition rate.

![Figure 5. Five kinds of static images of hand gesture](image)

4.2. Experimental Results and Analysis

Table 1 is the experimental result of the hand recognition rate. The total recognition rate of the hand contour image within 2m which Hu moment is used as the data feature is Accuracy = 51.73%. The total recognition rate of the data feature using hand skeleton map Hu moment is Accuracy = 79.34%.

Beyond 2m, Kinect cannot be used to recognize gestures accurately, so we use the hand recognition algorithm proposed in this paper. For the hand contour image using Hu moment as data feature, the total recognition rate is Accuracy = 68.98%. And for the hand skeleton image, the total recognition rate is Accuracy = 78.25%.

| Table 1. Hand gesture recognition rate |
|----------------------------------------|
| Recognition rate of hand gesture within 2m | Recognition rate of hand gesture beyond 2m |
| Contour | Rate/\% | Skeleton | Rate/\% | Contour | Rate/\% | Skeleton | Rate/\% |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 73.93 | 88.69 | 100 | 86.25 |
| 31.50 | 78.74 | 78.95 | 78.95 |
| 21.01 | 30.25 | 69.39 | 38.78 |
| 36.99 | 94.52 | 24.14 | 71.93 |
| 66.29 | 97.75 | 66.36 | 93.46 |
The above experimental results show that the recognition rate of the data feature of the hand skeleton image Hu moment is higher than that of the hand contour image Hu moment. The hand extraction method proposed in this paper effectively increased the hand recognition distance.

5. Conclusion and Future
In this paper, a method for extracting hand based on depth information is proposed. The Hu moment of hand contour and hand skeleton is used as data feature. We use SVM classifier to recognize five types of hand static gesture. It is found that the hand skeleton recognition rate is higher than the hand contour recognition rate.

Although the gesture recognition method implemented in the paper has a good recognition rate, the number of recognition gesture is still relatively less. At the next step, we will combine with dynamic gesture recognition or limb skeleton as the type of recognition, to realize a functionally rich somatosensory interaction system.

References
[1] Lin, Yen Yu , et al. "Depth and Skeleton Associated Action Recognition without Online Accessible RGB-D Cameras." IEEE Conference on Computer Vision & Pattern Recognition IEEE Computer Society, 2014.
[2] T.T.Thanh, F.Chen, B.Le, and B.Le,"Extraction of Discriminative Patterns from Skeleton Sequences for Accurate Action Recognition,” Fundamenta Informaticae, vol. 130, no.2, pp.247-261, 2014.
[3] Zhao, Mengyi , and Q. Jia . "Hand Segmentation Using Randomized Decision Forest Based on Depth Images." International Conference on Virtual Reality & Visualization IEEE, 2017.
[4] Kuang, Hailan , et al. "An Effective Skeleton Extraction Method Based on Kinect Depth Image." International Conference on Measuring Technology & Mechatronics Automation IEEE, 2018.
[5] Long, Jonathan , E. Shelhamer , and T. Darrell . "Fully Convolutional Networks for Semantic Segmentation." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.4(2014):640-651.
[6] Rossi, Matteo , et al. "Hybrid EMG classifier based on HMM and SVM for hand gesture recognition in prosthetics." (2015).
[7] Shen, Yujie , et al. "A Novel Human Detection Approach Based on Depth Map via Kinect." Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on IEEE, 2013.
[8] Zhang, T. Y , and C. Y. Suen . "A fast parallel algorithm for thinning digital patterns." Comm Acm 27.3(1984):236-239.
[9] Tarabek, Peter . A robust parallel thinning algorithm for pattern recognition. 2012.
[10] Liu, Yun , Y. Yin , and S. Zhang . "Hand Gesture Recognition Based on Hu Moments in Interaction of Virtual Reality." International Conference on Intelligent Human-machine Systems & Cybernetics IEEE, 2012.
[11] Huang, Zhihu Huang Zhihu , and J. L. J. Leng . "Analysis of Hu's moment invariants on image scaling and rotation." International Conference on Computer Engineering & Technology IEEE, 2010.