Research on Gesture Recognition Based on Improved Canny & K-Means Algorithm and CNN

Hai Zhang, Shixin Li* and Xiaoyu Liu

School of Electronic Engineering, Tianjin University of technology and Education, Tianjin, China

*Corresponding author e-mail: 1459399409@qq.com

Abstract. In view of the problem that the current dynamic gesture recognition system has decreased recognition rate and recognition speed due to the increase of sample size, in this study, we first use the improved Canny operator to detect the edge of the hand image. The improved algorithm includes two aspects: one is to generate HSV space image converted from RGB space and extract the V-component image; the other is to transform the traditional Gauss filter into a bilateral filter which integrates spatial distance and similarity to denoise; Then the improved K-means clustering algorithm is used to extract the feature points on the edge of the hand image. In the improved method, the pixel points at the peak of histogram are used as the initial clustering center, and the number of peak points is used as the number of classification; Then, the detection algorithm of convex hull and convex defect with geometric features is used to realize the effective fingertip tracking; Finally, CNN is used to achieve the precise sorting of fingertips. After verification, this synthesis algorithm can effectively solve the problems such as the difficulty to accurately detect the hand image contour, the poor fingertip tracking effect and the disorder of fingertip sorting caused by the similarity of background color and skin color, and the small difference between the length of fingers. With the increase of sample size, the recognition rate and speed have also improved, which can be widely used in dynamic gesture recognition and other fields.

1. Introduction

Gesture recognition can realize the naturalization and intelligence of human-computer interaction. It is widely used in computer and user interaction, human-computer interaction with robot, sign language recognition and so on. Generally, gesture recognition methods are divided into: (1) Using artificial neural network for gesture recognition. This method has the characteristics of classification and anti-interference. However, due to the weak ability of neural network to process time series, it is not ideal for dynamic gesture recognition. At present, it is only applied to static gesture recognition [1]. (2) The hidden Markov model is used for gesture recognition. A large number of state probability density and parameters to be estimated make training and recognition relatively slow [2]. (3) Hand gesture recognition technology based on geometric features has poor learning ability and efficiency. With the increasing sample size, its recognition rate has not improved significantly [3].
All the above algorithms about gesture recognition have their own scope of application. In this study, improved canny & k-means algorithm and convolutional neural network (CNN) are used for gesture recognition, which can effectively meet the requirements of real-time and accuracy. Canny edge detection is a technology to extract useful structural information from different visual objects and greatly reduce the amount of data to be processed. At present, it is widely used in image processing and computer vision [4]. Because of its simplicity and efficiency, K-means clustering has become the most widely used and famous clustering algorithm. Given a set of data points, each data point is divided into K clusters according to a certain distance function. At present, it has been widely used in data analysis, signal processing, machine learning and other fields. [5]. In recent years, convolution neural network is widely used as an efficient recognition method, which has become one of the research hotspots in many scientific fields, especially in two-dimensional image processing, machine vision and pattern recognition. Convolution neural network has been successfully used in handwritten character recognition, face recognition, human eye detection, pedestrian detection, robot navigation. Convolution neural network can recognize the changing patterns, and it is robust to simple geometric deformation. Convolution neural network has been widely studied in foreign countries, but its research and application in China is still in its infancy [6].

2. Hand gesture segmentation based on image edge detection and feature point extraction

2.1. Image edge detection algorithm based on improved canny operator

2.1.1. Canny operator correlation theory. Canny operator was proposed by Canny in 1986. Its edge detection algorithm is as follows: using Gaussian filter to filter and denoise to get smooth image; using finite difference of first-order partial derivative to calculate the gradient amplitude and direction of smooth image; using non maximum suppression for gradient amplitude; using double threshold algorithm to detect and connect edge. Because most of the image noise belongs to Gaussian noise, Gaussian filtering is widely used in image denoising. The traditional canny algorithm also uses Gaussian filter to eliminate image noise. Gaussian filtering is described by the following formula:

\[ h(x) = 1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(e)c(e, x)de \]

\[ k_d(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(e, x)de \]

Among them, C represents the Gauss weight based on the spatial distance, and kd(x) is used to unite the result.

In addition, HSV color space is more in line with people’s perception of color than RGB color space, so it is more conducive to image processing.

2.1.2. Improved canny operator. The improved algorithm includes two aspects: one is to generate HSV space image converted from RGB space and extract the V-component image; the other is to transform the traditional Gauss filter into a bilateral filter which integrates spatial distance and similarity to denoise [7]. (1) HSV space is suitable for people's visual characteristics. Firstly, HSV image from RGB color space is generated and the V component image is extracted; V-component represents the brightness information of the image, and it has nothing to do with the color information of the image, and the conversion matrix of V-component and RGB color space is: V=max(R, G, B). The V-component is not related to the other two components. Then, the image is denoised by two-sided filtering, the amplitude and direction of gradient are calculated by the first-order derivative finite difference, the gradient amplitude is suppressed by non-maximum value, and the edge is detected and connected by double threshold algorithm. After filtering, Sobel operator is used to get the vertical and
horizontal edges of the image, and double thresholds are set manually according to experience.

