International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.10 (2016), pp.151-158
http://dx.doi.org/10.14257/ijmue.2016.11.10.14

Corrigendum Algorithm for Gesture Recognition Based on Multiple Information Fusion and Kinect

Jinghui Wang[1], Wenqun Cao[2] and Mingzhi Niu[3]

[1]*Beijing Forestry University, Beijing, China*
[2]*Beijing Forestry University, Beijing, China*
[3]*Beijing Enfeisi technology co. ltd., Beijing, China*
E-mail: 18511371833@163.com

Abstract

Gesture recognition is an important and challenging task in the field of computer vision. Starting from the 3D shape of coding gestures, it puts forward a new kind of gesture recognition framework based on depth image. It extracts the space characteristics of a variety of 3D point cloud based on Kinect, including local principal components analysis on point cloud to get the histogram of main component, gradient direction histogram based on local depth difference and depth distribution histogram of local point cloud. Principal component histogram and gradient direction histogram effectively coding the local shape of gestures, depth distribution histogram compensates the loss of the shaping descriptor information. Through preliminary training of random forest classifier to filter the characteristics, and characteristics with less influence on classification results are removed, thus the computational costs are reduced. The filtered characteristics are used for training of random forest classifier again to classify gestures. Experiment is carried on two large-scale gesture data sets, for more difficult ASL dataset, the proposed method has improved the recognition rate of 3.6% then the best previous algorithm.

Keywords: *Gesture recognition; Kinect sensor; Depth image; Multiple spatial characteristics; Characteristics filtering*

1. Introduction

Gestures are a natural and intuitive way of human communication, with the popularity of computing technology, gesture recognition based on computer vision has become an important research subject in the field of human-computer interaction. On the other hand, the commercial gesture recognition system, such as Leap Motion [1] *etc.* has become the traditional an alternative way of human-computer interaction in recent years. However, the development of this field is very rapid, but gesture recognition is still a very difficult problem. This is caused by the inherent flexibility and complexity of gesture itself. In recent years, with the emerging depth sensor such as Kinect [2], gesture recognition task has become more and more convenient. First, due to the depth image is not sensitive to light condition, gesture segmentation method based on depth threshold [3] is more simple and has more robustness than traditional gesture segmentation based on skin color; second, compared with the traditional color image the depth image provides additional distance information, which converts the gesture recognition from 2D image recognition problem into a 3D object recognition problem. Third, the depth image does not contain the information of color and material of the object, thus it expresses the geometric shape of the object more purely, so it is more convenient for the researchers to extract the characteristic based on shape.

ISSN: 1975-0080 IJMUE

2. Multiple Spatial Characteristics

2.1. Gestures Character Description

A given depth image containing gestures $d = I(x,y)$, where x and y is the position coordinates of the pixels in the image, d is the corresponding depth, with the range from 0 to 255, all depth values equal to 255 pixels are regarded as the background pixels.This image has been made standardization on the center of the gesture and the main direction of gesture, and only contains the extracted gesture part after gesture segmentation.Preprocessing part will be introduced in Section 5 experiment.

Set the size of the image as $M \times N$, and evenly divided into n_b image blocks of ΔI, and $n_b = n_x \times n_y$. n_x and n_y denote the number of image block at x direction and y direction respectively. Set $\Delta x = M / n_x$, $\Delta y = N / n_y$, so the size of each image block is $\Delta x \times \Delta y$. This paper extracts three different characteristics based on spatial information, finally put all the characteristics of the image blocks combined into a long vector to be the characteristics of the whole image, as shown in Figure 1.

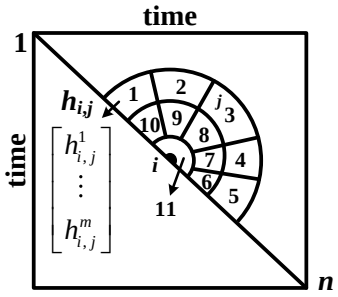

Figure 1. Local Characteristic Description

2.2. Principal Component Histogram

Principal component histogram was first put forward by Hossein *et al.*, used for gesture recognition.In order to describe three-dimensional shape of gestures, first convert depth image into point cloud in 3D space.For convenience, set $z=255-d$.So 3D point cloud Ω composed by all foreground pixels of this depth image can be represented as:

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

For any point $p \in \Omega$ in this point cloud, we define its local space as Ω_p, and satisfies

$$\Omega_p = \{q \mid ||q - p|| \leq r\} \tag{2}$$

Where, p and q transform to $(x, y, \lambda z)$, λ converted to the proportion of conversion parameters of depth and plane coordinate, r is the distance parameter, they need to be debugged in the experiment.

