International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.10, (2016), pp.89-96
http://dx.doi.org/10.14257/ijsip.2016.9.10.09

Robust Gesture Recognition with Kinect Data Acquisition

Jinghui Wang[1] and Mingzhi Niu[2]

[1]Beijing Forestry University, Beijing, China
[2]Beijing Enfeisi technology co. ltd., Beijing, China
18511371833@163.com

Abstract

To realize the gesture recognition of high precision ratio, the gesture recognition method of multi-model data fusion based on Kinect depth image is proposed, to implement the automatic splicing of models. First of all, the feature package model uses the speeded up robust feature (SURF) algorithm to replace the scale invariant feature transform (SIFT) algorithm to extract features, improve the real-time performance. Secondly, Hu moment is introduced to describe the global gesture features, further improving the recognition rate, the ray casting is used finally, and the obtained coordinate information is used to solve the rigid transformation between two point cloud models. Finally, the proposed data fusion method is verified through two experiments, the algorithm in this paper is better than the traditional support vector machine (SVM) method both in real time performance and recognition rate, and obtains better model splicing effect.

Keywords: *Gesture recognition; Kinect depth image; Model splicing; Data fusion*

1. Introduction

Different scale, rotation, perspective and other factors of the gesture image also make different images of the same gesture differ markedly, increasing the difficulty of recognition. In recent years, a lot of static gesture recognition methods based on depth image have been proposed. In accordance with different strategies of recognition process, these methods can be divided into two categories: top-down and bottom-up.

At present, Kinect has been widely used in the field of 3D scene modeling. This article will focus on the research of the method of depth data fusion, especially the data fusion based on voxel, which is applied to the Kinect depth image, to reconstruct 3D model, and analyze the deficiencies. However, only depending on the voxel to reconstruct fine 3D model can cause very serious memory resource consumption, and can't reconstruct the large 3D scene, therefore, this paper will expand Kinect reconstitution ability, put forward the multi-model splicing method, and reconstruct the large indoor 3D scene when ensuring the accuracy of reconstruction of 3D model. Kinect is the motion sensing device developed by Microsoft, and it belongs to a new type sensor called RIM (Range Imaging) camera [1], mainly used to capture the human body bone structure, to realize the idea of taking the body as the controller. Kinect consists of a transmission unit (pulse light, modulated light or structure light), the photosensitive sensor (CCD, CMOS, or APD), optical system and some drive circuits and computing units. Kinect device obtains the scene depth and image information synchronously, the scene texture mapping can be automatically accomplished, and the obtained depth data based on the Kinect device is structured point cloud data, so this paper adopts the data fusion method based on voxel to reconstruct 3D model.

ISSN: 2005-4254 IJSIP

2. Gesture Recognition Method

Each 3D model reconstructed by Kinect, should have a set of key frame data, and this paper use the data calculation to obtain the transformation relation between two 3D models. This process mainly includes feature extraction, feature merging, feature matching, calculation of transformation matrix, and fine matching.

2.1. Feature Extraction

Based on the existing RGB-D image, it is rather difficult to directly use point cloud data coordinate information to match the features of two images, by contrast, feature extraction and matching in the RGB image have better accuracy and robustness. Because the RGB image and point cloud data have been aligned, the extracted feature points in the RGB image mapped to point cloud data can be used as the feature of the point cloud data. The feature extraction based on SIFT operator has been widely used in image matching, because the traditional SIFT operator based on CPU can't satisfy the demand of the rapid processing, this paper uses the SIFT operator based on GPU, which can achieve the processing speed of 11 frames per second. It is noted that there is a large amount of invalid data in the point cloud data when SIFT feature is mapped to the point cloud data, therefore it is necessary to determine the validity of the data when mapping. Generally, invalid data appears in the following two cases, one is that the visual range of depth image is different from the visual range of color image, the existing area on the color images may be unable to receive the depth information on the depth image; The second is SIFT feature on the edge of an object, and we know that the point cloud data on the edge is not good. So these two reasons could lead to the point cloud data mapped to the invalid region. When SIFT feature corresponds to the invalid point on the point cloud data, SIFT point can be abandoned directly. In addition, solving the rigid transformation actually only needs 3-4 points, so too many SIFT features waste the time, 1000 SIFT features are saved at most on each image.

