International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.10, No. 1 (2017), pp.81-88
http://dx.doi.org/10.14257/ijsip.2017.10.1.09

Recognition Method Based on Size in Depth Image

Jian Zhang[1,2*]and Wanjuan Song[2,3]

[1]College of Sport Engineering and Information Technology, Wuhan Institute of Physical Education,Wuhan, Hubei,430079, China
[2]Computer School of Wuhan Universit ,Wuhan University,Wuhan, Hubei,430079, China
[3]College of Computer, Hubei University of Education,Wuhan, Hubei,430079, China
zjgo1979@163.com

Abstract

*According to the differences of the position and size of human parts in depth image, a method which represents human part size and position characteristic is designed. The random forest classifier is used for studying. Through the experiments, the relationship between tree count and depth is shown. And the scheme with excellent random forest classifier is selected. The recognition accurate rate of the classifier to the human pixel reaches 90%. For 320*240 depth image, the processing speed of Intel single-core 1.6 GHZ processor is 15ms/fps. And it has the real-time processing capacity.*

***Keywords**: Body Recognition; Part Size; Random Forest Classifier; Depth Image*

1. Introduction

The wholesome interactive human tracking has been applied [1] in the industries, such as game, man-machine interaction, safety engineering, Tele-presence and hygiene, etc. However, the technical research capable of recognizing the common body action and meeting the needed speed of real time interaction on the hardware cannot meet the requirements [9], [17-18]. Along the appearance of depth camera , the work is simplified greatly [3-7].

Human posture estimation is a research hot spot in the current computer vision field [1-8]. The human posture is estimated in the depth image [7]. Firstly, the error measurement between the model and the observation is defined. And the posture parameters are searched so that the error between the human model and the observation is minimum. The error function has multiple local limit values. During the error minimizing process, the intelligent optimization algorithm, such as particle swarm [9], hierarchy particle swarm [10], hierarchical evolutionary [2] and the like are used for solving global optimum. And the algorithm has poor real-time. In order to improve the speed, Mussi [9] realizes particle swarm algorithm on GPU. In view of real-time of the algorithm, the optimal posture is not likely to be searched and solved in the whole posture space. The posture of last frame is used as the initial value, which is searched [1] locally. The common method is the method [3],[11-12] based on Bayes frame and method [13-14] based on joint body registration. Siddiqui [17] pointed out that the method based on Bayes frame has better effect, however, it has slow speed. And the speed of the method based on joint body registration is quicker 90 times than the method based on Bayes frame. The method for estimating the posture by local searching has quick operation speed and can reach higher precision, however, it is initialized manually. When human moves seriously, and the method is easy to get into local optimum. PrimeSense requires the user to make out surrender posture to finish the

[*] Corresponding Author

initialization. In order to process the local optimum problem, Andriluka [15] uses the tracking by detection.

Microsoft Research Cambridge Shooton [4] regards the human posture estimation as the classification problem of each pixel. The probability of each part of each pixel belonging to the human model is calculated. The pixel is combined to obtain the pixel of each part. And then the assumption of the joint position is generated. The method obtains better effect and provides initial value to the local searching method. However, the method has slow calculation speed. The processing speed on eight-core processor is 5ms/fps. The method needs great sample size. And the Microsoft introduces that the sample size is TB level.

The text estimates the human posture by classifying each pixel. The representation human part size and position characteristics are designed based on the position distribution of each part in the human. And the type of each pixel of human surface is recognized by studying suitable sample. The content of the text is arranged as follows: the Section 2 describes the position of the representation part and the characteristics of size difference. In the Section 3, we test the performance of the classifier based on the characteristic training random forest classifier. And the content of the text is summarized in Section 4.

.

2. Characteristic Description

2.1. Depth Image Data Acquired

It needs to collect a great number of human posture images for training to recognize and test the common postures. There is no public depth image human posture standard library at present, so a small sample library is built by PrimeSense depth image sensor. As shown in Figure 5, the hands can generate various different actions only, so the actions of human cannot be counted fully. In order to avoid a great of repeated work in the action selection stage, the action samples with obvious change are selected to generate the depth pictures corresponding to different shapes and sizes of human. The bit depth of the picture is 16bit; and the gray value of the image is 0-65535. When the gray value is 0, the distance is the near effective distance; when the value is 65535, the distance is the longest distance supported by PrimeSense depth image sensor.