Points in Ω_p has certain descriptive power on gestures surrounding, so it conducts principal component analysis on Ω_p.

Set n_p as the number of points in Ω_p, then covariance matrix C of point in Ω_p can be expressed as

$$C = \frac{1}{n_p} \sum_{q \in \Omega_p} (q - \mu)(q - \mu)^T \tag{3}$$

where

$$\mu = \frac{1}{n_p} \sum_{q \in \Omega_p} q \tag{4}$$

Make characteristic decomposition on C, then we have

$$CV = EV \tag{5}$$

E is the diagonal matrix, includes three characteristic values $\lambda_1 \geq \lambda_2 \geq \lambda_3$. V contains characteristic vectors $[v_1\ v_2\ v_3]$ of three characteristic values. v_1 indicates the direction of the maximum variance, v_3 indicates the normal vector of the surface of 3D point cloud. They contain the local shape information of 3D point cloud. In order to carries on the quantification and coding on it, this paper defined two projection methods, so as to avoid 180° ambiguity existed in characteristic vector, we regulate the component of characteristic vector on z axis must be nonnegative, otherwise each dimension will be inverted before projection:

2.3. Depth Distribution Histogram

Principal component histogram and gradient direction histogram are shape descriptor, for they do not have robustness on depth value changes of the same shape, and for the image block ΔI, its depth value distribution also contains local information of the gesture. Depth distribution histogram solved the problem that shape descriptor is sensitive to the depth changes, and added the depth distribution information of gestures.

Select minimum depth of all foreground pixels $d_{min} > 0$ and maximum depth d_{max}. And $d_{max} - d_{min}$ is divided evenly into N_d segments, and the size of each segment is

$$\Delta d = (d_{max} - d_{min}) / N_d \tag{6}$$

For all foreground pixels in image block ΔI, we determine its segment num according to its depth value, and construct depth distribution histogram H_{dd} based on this.

$$num = \left\lceil \frac{I(x,y)}{\Delta d} \right\rceil \tag{7}$$

$$H_i(num) = H_i(num) + 1, \forall (x, y) \in \Delta i \tag{8}$$

Finally, the characteristic of depth distribution histogram of the whole image is

$$H_{dd} = [H_1, H_2, \ldots, H_{n_b}] \in R^{N_d n_b} \tag{9}$$

3. Kinect Data Gesture Recognition Steps

3.1. Data Acquisition

After extraction of various space characteristics, this paper fuses them into a long vector as global characteristics of the whole image. Due to characteristics included HOPC and HOG dense operator can lead to character dimension too high, and "dimension disaster" caused expensive computational cost, so it needs dimension reduction by characteristics filtering. This paper adopts the method of preliminary training of random forest to measure the importance of characteristics, so as to select a discriminant characteristic.

Because of the large noise in Kinect depth data, hands can be positioned as the most reliable marks in such 3D gesture model, based on 2D+3D algorithm [10], it can detect fingertips of 3D gesture under different expressions and gestures. However, in this case, the 3D data is high resolution, this paper assumes that the tip position has been detected approximately. Because the fingertip only need face clipping and rough alignment, therefore, as long as the detected point is closed enough to the real position, system can work normally.

3.2. Model Building

Given the tip of the finger position, 3 D gesture cutting can be done easily, this algorithm using a sphere with a radius of 8 cm to cut gestures, first of all to finger point cloud into the origin, then remove those points away from the origin more than 8 cm, thus can get only face 6 D point cloud surface area.