2.2. Feature Merging

A series of feature points are extracted from each image through feature extraction, but because there is the overlap between images, and repeat scanning may be conducted on the same area when scanning, there are many repeated features in a set of images, and data redundancy can be reduced through feature merging, improving the efficiency of the follow-up feature matching.

We know the SIFT feature descriptor has 128 directions, that is to say, the single SIFT feature can be regarded as the 128-dimensional vector. That SIFT features are used directly to conduct the feature merging consumes more resources (memory resources and computation resources), and the merging effect is not necessarily good. Through feature extraction and feature mapping, SIFT feature points in the point cloud data have the corresponding points (x, y, z), which we call 3D mapping point in order to facilitate the subsequent narrative. Three-dimensional mapping point is obtained from the voxel model, so if the SIFT features are the same, the coordinates of the corresponding three-dimensional mapping point should be very close under the unified coordinate system, as shown in Figure 1. Three-dimensional mapping point is used to carry out the feature merging based on this feature.

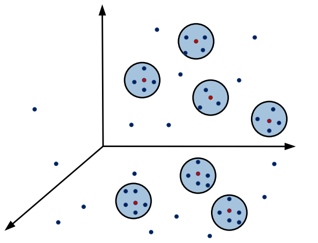

Figure 1. Feature Merging based on the Spatial Relationship

The process of feature merging is the process of proximal point clustering. What is the nearest neighbor in European space and mathematical description is given here: given a multidimensional space RK, a vector in the PK is a sample point, and the finite set of these sample points is called the sample set; given the sample set E , and a sample point $s^{'}$, the nearest neighbor of $s^{'}$ is arbitrary sample point $s \in E$ meets $Nearest\left(E,s^{'},s\right)$. Nearest is defined as follows:

$$Nearest\left(E,s^{'},s\right) \Leftrightarrow \forall s^{''} \in E\left|s^{'}-s\right| \le \left|s^{'}-s^{''}\right| \tag{1}$$

The distance metric in the above formula is the Euclidean distance, that is,

$$\left|s^{'}-s\right| = \sqrt{\sum_{i=1}^{k}(s_i - s_i^{'})^2} \tag{2}$$

Thereinto, s_i is the i th dimension of vector s . As can be seen from the above formula, search is a linear process, but because of the large amount of data, the speed of search point by point is certainly slower, K-D tree and Octree can be used in the test to partition the sample set, and then proximity search is conducted. This thesis uses the mature code library FLANN, which implements the nearest neighbor search based on K-D tree, and has been widely used in many software.

2.3. Feature Matching and Transformation Matrix

After the SIFT feature vector of two images is generated, the Euclidean distance of feature vector of key points is adopted in the next step as the similarity judgement measures of the key points in two images. Take the key point in the image 1, and find out the first two key points in the image 2 with the nearest Euclidean distance to the key point in the image 1. If the nearest distance divided by the adjacent distance is less than a certain ratio threshold in two key points, a pair of matching points is received. If the ratio threshold is reduced, the number of SIFT matching points will decease but will be more stable. In addition, because SIFT features of all photos are used when matching, although some SIFT features have been merged, the SIFT features in two groups of photos are still very rich, the realized SIFT matching algorithm by GPU improves the efficiency.

However, there is still no matching in the SIFT matching, so RANSAC [17] method is usually used to solve out the appropriate model parameters through "inlier" data. Because the model parameter gets "outlier" not suitable for this model, that is the abnormal point.

2.4. Fine Matching and Data Splicing

The more reliable transformation matrix can be obtained by RANSAC algorithm, but using transformation matrix can only implement the rough matching between the two sets of point cloud data, if using transformation matrix to match two sets of point cloud data can't get the perfect splicing effect, so ICP algorithm is used on the basis of the rough matching to realize the fine matching between the point cloud data. It should be noted that two sets of point cloud data involved in the fine matching is the complete point cloud data generated by voxel, and the complete point cloud data is more rich, and more reliable when ICP algorithm is used relative to the point cloud data corresponding to the single RGB image. A new matrix can be obtained by ICP matching, which is multiplied by the original rough registration matrix and the final finishing registration matrix is generated. The matrix is applied to two sets of point cloud data, and the precision splicing can be implemented.