Figure (1). Depth Image Sample

2.2. Problem Analysis of Depth Image Characteristics Extraction

In characteristics extraction, the excellent characteristics have higher distinction degree to different samples; and the dimension and calculated amount of the characteristics are reduced as far as possible. The gradient and point characteristics are two common characteristics in the characteristics extraction of visible light body recognition. The gradient characteristic comprises Canny operator [16], Laplace-Gaussian [14] operator and direction gradient histogram [5] (HOG) and the like. The front two operators can detect the points at the edge of the image, namely, the points of which the gray value changes greatly. The two methods are likely to divide the image into several area blocks not communicating with each other. The body recognition in complex scene shall process the image further to remote the interference of other noise. If the classical morphological method is combined with the threshold segmentation, it can remove a small part of noise and damage the form of the detection target. The direction gradient histogram is the classical method in human detection and recognition; it has the advantages of high processing precision and good detection effect and the shortages of high dimension and great calculation overhead; sometimes, it needs thousands of dimensions, so it is hard to ensure the real time processing. On the other hand, the common point characteristic comprises angle , SIFT [6] and the like. Although it has low dimension, the point characteristic is hard to adapt to the changeable postures under chaos background; the point characteristic needs cluster operation, so it enlarges the difficulty of solving the problems. It is not a good solution to adopt the gradient characteristic or point characteristic singly.

2.3. Method

The text adopts boardcard human mode, which is divided into body, head, left upper arm, left lower arm, right upper arm, right lower arm, left thigh, left leg, right thigh, right leg and the like and as shown in Figure 2.

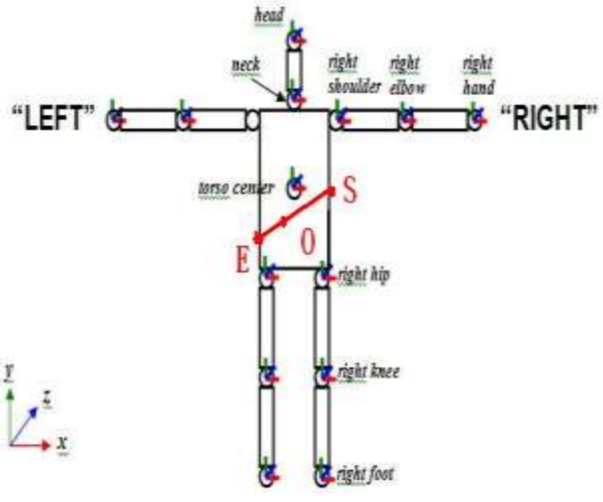

Figure (2). Sketch Map of Scanning Line

Each part has different distribution positions. For example, in most cases, the head is above the body. And the left upper leg, left lower leg, right upper leg and right lower leg are distributed below the body. The human parts have different sizes, for example, the body size is greater than that of left and right lower arms.

Based on the above characteristics, the text designs the representation human part position and size characteristic.

(1) Human part position characteristic

The three-dimensional coordinate of the pixel is directly used for describing the position characteristic; assuming that the pixel coordinate is (x,y,z), and the characteristic vector of the representation pixel position is $[x,y,z]^T$.

(2) Human part size characteristic

For the pixel in the human depth image, the text takes the length passing through the scanning line of the pixel as the size of the part of the representation pixel. In order to avoid interpolation, 0 degree, 45 degrees, 90 degrees, 135 degrees, 180 degrees, 225 degrees, 270 degrees and 315 degrees of scanning lines are used. m degrees of direction and m+180 degrees of direction form a straight line. The scanning line is defined all pixels from the first jump in m degrees direction to the first jump of m+180 degrees direction; the jump refers to the two adjacent pixels along the scanning direction. The background pixel is changed to the human pixel or the human pixel is changed to the background pixel. As shown in figure 1, the first jump pixel passing through 45 degrees direction of O point is S. The first jump pixel in 225 degrees direction is E. And a scanning section passing through O is SE.