Iterative closest point (ICP) algorithm is based on a precise alignment technique, and its computational cost is very large, because different objects have different face shapes, reference gesture model must be the reliable expression of general 3D gesture, and can not be constructed with high noise level of the Kinect 3D data. The proposed algorithm, therefore, through alignment scanning, and resampling on uniform grid, then take their average to build reference gestures, the reference gesture shall be 64 points between the center of two eyes, and points on the ligature from lips center to the eyes are also 64, the complete gestures have 128×128 points, reference gesture model used in the experiment as shown in Figure 2, all gestures including training data and query gestures use six ICP iteration to get the reference gesture.

Figure 2. Reference Gestures

After gesture correction, through X value of the original point cloud replaced with opposite (-X) to create a mirror point cloud. However, not all mirror points are useful, because the purpose of this study was to fill the missing data. In ideal condition, the positive gestures don't need to add points, and all points should be reflected in a profile view. For this, each mirror point, this paper calculates Euclidean distance of the nearest point at the origin cloud (XY value only), if the distance is less than threshold δ, so mirror points are removed, when, and only when there is no neighborhood points on a location to add mirror points. One shall note that, do not use Z when calculating the distance, because the difference of Z is usually caused by palm symmetry instead of missing data. Then to merge the rest of the mirror points and origin cloud.

3.3. Smoothing

Threshold δ can be spatial resolution based on sensor or point cloud itself, this value can be user defined. Depending on the initial sample density, too high δ value will produce a noise surface, while too low value is useless for symmetrical filling. Experiment has shown that different δ value taken from 1 to 5 mm had less influence on performance, when δ = 2mm a good balance can be achieved.

Resampling has three main objectives: ① it can remove the noise surface generated by Kinect sensor smoothly, and symmetric filling; ② it may fill loopholes still existed after symmetric filling; ③ it reduces the influence of gesture alignment error on 2D grid caused by ICP registration. For this purpose, the algorithm will fit smooth surface to point cloud (XYZ), the algorithm will use similar instead of interpolation to fit curved surface to point, using a smoothing factor (or fitting) to perform curved surface fitting, surface sudden bending is not allowed, so as to alleviate the influences of noise and outliers. For every gesture, 128×128 points were uniform resampling, from the minimum X and Y to the maximum X and Y, the advantage of resampling from the minimum to the maximum is that, it can align the faces on 2D grid. There is no smooth texture, because it's not noisy, smooth will only make it become blurred. After resampling, X and Y grid will be discarded, and depth and four 128 128 matrixes can be obtained, in order for further processing, they continue to the next sampling as 32×32 size.

4. Experiment

4.1. Data Set

This paper carried on experiments on two gesture data sets of NTU data set [10] and ASL dataset [13].Two data sets were collected from the depth gesture image of Microsoft Kinect, ASL dataset contained color gesture image, but this paper didn't use it.

NTU data set contained 1000 images, including 10 different types of gestures (from 0 to 9).Image acquisition from 10 individuals, that was, collected 10 images of each gesture on each person.The original image contained person and background, through gestures segmentation, gestures contained in the data set as shown in Figure 3.

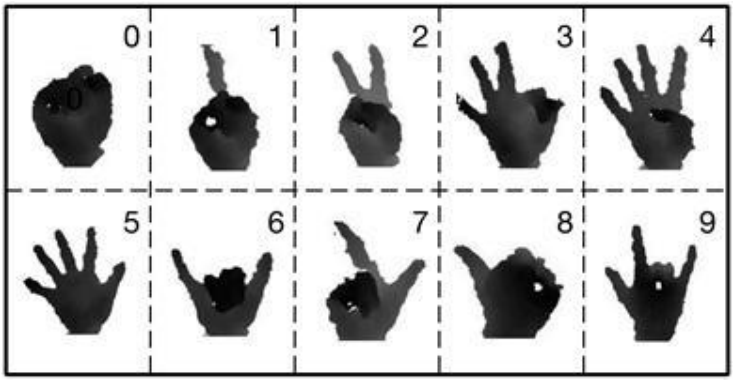

Figure 3. NTU Hand Digits Dataset [10]

ASL data set contained 60000 pieces of divided gesture images, contained 24 letters of American sign language (ASL) (from a to z, remove two dynamic hand gestures of j and z), collected from five people. Compared with NTU data set, the difference between gestures in this data set was smaller, and intra-class difference was bigger, which made the classification more difficult. The gestures of the data set as shown in Figure 3.