2.5. Ray Casting

Because only the RGB image is stored during the scanning mentioned above, and the original depth image quality obtained by Kinect device is poorer, the depth image generated in the voxel model has high precision, and the ray casting technology needs to be used. The basis of ray casting projects a light beam from the center of projection, until the light reaches the surface of the nearest object which stops it from continuing to propagate. The ray casting algorithm is used to determine the XYZ coordinate information corresponding to each pixel on the RGB image here, and the obtained coordinate information is mainly used to solve the rigid transformation between two point cloud models.

The camera position parameter corresponding to the key frame is acquired while acquiring the RGB image in the acquisition process of key frames, and the XYZ coordinate corresponding to each pixel on the RGB image can be obtained from the voxel model with the ray casting algorithm by using the parameters. Assuming that the coordinate of the pixels on the RGB image is $\mathbf{\upsilon} = (r, c)$, the internal parameter of the camera is K, and the global position of the camera is $T_{g,k}$, then the ray corresponding to the pixel is $\mathbf{r} = T_{g,k}\mathbf{K}^{-1}\mathbf{\upsilon}$, the voxel is traversed along $\mathbf{r}$ in the voxel, until meeting the voxel recorded at $D(x) = 0$, and the returned XYZ coordinate here is the three-dimensional coordinates corresponding to the pixel.

3. Experiment Result and Analysis

3.1. Experiment Preparation

The gestures sample library is established first before the experiment begins, and this paper collects the samples of six gestures as shown in Figure 2, respectively marked as "fist", "good" and "ok", "number 6", "pointing to" and "praise". Each gesture collects 200 images, according to different light conditions, different hands, different sizes and different rotation angles. 100 images are used for training, to make the support vector machine (SVM) classification model. These images take white wall as the background, and there are only gestures in the images, so that we can ensure the extracted features come from the gesture area. Other 100 images are used for testing, and these images have no background restrictions, including simple background, complex background of approximate skin color object. There are some images from Sebastien Marcel [13] gestures library, the hand area needs to be segmented on the basis of facial elimination, skin color detection and hand shape contour comparison algorithm [2] before testing.

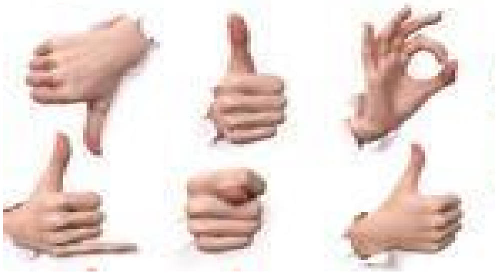

Figure 2. Sample Images of Six Gestures

First the gesture image is described based on the fusion feature algorithm proposed in this paper in the training phase, and then different labels I are assigned to each gesture, such as "fist" gesture corresponding to I = 1, "palm" corresponding to I = 2, *etc.*, and then <Label I, F vector space> is input to the multi-classification SVM classifier for training. In this paper, multi-classification SVM uses LibSVM open source package [14], and kernel function chooses the radial basis function. Because the size of the visual dictionary and the weight of fusion feature will affect the recognition rate at training, their values need to be obtained through the experimental experience. Finally the trained SVM classification model is saved in the xml file, convenient for direct calling in the testing phase.

The computer used in the experiment is Intel Core i3 CPU, 380 M, 2.53 GHz, 2 GB memory, collecting images uses ordinary webcam (Lenovo EasyCamera), and the resolution ratio is 640×480. Software environment is Visual Studio 2010, open source Visual package Opencv 2.4.3, Boost Filesystem Library Version 3 supporting file system operation and MATLAB 2010b simulation platform.

3.2. Experiment Result and Analysis

Experiment 1 the choice of the number of clustering centers. This experiment chooses 100 training images and 100 test images from the sample library for each gesture. We find through the extraction experiment of SURF feature points that "palm" gesture area is the largest, containing 179 feature points, so k value should be greater than 179. k value starts from 300 in the experiment, and increment is 100 each time, until 1000, and the fusion feature weights are 0.5. 6 gestures in the sample library are trained and recognized respectively, and recognition accuracy of each gesture under the k value is counted, and finally the average recognition rate under different k values is calculated. As can be seen from table 1, when the k value is 600, the average recognition accuracy is the highest.