The length of the scanning section is defined as the Euclidean distance of two points of the scanning section. The beginning pixel is three-dimensional coordinate (x_s, y_s, z_s). The ending pixel is three-dimensional coordinate (x_e, y_e, z_e), the scanning size is as follows:

$$d = \sqrt{(x_s - x_e)^2 + (y_s - y_e)^2 + (z_s - z_e)^2} \tag{1}$$

For each pixel, the text adopts the length of four scanning sections as the size of the part of the pixel; the size characteristic vector of the representation size is $(d_0, d_{45}, d_{90}, d_{135}, x, y, z)^T$. d_0 shows the size of the scanning section passing through the level direction of the pixel. d_{45} shows the size of the scanning section passing through 45 degrees direction of the pixel. d_{90} shows the size of the scanning section passing through the vertical direction of the pixel. and d_{135} shows the size of the scanning section passing through 135 degrees direction of the pixel.

(3) Characteristic normalization

Each pixel obtains the characteristic vector $(d_0, d_{45}, d_{90}, d_{135}, x, y, z)^T$, the characteristic vector is subjected to size normalization and position normalization. During the position normalization process, PCA analysis is carried out firstly. And then PCA is reprojected again. And the reprojected coordinate is divided by the height. The size normalization is divided by the height by the length of the scanning section, namely, $(d_0 | \text{height}\ , d_{45} | \text{height}\ , d_{90} | \text{height}\ , d_{135} | \text{height}\ )^T$. Here, the height is obtained by the eigenvalue of maximum of the covariance matrix in PCA process. We think that the height is in direct proportion to the square root of the eigenvalue of maximum, namely,

$$height = k \times \sqrt{\lambda_{\max}} \tag{2}$$

The sample is subjected to normalization analysis to obtain k=4.

3. Experiment and Analysis

In this section, the human pixel part is marked by joint body registration algorithm. The performance of the classifier is tested by three testing collections. And the performance of the classifier is analyzed. In addition, we analyze the limitation and reasons of the characteristics of this text.

(1) Sample mark

The text uses PrimeSense depth image sensor to collect the motion process of human. The human posture is solved by the joint body registration algorithm. For the pixel in the human point cloud, assuming that the part of the pixel is the human part closet to the pixel, the sample mark is rough. And the pixel at the joint is such. In order to avoid accurate mark, we use the non-boundary pixel of each part as the sample during the training process.

(2) Test set description

We design three test sample sets. In sample set 1, the body is direct and the arms move, there are 600 testing images. Partial testing sample images are as shown in Figure (3). (a). The walking action is in sample set 2. There are 500 testing images. Partial testing sample images are as shown in Figure (3). (b). Human can move freely in sample set 3. The action approaches to dancing. There are 1,000 testing images. And partial testing sample images are as shown in Figure (3). (c). We do not process the shielding conditions, so three sample sets have no image with serious shielding.

Figure (3). (a) Test Set 1: Body does not Move; Arms Move;

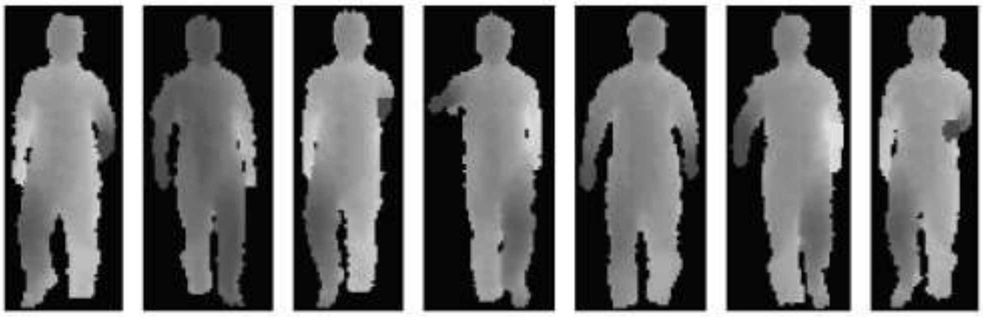

Figure (3). (b) Test Set 2: Walking

Figure (3). (c) Test Set 3: Freely Moving, Approaching to Dancing

(3) Performance of classifier

The random forest classifier is a multi-class classifier with the advantages of quick studying and identifying speed and not overfitting. And the text adopts the random forest classifier.