(2) Bilateral filter. Gauss filter only considers the spatial position relationship between pixels, which leads to the loss of edge information. For this reason, the bilateral filtering adds another weight part to keep the edge information, which is realized by the following expression:

\[ h(x) = 1/k(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\varepsilon)s(f(\varepsilon), f(x))d\varepsilon \]

(3)

\[ k(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(f(\varepsilon), f(x))d\varepsilon \]

(4)

Among them, s represents the Gaussian weight based on the pixel shear similarity, and kr(x) is used to unite the result.

Combined with the above two parts, we can get a bilateral filter which combines spatial distance and similarity:

\[ h(x) = 1/k(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\varepsilon)c(\varepsilon, x)s(f(\varepsilon), f(x))d\varepsilon \]

(5)

\[ k(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(\varepsilon, x)s(f(\varepsilon), f(x))d\varepsilon \]

(6)

Bilateral filtering includes two Gaussian basis filtering functions. Therefore, when processing the adjacent pixel values in the region, not only the geometric proximity of each pixel value is considered, but also the similarity of the brightness of each pixel value is considered. So it has double filtering function.

2.2. Improved k-means clustering algorithm for feature point extraction

2.2.1. Principle of K-means clustering algorithm. K-means clustering algorithm is based on distance similarity. The smaller the distance between the two sample data, the more similar it is. It can be divided into the same family. Finally, all the sample data with similar distance constitute different clusters, and get compact and independent categories. Firstly, K initial clustering centers are randomly selected in a given data set. Then, according to the principle of close distance, the distance between K clustering centers is calculated by appropriate distance formula. At this time, the data is divided into clusters where the nearest center is located, forming a cluster containing the cluster center and all data objects assigned to the center. After all the data are allocated successfully once, the average value of all the data objects in each cluster is calculated repeatedly to get a new cluster center, which iterates in turn until a certain termination condition is met, indicating that at this time, all the data objects are classified and K clusters are obtained. K-means clustering algorithm for image segmentation and edge feature point extraction is intuitive, fast and easy to implement. The essence of K-means clustering algorithm applied to image is to quantify the color of image.

2.2.2. Improved K-means clustering. The traditional K-means clustering algorithm usually uses random selection method. The disadvantage of such selection method is that it does not combine the characteristics of the image itself. This randomness makes it necessary to go through many iterations to find the real clustering center, which requires a lot of calculation and the clustering effect is not ideal.

In the improved method, the pixel points at the peak of histogram are used as the initial clustering center, and the number of peak points is used as the number of classification, iterating in turn until the
final clustering center is calculated, and then the pixel points are classified until the clustering is completed[8]. In this study, V-component histogram of bus image is selected to illustrate, as shown in Fig.1.

![Figure 1. V-component histogram of bus image.](image)

In the histogram of V-component, there are three obvious peaks, but considering that the first peak is close to the second peak, and the number of classification should be consistent with the number of h-component, so the number of classification is also selected as 2, and the pixel points 0.23 and 0.76 corresponding to the two peaks are taken as the initial clustering center. In this way, the K-means clustering of clustering centers can greatly reduce the amount of calculation and make the clustering results more ideal.

3. **Fingertip tracking based on convex hull and convex defect detection of contour points**

3.1. **Detection of convex hull of contour point**

First, the definition of convex hull is given: for a set D, all the linear combinations of any finite points in D are called convex hull of D, and the intersection of all convex sets containing D can also be considered.

The convex hull of point set Q refers to a minimum convex polygon, which satisfies the point in Q either on the edge or in the polygon. The polygon represented by the red line in Fig.2 is the convex hull of point set Q = \{p0, p1... p12\}.

![Figure 2. Convex hull of point set Q.](image)

So for a group of points on a plane, we can find the smallest convex polygon that contains all points, which is the convex hull problem. This can be vividly thought of as follows: place some immovable wooden piles on the ground, circle them as tightly as possible with a rope, and they are convex shaped, which is convex [9]. In this algorithm, the convex hull is detected by approxpolydp function.
3.2. Detection of contour convex defects

As shown in Fig.3, the black outline is the convex hull obtained by the `approxpolydp` function above, and the part between the convex hull line and the palm is the convex defect. Each convex defect area has four characteristic quantities: start point, end point, farthest point from convex hull line (farPoint), and the distance between furthest point and convex hull line (depth), which can be understood in combination with Fig.4 [10].