Table 1. Gesture Recognition Accuracy Under Different k Values

k	Gesture						
	Gesture1	Gesture2	Gesture3	Gesture4	Gesture5	Gesture6	Average recognition rate(%)
300	93	94	93	93	92	91	92.67
400	95	92	89	90	92	91	91.50
500	97	94	95	92	93	95	94.33
600	98	96	93	97	94	96	95.67
700	98	95	94	96	92	96	95.17
800	96	94	93	91	95	94	93.83
900	94	92	94	88	91	93	92.00
1000	93	94	92	93	89	93	92.33

Experiment 2 the choice of the fusion feature weight. The above experiment obtains the algorithm recognition ratio when fusion weight is 0.5 respectively, and does not take into account the influence degree of two features on the recognition result. This experiment chooses the optimal number of clustering center in experiment 1 on the basis of the above experiment, namely $k = 600$, and tests the recognition performance of the algorithm in this paper under the change of Hu moment feature and BoF - SURF feature fusion weights w_1, w_2. Thereinto, $w_1 + w_2 = 1$, when $w_1 = 0$, it means only BoF - SURF feature recognition algorithm can be used, and when $w_1 = 1$, it corresponds to the recognition rate of Hu moment feature. As can be seen from the experiment result of Figure 3, when the weight of Hu moment is $w_1 = 0.4$, and the weight of BoF-SURF feature is $w_2 = 0.6$, the best recognition performance obtained by the algorithm in this paper is 96.33%.

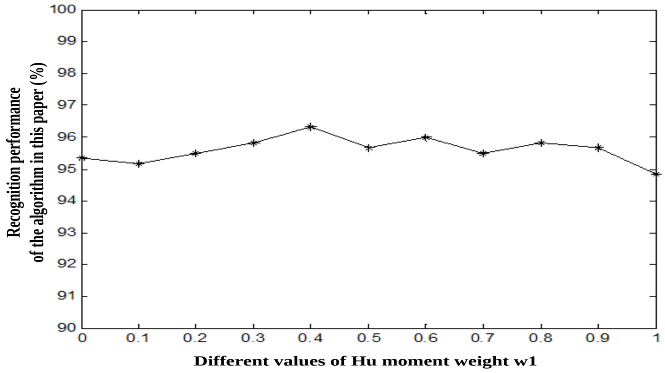

Figure 3. Algorithm Performance in this Paper at Different Fusion Weight Values

Experiment 3 comparison of the algorithm in this paper with other algorithms. This experiment compares the average recognition rate and the average recognition time of single image of BoF - SURF + SVM, BoF SIFT + SVM, Hu + the SVM and the algorithm in this paper under the condition of changing the number of training samples. The calculated time includes feature extraction, quantification and recognition process. As can be seen from the experiment results of Figure 4, and 2, that this algorithm can obtain higher recognition effect under the condition of less training samples, whose performance is better than that when two features are used alone, and whose recognition rate is higher than that of BoF-SIFT+SVM algorithm, with less time. Compared with the average recognition rate 88.63% of fusion feature in the literature [7], the fusion feature performance of this paper is better.

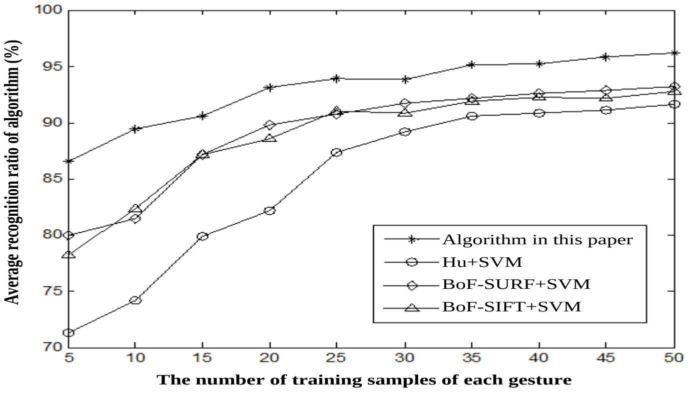

Figure 4. Performance Comparison of Algorithm in this Paper with Other Algorithms

Table 2. Average Recognition Time of Algorithm

Method	Average recognition time (s/picture)
Hu+SVM	0.042
BoF-SURF+SVM	0.761
Algorithm in this paper	0.838
BoF-SIFT+SVM	3.670

The recognition effect of some gestures by the algorithm in this paper is as shown in Figure 5. As can be seen from the figure that the algorithm in this paper can carry out the gesture recognition when there is face, similar skin color object and others in the background, which means the algorithm has good adaptability and robustness for the environment change.