In the random forest, the main parameters comprise number of tree in the forest and depth of each tree. Trees (3-9) with different amounts and training classifier with different depth (6-12 layers) are used. And the performance of the classifier is tested by three test sets. The identifying speed is in direct proportion to the number of the tree and the depth of the tree, the text focuses on the indictor of the identifying rate. Under different parameters, the resolution ratio of the classifier obtained by three test sets is as shown in Figure 3.

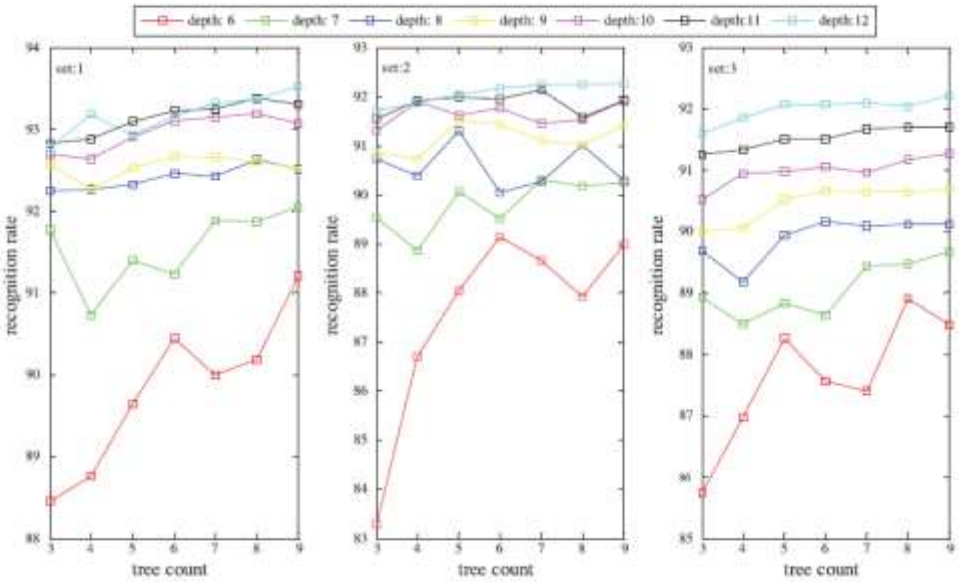

Figure (3). The Resolution Ratio of the Classifier Obtained by Three Test Sets Under Different Parameters

The tree depth plays an important role to the resolution ratio. And the number of the tree can prevent the overfitting during the training process. It can be found from Figure 4, that the resolution ratio of the classifier to three test sets is above 90% for any number from 3 to 9 when the tree depth reaches 9 layer or above. The resolution ratio is improved along the increased depth of the tree. And the phenomenon is obvious in test set 3. When test set 3 is used for testing the classifier, the resolution ratio of seven classifiers whose depth is k+1 is higher or approaches to the greatest resolution ratio of the seven classifiers whose depth is k. This is because that the test set 3 comprises each kind of human posture. It requires complex classifier and greater depth.

The recognition rate and the recognition speed are considered fully. Four trees and ten-layer random forest classifier are adopted. And the resolution ratio of the three test sets is 92.53%, 91.86%, 90.88%.

The recognition speed is tested on Intel single-core 1.6 GHZ processor without considering the extraction time of characteristic. For 320*240 depth image, the average characteristic extraction and recognition speed is 1.86us/pix. When the human is closed to the depth sensor 2.5m, the human who is 1.7m has 8,000 points, and the processing time is 15ms.

Figure 4, show some results.

