Gesture Recognition with Multiple Spatial Feature Fusion

Qinmeng[1, a],Heguoping[2]

[1.]Linyi university,Shandong,linyi,276000

[2.]shandong academy of sciences,Shandong,jinan,250014

[a]email:qinmeng123sd@163.com

Keywords: Multiple Space, Gesture Recognition, Depth Image, Feature Filtering

Abstract. Hand gesture recognition is an important topic in the field of computer vision, it has a wide range of applications, such as interactive games and sign language recognition, etc.. With the launch of the depth sensor, the task of hand gesture recognition becomes more simple. In recent years, there are a number of methods to try to extract features in the depth image, which is to be an effective expression of some kind of gesture. However, due to the inherent flexibility and complexity of the gesture, the recognition performance of the existing algorithms on large data sets is still not satisfactory. Local shape is presented in this paper a novel method based on multi spatial feature fusion to recognize static hand gesture depth image, namely on the 3D point cloud which were the local principal component analysis. It extract the local gradient information and local points cloud depth distribution, the effective information coding the gesture, we put local features which are the concatenation of the gesture image features and the classification results of the random forest classifier of features are filtered to remove the result of classification which did not influence the characteristics. We adopt filtered features to train the random forest again to recognize gestures. Compared with the existing algorithms, this method can effectively improve the recognition rate of the two large gesture data sets.

Introduction

In this work, we extracted multiple spatial characteristics in the gesture of the depth of the image , including local spaces of 3D point cloud by principal component analysis (PCA) feature vectors and between pairs of adjacent to the point cloud point calculation depth difference of gradient information and local spatial gesture image depth distribution. These features are projected onto the corresponding space after quantization, that is, encoding local shape of the gesture. After that, we will feature these features through the random forest (RDF) classification results for feature filtering. In the end, we train the random forest to recognize the gesture. Compared with the previous methods, our method has more distinguishing features. Experiments were conducted on Hand Digits Data set [10] NTU and Spelling Data set Finger two data sets and compared with the current popular gesture recognition method. The results show that the method of multi spatial feature fusion has achieved a better recognition effect.

Multiple Spatial Features

Given a depth image of the gesture d = I (x, y), where x and y are for the pixel position coordinates of the image, d for the corresponding depth values, ranging from 0 to 255, all of the depth value is equal to 255 of the pixels which are background pixels. The image has been standardized to the center of the gesture and the main direction of the gesture and it only contains the gesture after the segmentation extracted part. The reprocessing part will be introduced in section fifth.

Set the size of the image I $M \times N$,We divide it into $n_b$image patches△I. Among them. $n_b = n_x \times n_y$the number $n_x n_y$of image blocks in X direction and Y direction, respectively. $\Delta x = M / n_x$, $\Delta y = N / n_y$A delta x, delta y, then each image block size for delta x and delta y. We're going to extract $\Delta x \times \Delta y$ each△I from three different based on the characteristics of spatial

information and finally to all the features of the image blocks combined into one long vector as the characteristics of the overall image, as shown in Figure 1.

Multi Layer Spatial Principal Component Histogram

In order to describe the 3D shape of the gesture, we first transform the depth image into the point cloud in the three-dimensional space. For the sake of convenience, we set up z=255-d. Then all of the foreground pixels in a depth image can be represented by a 3D point cloud Ω:

$$\Omega = \{(x, y, z) | z \neq 0\}, \quad (1)$$

For any one point P of the point cloud $p \in \Omega$, we define its local space as Ω_p:

$$\Omega_p = \{q | \, ||q - p|| \leq r\}, \quad (2)$$

p和q转换为$(x, y, \lambda z)$, λ is the depth of the value and the plane coordinates of the conversion of parameters, r is the distance parameter, they need to debug in the experiment.

We consider the Ω_p point in the shape of a gesture in the p neighborhood that has the ability to describe, so we performed Ω_p a principal component analysis (PCA).