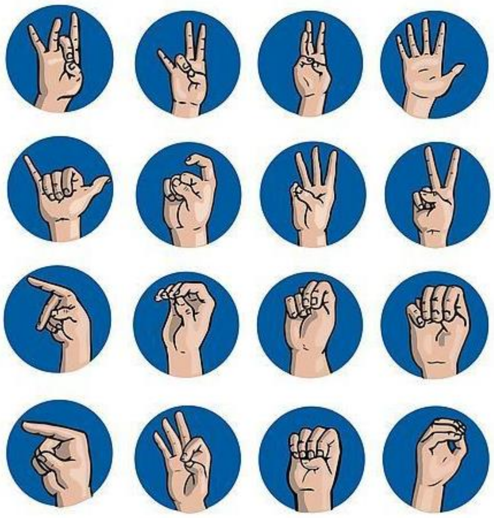

Figure 5. Gesture Recognition Effect Under Different Circumstances

4. Conclusion

This paper puts forward the robust gesture recognition method based on Kinect depth image data. First the gesture image feature is extracted through Kinect, then the improved K - Means clustering method is used to build the visual dictionary, and the gesture image is quantifies as BoF - SURF word frequency vector, and finally Hu moment vector and word frequency vector are input to SVM for training and classification by a certain weight fusion to describe gesture images. The experiment results show that the algorithm in this paper can achieve better recognition performance and higher real-time performance. What needs to be improved is: 1) Due to the use of different sample libraries and different experiment settings, the choice of the optimal number k of clustering center is also different, so you need to further study how to choose the optimal value k. 2) The fusion method of two features in this paper is relatively simple, then the relationship between two features will be studied in depth, to choose the better fusion method, and further improve the recognition rate.

References

[1] 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.

[2] K. Wang, X. Zhou and T. Li, "Optimizing load balancing and data-locality with data-aware scheduling", Big Data (Big Data), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 119-128.

[3] L. Zhang, B. He and J. Sun, "Double Image Multi-Encryption Algorithm Based on Fractional Chaotic Time Series", Journal of Computational and Theoretical Nanoscience, vol. 12, **(2015)**, pp. 1-7.

[4] T. Su, Z. Lv and S. Gao, "3d seabed: 3d modeling and visualization platform for the seabed", Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 1-6.

[5] Y. Geng, J. Chen, R. Fu, G. Bao and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine", IEEE transactions on mobile computing, vol. 1, no. 1, **(2015)** April, pp. 1-15.

[6] Z. Lv, A. Halawani and S. Feng, "Multimodal hand and foot gesture interaction for handheld devices", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1, **(2014)**, pp. 10.

[7] G. Liu, Y. Geng and K. Pahlavan, "Effects of calibration RFID tags on performance of inertial navigation in indoor environment", 2015 International Conference on Computing, Networking and Communications (ICNC), **(2015)** Febuary.

[8] J. He, Y. Geng, Y. Wan, S. Li and K. Pahlavan, "A cyber physical test-bed for virtualization of RF access environment for body sensor network", IEEE Sensor Journal, vol. 13, no. 10, **(2013)** October, pp. 3826-3836.

[9] W. Huang and Y. Geng, "Identification Method of Attack Path Based on Immune Intrusion Detection", Journal of Networks, vol. 9, no. 4, **(2014)** January, pp. 964-971.

[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. He, Y. Geng, F. Liu and C. Xu, "CC-KF: Enhanced TOA Performance in Multipath and NLOS Indoor Extreme Environment", IEEE Sensor Journal, vol. 14, no. 11, **(2014)** November, pp. 3766-3774.

[12] N. Lu, C. Lu, Z. Yang and Y. Geng, "Modeling Framework for Mining Lifecycle Management", Journal of Networks, vol. 9, no. 3, **(2014)** January, pp. 719-725.

[13] Y. Geng and K. Pahlavan, "On the accuracy of rf and image processing based hybrid localization for wireless capsule endoscopy", IEEE Wireless Communications and Networking Conference (WCNC), **(2015)** March.

[14] X. Li, Z. Lv and J. Hu, "Traffic management and forecasting system based on 3d gis", Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on, **(2015)**, pp. 991-998.

[15] S. Zhang and H. Jing, "Fast log-Gabor-based nonlocal means image denoising methods", Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 2724-2728.

[16] J. Hu and Z. Gao, "Distinction immune genes of hepatitis-induced heptatocellular carcinoma", Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.

Author

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.