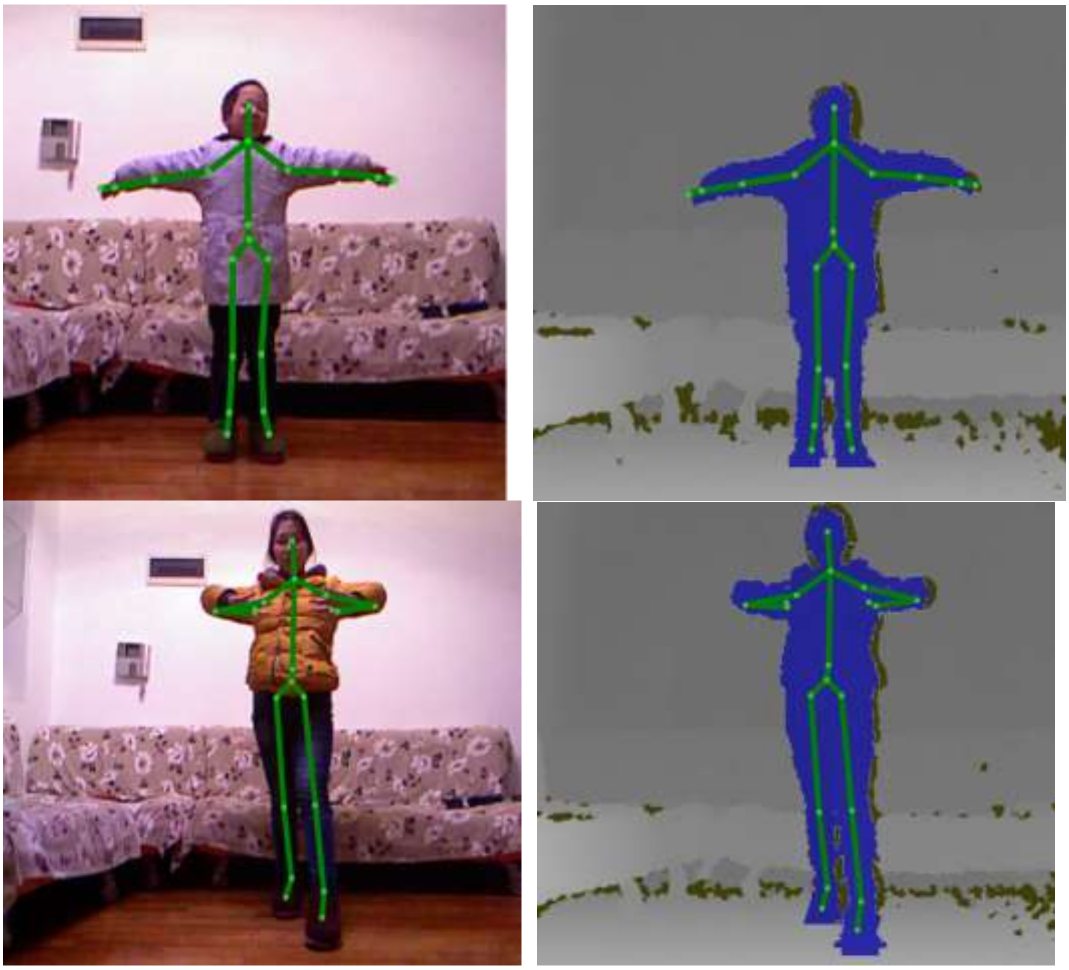

Figure (4). Some Detection Results

4. Conclusion

The text estimates the human posture by classifying each pixel. The representation human part size and position characteristics are designed based on the position distribution of each part in the human. And the type of each pixel of human surface is recognized by studying suitable sample. Based on the differences of the position and size of human parts, the text designs a representation human part size and position characteristic. The random forest classifier is used for studying. Through the test, the scheme with excellent random forest classifier is selected. The recognition accurate rate of the classifier to the human pixel reaches 90%. For 320*240 depth image, the processing speed of Intel single-core 1.6 GHZ processor is 15ms/fps. And it has the real-time processing capacity.

Acknowledgment

This work was financially supported by Science and Technology Department of Hubei Province, China (No.2014CFB383) and Wuhan Institute of Physical Education Foundation (No.2016QZ07).