The covariancen_p matrix $\Omega_p$can be expressed as the number of internal points$C$.

$$\mu = \frac{1}{n_p} \sum_{q \in \Omega_p} q \quad , \quad (3)$$

in which

$$\mu = \frac{1}{n_p} \sum_{q \in \Omega_p} q \quad , \quad (4)$$

The characteristics C of the decomposition, it comes to

$$CV = EV \quad , \quad (5)$$

For diagonal matrix, $\lambda_1 \geq \lambda_2 \geq \lambda_3$it contains three characteristic values and three characteristic values to deal with the feature vector$[v_1 \, v_2 \, v_3]$, which $v_1$indicates the direction of the maximum variance and $v_3$the 3D point cloud in the surface normal vector [18]. They all contain the local shape information of the 3D point cloud. In order to encoding, we define two types of projection, see figure 2. In order to avoid the 180 degree ambiguity in feature vector. We require that the component of the feature vector$z$ must be non negative on the axis. After the projection:

Positive twenty plane projection:

$$\left(\frac{\pm 1}{\gamma}, \frac{\pm 1}{\gamma}, \frac{\pm 1}{\gamma}\right), \left(0, \frac{\pm \varphi^{-1}}{\gamma}, \frac{\pm \varphi}{\gamma}\right)$$

$$\left(\frac{\pm \varphi^{-1}}{\gamma}, \frac{\pm \varphi}{\gamma}, 0\right), \left(\frac{\pm \varphi}{\gamma}, 0, \frac{\pm \varphi^{-1}}{\gamma}\right)$$

In order to encoding feature vector, we need to quantify the three-dimensional space. In the regular polyhedron, the twenty faces are the most, the loss of information at least. We projected the feature vector to the center of each of the twenty faces. It is known that the vector of the center of the center to the origin of the twenty faces can be expressed as

In order to encoding feature vector, we need to quantify the three-dimensional space. In the regular polyhedron, the twenty faces are the most, the loss of information at least. We projected the feature vector to the center of each of the twenty faces. It is known that the vector of the center of the center to the origin of the twenty faces can be expressed as

$$P_i = B_{20}{}^T v_i \in R^{2C} \quad , \quad (6)$$

We changed all of the projection components of less than 0 to 0, obtaining the value P_i',then normalized and scaled according to the corresponding characteristic value.

$$h_p^i = \lambda_1 P_i' / ||P_i'|| \quad , \quad (7)$$

After that, all the feature vectors are connected to a long vector and the feature of P is obtained.

$$H_P = [h_p^1, \, h_p^2, h_p^3] \in R^{61} \quad , \quad (8)$$

Three plane projection:

$(\pm 1,0),(0,\pm 1),(\pm\sqrt{2}/2,\pm\sqrt{2}/2)$

Another way of projection is to the feature vector in the projection of a plane XY, YZ, XZ, which consists of the coordinate axes. We divide the plane of 360 degree into 8 blocks according to the degree of 45 and correspond to 8 vectors:

Then for its matrix form $B_8 \in R^{2\times 8}$, we use v_i the corresponding X axis and the Y axis component to form a new vector, then the vector in the XY plane projection v_i^{xy}

$$P_i^{xy} = B_8{}^T v_i^{xy} \in R^8, \tag{9}$$

For the YZ and XZ planes, the feature vectors are not negative in the Z direction as before, so we take B_8 the second component as the non negative five vectors to form a set of basis B_5. It is

$$P_i^{yz} = B_5{}^T v_i^{yz} \in R^5 , \tag{10}$$

$$P_i^{xz} = B_5{}^T v_i^{xz} \in R^5 , \tag{11}$$

The projection connection of three planes is obtained by

$$P_i = [P_i^{xy}, P_i^{yz}, P_i^{xz}] \in R^{18} , \tag{12}$$

After the same standardization and the characteristic value, we get $H_P \in R^{54}$.

After according to the two projection methods H_F, we calculate the point p where the image block principal component histogram ΔI

$$H_{\Delta I} = \sum_{p\in\Delta I} H_F , \tag{13}$$

And in order to carry out standardization $H_{\Delta I} \leftarrow H_{\Delta I}/||H_{\Delta I}||$, We connect the principal component histogram of all image blocks into a large vector as the principal component histogram of the whole image.

$$H_{pca} = [H_{\Delta 1}, H_{\Delta 2}, \ldots, H_{\Delta n_b}] \in R^{20n_b}, \tag{14}$$

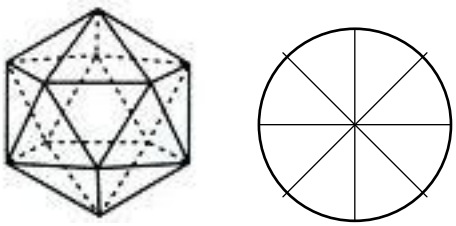

Fig 1. Twenty Normal Plane and XY Plane

Experiment

Data set

In this paper, Hand Digits Data set NTU [10] and Spelling Data set [13] Finger on the two hand gesture data sets were carried out. Two data sets are collected from the depth of the Microsoft Kinect hand gesture images, Spelling Data set Finger also contains the color of the gesture image, but this is not used in this paper.