4.2. Parameter Setting

In order to compare with the current algorithm, this paper adopted the same cross validation strategy of literature [12] and [13], namely, independence between objects and co-dependent between objects. For the samples collected from N individuals, independence between objects indicated that, with N - 1 individuals as the training sample set, 1 individual as a test set, repeated for N times to make the training set and test set covering all situations, and then take the average accuracy; co-dependent between objects indicated that, all N individuals randomly and evenly divided into two parts, one part as the training set, the other part as the test set. Also take average accuracy after repeated N times.

For the original depth image, the required preprocessing steps, including gesture segmentation, image scale standardization, the main direction standardization of gestures. This paper adopted the method of limited depth threshold for gesture segmentation: hands were regarded as the object most close to the depth camera, and selected a certain depth range of pixels as point cloud of gestures, and mapped the depth value to 0 to 255 of gray space, to generate gesture image. For NTU data set, we also made more accurate gesture segmentation by calculating the palm range. Due to the image size was differ, the standardization of image size was required, after experiment, we selected the best image size to be pixels of 120 height, pixels of 100 width. The standardization of main direction of gestures can reduce inner class difference caused by the in-plane rotation, this paper set

the direction of the principal component of foreground pixels found by PCA as the main direction of gestures, and rotating the image make the y axis as the main direction.

In the experiment,the size of the selected image block was10×10, so the number of the image blocks was 120.For principal component histogram, positive icosahedron projection had 7200D, three plane projection had 6480D.Gradient direction histogram had 960D, depth distribution histogram had 1200D.

After the characteristics filtering, we selected 2000 characteristics with high importance.

4.3. Experimental Results

Through systematic contrast experiments, the recognition rate of proposed method and the current gesture recognition algorithm on the two data sets as shown in Table 1 and Table 2. As can be seen from the table that, on the two data sets, the proposed method has obtained better recognition rate than that of current method.

Table 1. Recognition Rate of Each Method Used on NTU Data Set

method	independent between objects	co-dependent between objects
Ren[10]	0.939	N/A
HOG	0.931	0.964
H3DF[12]	0.955	0.992
Our method 1	0.972	0.994
Our method 2	0.963	0.992

The proposed method 1 indicates positive icosahedron projection, the proposed method 2 refers to the three-plane projection. In addition, the experiments of two data sets, the recognition rate of two projection methods are similar, the effect of positive icosahedron projection is better than that of three-plane projection.

Table 2. Recognition Rate of Each Method Used on ASL Data Set

method	independent between objects	co-dependent between objects
Ren method	N/A	N/A
Bowden method	0.480	N/A
HOG	0.634	0.970
H3DF method	0.713	0.979
This method No.1	0.757	0.977
This method No.2	0.759	0.972

This paper regarded the gestures as a 3D object to extract characteristics, without depending on the particular perspective and gestures, so there would not appear the condition of some gesture not supported. Method most close to the proposed method was [12], which also encoded the normal vector of 3D object surface.

Figure 4. Finger Spelling Dataset [13]

Table 3. Recognition Rate of Posture and Gesture Changes(%)

	dissymmetry			symmetry		
gesture	D	T	fusion	D	T	fusion
positive	100	100	100	100	100	100
rotate ±30°	49.5	98.1	93.6	88.3	99.8	99.4
rotate ±60°	14.9	80.4	55.1	87.0	97.4	98.2
rotate ±90°	1.0	39.4	14.4	74.0	83.7	84.6
Tilt ±60°	77.2	91.3	90.0	81.6	89.1	92.8
average	46.2	87.6	77.0	85.4	95.0	96.3

This paper effectively expressed the 3D shape through local principal components analysis, and integrated the characteristics of more identifying information, thus the classification accuracy was improved. Table 3 provided the confusion matrix on ASL data set, data from Figure 4, it reflected the percentage relationship of real category of sample and predicted category. It can be seen from the figure that, even though the proposed method improved the recognition rate, but for some gestures with similar appearance, such as the letter M and N, P and Q gestures, recognition error rate was still high.