![Figure 3. Convex defect of hand image.](image)

![Figure 4. Four characteristic quantities of convex defects.](image)

Many objects in the scene can also detect convex defects. In order to detect the convex defects of the desired hand more accurately, it is necessary to filter out some irrelevant areas, such as the angle between the straight line formed by the start point and the farthest point and the straight line formed by the end point and the farthest point cannot be too large. According to the actual situation, the maximum upper limit of the empirical value is 95 degrees. Then calculate the distance PD1 from the starting point to the farthest point and the distance PD2 from the ending point to the farthest point. PD1 and PD2 should not be too short, they must be greater than 1/5 of the height of the hand. At the same time, the proportion between PD1 and PD2 should be close, not too large or too small, and the proportion should be in the range of 0.5-1.5. These steps are realized in this algorithm through `eliminatedefects` function.

In this algorithm, the function `removedundantendpoints` is used to remove the endpoint of another convex defect which is very close to the start point.

Then confirm whether it is a hand. If the number of fingertips is greater than 5, or the ratio between the height and width of the contour is too large, or the width of the contour is less than 20, it is not considered as an effective hand area. If it is an effective hand area, start to detect the fingertips, through the `getfingertips` function, as long as you traverse the detected convex defects of the hand.

After the fingertip is detected, we remove some redundant fingertips through the `removedredundantfingertips` function, that is, if the distance between two fingertips is too small or less than 10, then the redundant fingertips are removed, and the remaining fingertips are the final effective fingertips [11].
4. Finger sorting based on convolution neural network

Basic features of CNN [12]: partial connection, shared weight and downward collection of samples. Through partial connection and sharing of weights, the training parameters can be greatly reduced, and the stability of the model can be improved by down sampling. It can be seen that CNN generally consists of convolution layer and subsampling layer. As the classification task in this study is relatively simple, the simplified network structure is shown in Fig.5.

![Figure 5. Simplified network structure.](image)

The input layer is the input layer of gesture feature fusion image with input of 32×32, layer1-3 is the convolution layer, layer4-5 is the full connection layer, and the last layer is the output layer. Some of these parameters, such as convolution kernel and bias, are generated randomly, and the updating of these parameters is based on the forward propagation and back propagation algorithm to train the network.

The essence of convolution filtering is to let the convolution kernel slide through the image matrix. The convolution kernel multiplies the elements of relative positions on the image, and then adds the results to get a result value. Finally, the convolution result is obtained by activation function. When the convolution kernel traverses the whole image, the feature extraction is finished and a new feature map is obtained. At the same time, the steps of convolution kernel sliding also have the following relations with the feature matrix finally obtained:

\[ a_{i,j} = f \left( \sum_{m=0}^{M} \sum_{n=0}^{N} w_{m,n} x_{i+m,j+n} + w_{b} \right) \]  \hspace{1cm} (7)

\[ f(x) = \max(0, x) \]  \hspace{1cm} (8)

\[ W_2 = \left( W_1 - F + 2P \right) / S + 1 \]  \hspace{1cm} (9)

\[ H_2 = \left( H_1 - F + 2P \right) / S + 1 \]  \hspace{1cm} (10)

Formula (7) is convolution calculation, Formula (8) is activation function, and Formula (9) and Formula (10) are convolution variation. Among them, \( x_{i,j} \) is the element in row \( i \) and column \( j \) of the image, \( w_{m,n} \) is the weight in row \( m \) and column \( n \) of the convolution kernel, \( w_b \) is the offset term of the convolution kernel, \( f \) is the activation function, that is, the relu function, \( W_2, H_2 \) are the width and height of Feature Map after convolution, \( W_1, H_1 \) are the width and height of image before convolution, \( F \) is the width of filter, \( P \) is the number of circles 0 filled around the original image, and \( S \) is the stride.

After the convolution filtering, the feature matrix of the down sampled image is used to reduce the calculation amount, and at the same time, the over fitting caused by too many features is avoided, which enhances the robustness of the network structure to the displacement. The specific convolution and subsampling calculations are as follows:
\[
\begin{align*}
    b &= P \begin{pmatrix}
        a_{i,j} & \ldots & a_{i+p,j} \\
        a_{i,j+1} & \ldots & a_{i+p,j+1} \\
        a_{i,j+q} & \ldots & a_{i+p,j+q}
    \end{pmatrix}
\end{align*}
\]

Where, \( a_{ij} \) is the element of row \( i \) and column \( j \) after convolution; \( P \) is the lower sampling function, generally Max Poling or Mean Poling, and Max Poling is used in this paper.