References

[1] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture", Computer Vision and Image Understanding, vol. 81, **(2001)**, pp. 231–268.

[2] C. Robertson and E. Trucco, "Human body posture via hierarchical evolutionary optimization", Image and Vision Computing, vol. 28, no. 11, **(2010)**, pp. 1530-1547.

[3] Y. D. Zhu and K. Fujimura, "A bayesian framework for human body pose tracking from depth image sequences", Sensors, **(2010)**.

[4] J. Shotton, "Microsoft Res.", Cambridge, UK;Efficient Human Pose Estimation from Single Depth Images.Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 12, **(2013)** December, pp. 2821 - 2840.

[5] E. KALOGERAKIS, A. HERTZMANN and K. SINGH, "Learning 3D mesh segmentation and labeling", ACM Transactions on Graphics, vol. 29, no. 3, **(2010)**, pp. 101-102.

[6] D. LOWE, "Distinctive image features from scale-invariant key points", International Journal of Computer Vision, vol. 60, no. 2, **(2004)**, pp. 91.

[7] K. Berger, "The role of RGB-D benchmark datasets: an overview", Comput Res Reposit, **(2013)**, pp. 4321–4326

[8] K. Berger, "A state of the art report on multiple RGB-D sensor research and on publicly available RGB-D Datasets. In: Computer vision and machine learning with RGB-d sensors", Springer International Publishing, **(2014)**, pp. 27–44.

[9] L. Mussi, S. Ivekovic and S. Cagnoni, "Markerless articulated human body tracking from multi-view video with GPU-PSO", Proceedings of the 9th International Conference on Evolvable Systems: from Biology to Hardware, **(2010)**.

[10] J. Vijay, S. Ivekovic and E. Trucco, "Articulated human motion tracking with HPSO", Proceedings of The Fourth International Conference on Computer Vision Theory and Applications, **(2009)**, pp. 531-538.

[11] H. Sidenbladh, M. J. Black and D. J. Fleet, "Stochastic tracking of 3d human figures using 2d image motion", Proceedings of the 6th European Conference on Computer Vision-Part II, **(2000)**, pp. 702–718.

[12] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor", Computer Vision and Pattern Recognition Workshops (CVPRW), **(2010)**.

[13] A. Hernández-Vela, N. Zlateva and A. Marinov, "Graph cuts optimization for multi-limb human segmentation in depth maps", Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, **(2012)**, pp. 726-732.

[14] H. Jiang and J. Xiao, "A Linear Approach to Matching Cuboids in depth Images. Computer Vision and Pattern Recognition (CVPR)", 2013 IEEE Conference on, **(2013)** June, pp. 2171-2178.

[15] M. Andriluka, S. Roth and B. Schiele, "Monocular 3D pose estimation and tracking by detection", Computer Vision and Pattern Recognition Workshops (CVPRW), **(2010)**.

[16] A. Hernández-Vela, N. Zlateva and A. Marinov, "Graph cuts optimization for multi-limb human segmentation in depth maps", Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, **(2012)**, pp. 726-732.

[17] X. Chang, Y. Yang, E. Xing and Y. Yu, "Complex event detection using semantic saliency and nearly-isotonic SVM", In Proceedings of the 32nd international conference on machine learning (ICML-15), **(2015)**, pp. 1348–1357.

[18] H. Gjoreski, M. Luštrek and M. Gams, "Accelerometer placement for posture recognition and fall detection", In Intelligent environments (IE), 2011 7th international conference, **(2011)**, pp. 47–54.

Authors

Jian Zhang, He is a lecturer of College of Sport Engineering and Information Technology, Wuhan Institute of Physical Education,Wuhan, Hubei, China. Now he is pursuing the Ph.D. degree in Computer School, Wuhan University. His research interests include image processing and pattern recognition.

Wanjuan Song, She is a lecturer of College of Computer, Hubei University of Education,Wuhan, Hubei, China. Now she is pursuing the Ph.D. degree in Computer School, Wuhan University. Her research interests include machine learning and pattern recognition.