5. Conclusion

This paper proposed a new gesture recognition method based on multiple spatial characteristics Kinect sensor data.The principal component histogram and gradient direction histogram described the shape of gestures in different scales, and depth distribution histogram embodied the depth distribution of gestures.On this basis, this paper calculated characteristics importance through the preliminary training of random forests and filtered characteristics.Experiment was carried on two large-scale gesture data sets, the results showed that, compared with the present popular gesture recognition algorithm, the proposed method can effectively improve the recognition effect. We will give consideration on how to extract characteristics of more discriminant information or using convolution neural network method to learn characteristics of gestures image automatically, so as to improve the recognition rate, and expand the existing method to dynamic gesture recognition based on depth video issues in the future.

References

[1] T. Su, W. Wang and Z. Lv, "Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve", Computers & Graphics, vol. 54, **(2016)**, pp. 65-74.
[2] W. Gu, Z. Lv and M. Hao, "Change detection method for remote sensing images based on an improved Markov random field", Multimedia Tools and Applications, **(2015)**, pp. 1-16.
[3] Z. Lv, A. Tek and F. D. Silva, "Game on, science-how video game technology may help biologists tackle visualization challenges", PloS one, vol. 8, no. 3, **(2013)**, pp. 57990.
[4] Z. Chen, W. Huang and Z. Lv, "Towards a face recognition method based on uncorrelated discriminant sparse preserving projection", Multimedia Tools and Applications, **(2015)**, pp. 1-15.
[5] Y. Lin, J. Yang and Z. Lv, "A Self-Assessment Stereo Capture Model Applicable to the Internet of Things", Sensors, vol. 15, no. 8, **(2015)**, pp. 20925-20944.
[6] J. Yang, S. He and Y. Lin, "Multimedia cloud transmission and storage system based on internet of things", Multimedia Tools and Applications, **(2015)**, pp. 1-16.
[7] Z. Lv, T. Yin and Y. Han, "WebVR——web virtual reality engine based on P2P network", Journal of Networks, vol. 6, no. 7, **(2011)**, pp. 990-998.
[8] J. Yang, S. He and Y. Lin, "Multimedia cloud transmission and storage system based on internet of things", Multimedia Tools and Applications, **(2015)**.
[9] C. Guo, X. Liu and M. Jin, "The research on optimization of auto supply chain network robust model under macroeconomic fluctuations", Chaos, Solutions & Fractals, **(2015)**.
[10] X. Li, Z. Lv and J. Hu, "XEarth: A 3D GIS Platform for managing massive city information", Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2015 IEEE International Conference on. IEEE, **(2015)**, pp. 1-6.
[11] J. Yang, B. Chen and J. Zhou, "A Low-Power and Portable Biomedical Device for Respiratory Monitoring with a Stable Power Source", Sensors, vol. 15, no. 8, pp. 19618-19632.
[12] G. Bao, L. Mi, Y. Geng and K. Pahlavan, "A computer vision based speed estimation technique for localiz ing the wireless capsule endoscope inside small intestine", 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)**.
[13] X. Song and Y. Geng, "Distributed community detection optimization algorithm for complex networks", Journal of Networks, vol. 9, no. 10, **(2014)**, pp. 2758-2765.
[14] D. Jiang, X. Ying and Y. Han, "Collaborative multi-hop routing in cognitive wireless networks", Wireless Personal Communications, **(2015)**, pp. 1-23.
[15] J. Hu and Z. Gao, "Modules identification in gene positive networks of hepatocellular carcinoma using Pearson agglomerative method and Pearson cohesion coupling modularity", Journal of Applied Mathematics, vol. 2012, **(2012)**.

Authors

Jinghui Wang, received her master's degree in Animation from the University of Technology Sydney in Australia. She is currently a lecturer in Beijing Forestry University and studying for a PhD in the Information College. Her research interest is mainly in the area of Computer Software, digital forestry technology, and digital media. She has published several research papers in scholarly journals in the above research areas and has participated in several books.