5. Experimental results and analysis
The algorithm uses self-built database, image acquisition and pre-processing and other multi-module communication based on the Robot Operating System (ROS) platform. In addition, it also combines Logitech camera to take indoor images and DJI Phantom P3P to take outdoor images. The collected pictures contain 31 kinds of gesture data sets designed above, each of which has 2000 pictures. Fig.6 is a partial result diagram selected from the dynamic gesture recognition video.

![Figure 6. Fingertip tracking and sorting renderings.](image)

In this algorithm, we want to realize the effective tracking and precise sorting of fingertips in the dynamic situation (from thumb to little thumb, the label starts from 0). From the above figure, when we extend five fingers, we can detect 0-4 from thumb to little thumb in turn; When the thumb is closed, the index finger is detected as 0 and pushed back in turn; when the thumb and the little thumb are closed, the other three fingers are detected as 0, 1, 2 and so on. According to the permutation and combination, the index of extended hand can be 1, 2, 3, 4, and 5. There are 31 permutations.

In this dynamic gesture recognition algorithm, the number of input samples is as high as 60000, and the number of output sample elements is 31. The sample size is relatively high. However, from the live dynamic video, it can be seen that the whole recognition process is smooth and accurate, and there is no jam or recognition error. It can be seen that the recognition rate and recognition speed are greatly improved while the number of samples is improved.

6. Conclusion
In this study, when Improved Canny operator is used for edge detection of hand image, the first step is to generate HSV space image converted from RGB space and extract the V-component image, and then change the denoising method from traditional Gaussian filter to bilateral filter that integrates spatial distance and similarity. The experimental results show that the improved canny algorithm can get the edge information of each image clearly, the edge location is relatively accurate, and the details are well processed. When K-means clustering algorithm is used to extract feature points from the edge of the hand image, the pixel points at the peak of histogram are used as the initial clustering center, and the number of peak points is used as the number of classification, which is not to randomly determine the number of classification and clustering center, which greatly reduces the amount of calculation and
makes the clustering results more ideal. The detection algorithm of convex hull and convex defect of contour points makes fingertip tracking more effective. Finally, more than 60000 pictures of 31 kinds of finger sorting graphs are input into the set convolution neural network for training and learning, which realizes the accuracy of fingertip sorting.

Acknowledgments

This work was financially funded by Tianjin graduate research innovation project.

References

[1] STERGIOPULOU E, PAPAMARKOS N. Hand gesture recognition using a neural network shape fitting technique [J]. Engineering Applications of Artificial Intelligence, 2009, 22(8): 1141–1158.

[2] MURTHY G R S, JADON R S. Hand gesture recognition using neural networks [C]// Advance computing conference. Patiala, India: IEEE, 2010: 134-138.

[3] LIU Yun, YIN Yanmin, ZHANG Shujun. Hand gesture recognition based on HU moments in interaction of virtual reality [C]// 4th international conference on intelligent human-machine systems and cybernetics. Nanchang, China: IEEE, 2012: 145-148.

[4] Parthasarathy, G Ramanathan, K Anitha, Y Justindhas. Predicting Source and Age of Brain Tumor Using Canny Edge Detection Algorithm and Threshold Technique [J]. Asian Pacific journal of cancer prevention: APJCP, 2019, 20(5).

[5] Fuente-Tomas Lorena de la, Arranz Belen, Safont Gemma, Sierra Pilar. Classification of patients with bipolar disorder using k-means clustering [J]. PloS one, 2019, 14(1).

[6] LINHI, HSU M H, CHEN W K. Human hand gesture recognition using a convolution neural network [C]// 2014 IEEE International Conference on Automation Science and Engineering (CASE). USA: IEEE, 2014: 1038-1043.

[7] Yanlu Xu, Yan Ma. Image edge detection algorithm based on improved Canny operator [J]. Fujian Computer, 2019, 35(05):5-7.

[8] Yubo Li. Research on the selection of initial clustering center of K-means algorithm [D]. Lanzhou Jiaotong University, 2012.

[9] Navjot Singh, Rinki Arya, R.K. Agrawal. A convex hull approach in conjunction with Gaussian mixture model for salient object detection [J]. Digital Signal Processing, 2016, 55.

[10] Abbas Al-Refaie, Ali Diabat. Optimizing convexity defect in a tile industry using fuzzy goal programming [J]. Measurement, 2013, 46(8).

[11] Youssef, M.M, Asari, K.V, Tompkins, R.C. Hull convexity defects features for human activity recognition [J]. IEEE(AIPR), 2010.

[12] Liang Chang, Xiaoming Deng, Mingquan Zhou. Convolution neural network in image understanding [J]. Journal of automation, 2016, 42(09): 1300-1312.