Hand Digits Data set NTU contains 1000 pictures, containing 10 different types of gestures (from 0 to 9). Image collected from 10 people, that is, each person to collect 10 pictures of each. The original image contains a person and background, after the segmentation of the gesture, the data set contains the gesture as shown in figure 3.

Fig 2. Hand Digits Data Set [10] NTU

Spelling Data set Finger containing 60000 images have been segmented, including the American sign language set (ASL) in the 24 letters (from a to Z, remove the J and Z two dynamic gestures), collected from 5 people. Compared to Hand Digits Data set NTU, the data set of the data between the smaller difference, the difference within the class is greater, making the classification more difficult. The data set of gestures is shown in Figure 4.

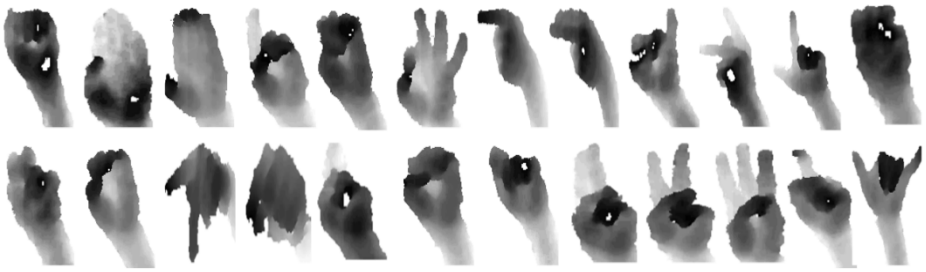

Fig 3. Spelling Data Set [13] Finger

Experimental content

For the original depth image, we first make a pre-processing step, including gesture segmentation, image scale standardization and the direction of the main direction of the standardization. In this paper, we adopt the method of limiting depth threshold to segment: that is, the hand is the nearest object from the depth camera and select the pixels in a certain range of greater than the minimum depth value. Then we map the depth value to 255 to 0 of the gray space and generate the gesture image. For the NTU data set, we also make a more accurate gesture segmentation by calculating the range of the palm. Because of the different size of the image, we have standardized the image size 120×100. After the experiment, we have chosen the best image size. In order to reduce the differences between different images with a gesture (mainly in the plane differential rotation), we conducted a gesture for the main direction of the standardization, namely by PCA to find the foreground pixels of principal components, and rotate the image so that its height direction and image parallel.

In the experiment, we selected the image block size of $\Delta I\,10 \times 1($, so the number of the image blocks is 120. For the principal component histogram, the positive twenty plane projection has 7200 dimensions and the three plane projection has 6480 dimensions. The gradient direction histogram has 960 dimensions and the depth distribution histogram has 1200 dimensions.

After feature filtering, we choose 2000 important features of high importance.

Conclusion

In this paper, we propose a new method of gesture recognition based on multi spatial feature fusion. These spatial features describe the shape and distribution of gestures in the local space and we make the feature filtering to retain the features of the discriminant information so as to reduce the computational overhead. We performed experiments on two large gesture data sets and compared with the popular methods, our method effectively improves the recognition performance. In the future, we will consider how to improve the features to make it more discriminative, such as the use of convolution neural network and other methods to automatically learn the characteristics of gestures. And how to reduce the computational overhead of feature extraction and find the low

dimensional expression of a certain gesture, finally reducing the dimension and so on.

Reference

[1] He, J., Geng, Y., Wan, Y., Li, S., and Pahlavan, K. (2013). A cyber physical test-bed for virtualization of RF access environment for body sensor network. Sensors Journal, IEEE, 13(10), 3826-3836.

[2] Lv Z, Tek A, Da Silva F, et al. Game on, science-how video game technology may help biologists tackle visualization challenges[J]. PloS one, 2013, 8(3): 57990.

[3] Su T, Wang W, Lv Z, et al. Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve[J]. Computers & Graphics, 2016, 54: 65-74.

[4] Jinyu Hu, Zhiwei Gao and Weisen Pan. Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation[J]. Journal of Applied Mathematics, 2013 (2013).

[5] Shuang Zhou, Liang Mi, Hao Chen, Yishuang Geng, Building detection in Digital surface model, 2013 IEEE International Conference on Imaging Systems and Techniques (IST), Oct. 2012