3D Depth Camera Based Human Posture Detection and Recognition Using PCNN Circuits and Learning-Based Hierarchical Classifier

Hualiang Zhuang, Bo Zhao, Zohair Ahmad, Shoushun Chen and Kay Soon Low
School of Electrical and Electronic Engineering, Nanyang Technological University
Singapore, 639798
Email: eechenss@ntu.edu.sg

Abstract — **A new scheme for human posture recognition is proposed based on analysis of key body parts. Utilizing a time-of-flight depth camera, a pulse coupled neural network (PCNN) is employed to detect a moving human in cluttered background. In the posture recognition phase, a hierarchical decision tree is designed for classification of body parts so that the 3D coordinate of the key points of the detected human body can be determined. The features described in each individual layer of the tree can be chained as hierarchical searching indices for retrieval procedure to drastically improve the efficiency of template matching in contrast to conventional shape-context method. Experimental results show that the proposed scheme gives competitive performance as compared with the state-of-the-art counterparts.**

Keywords-human posture recognition; depth image; PCNN; decision tree.

I. INTRODUCTION

Real-time human posture recognition and human activity detection are tremendous challenges to computer vision systems. The related topics have attracted much research in recent years due to their academic significance and valuable applications, and an impressive series of research work has been reported in this field [1-3]. In general, those approaches first detect moving objects by the analysis of video stream, and then extract human silhouettes using background subtraction techniques [4]. Then posture profiling is conducted based on frame-by-frame posture classification algorithms [5]. One of the obvious difficulties for these approaches is the effect of varying illumination, appearance of dress and the cluttered background. To circumvent these obstacles, one has to manage the camera configuration adaptive to the varying illumination, and code sophisticated algorithms for analyzing attire texture, and design robust background models for extracting the subject [6]. These tedious computations become a bottleneck for the overall processing speed, and may make the algorithm unsuitable for more real-time applications.

Moreover, the conventional 2D imaging scheme is sensitive to varying view points. This problem motivates researchers to investigate 3D image analysis for posture recognition [7-8]. For the 3D imaging sensory part, stereovision and depth measurement are two typical techniques. The crucial shortcoming of the stereovision technique is that its accuracy decreases drastically with the increasing distance from camera head to the object. In contrast, depth camera can provide significant improved measurement accuracy within the working range [9]. Hence we will adopt a time-of-flight depth camera, SwissRanger SR4000, as a 3D imaging system for this research.

Human detection is the prelude of human posture analysis. Based on depth images, the foreground human figure is easier to be segmented in 3D space in contrast to 2D projection. However, simple thresholding is not always viable for applications with varying and cluttered backgrounds. To improve this case in this research, the foreground human figure will be detected based on a self-organizing pulse coupled neural network (PCNN) [10] for segmentation and temporal difference between sequential depth images for moving human detection. One of the salient points of this design is that the temporal-pulse-based PCNN computing could utilize the consistent domain of signal from the time-of-flight depth sensor. This provides a seamless interface between the image sensor and the processing circuits. Another point of this design is that the proposed PCNN is a parallelism-oriented network of circuits different from conventional soft computing. It is expected to be implemented on hardware platform in massive parallel so that the processing speed can be improved to real-time level to handle cases with time varying background. For the posture recognition phase, the conventional feature based template-matching algorithm could have high time complexity if the

number of templates is too large. In this paper, learning based human body recognition algorithm will be designed to establish decision trees to classify body parts in the input frame of depth image, so that the joint points of human body can be located as the posture features. The hierarchical design of the decision tree circumvents over-fitting issue of the shape context algorithm and the time-consuming issue of the template-matching approach. This leads to a scheme with competitive performance and fast processing speed for recognizing the human posture in real time.

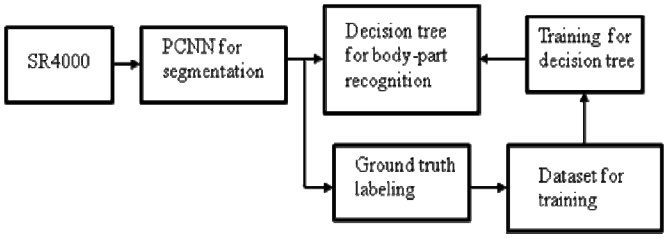


Fig. 1. Diagram for the scheme.

II. Overview of the System

A. Diagram of the scheme

The overview of the scheme is illustrated in Fig. 1. Human detection is the first step of the human posture analysis. The foreground human figure will be detected using a self-organizing PCNN based segmentation. The human figure is then fed to decision trees to perform body-part classification and localization for posture recognition. Each of these decision trees has hierarchical structure. It is established beforehand by using off-line training based on the local shape context features of labeled body-parts.

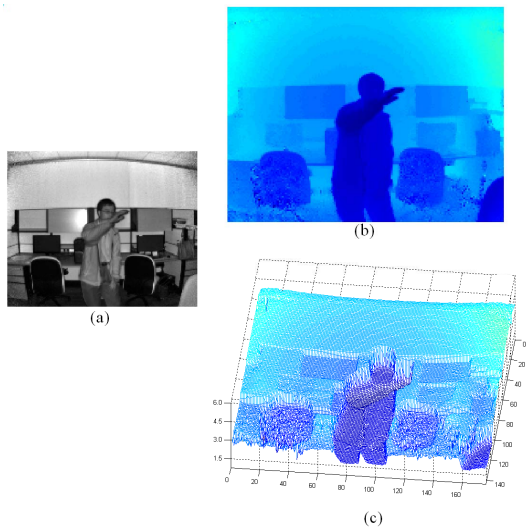


Fig. 2. Visualized depth image. (a) intensity image; (b) depth image visualized using pseudo color map; (c) 3D view.

B. Depth image

The depth image acquired by SR4000 is a 176×144 pixel array, where each pixel is the value of the distance from the reflective point to the camera head. Fig. 2 illustrates an example depth image, where Fig. 2(a) is the intensity image, and 2(b) visualizes the 176×144 array of depth values using pseudo color map. Fig. 2(c) is the 3D view reconstructed from the depth data.

C. Extracting Human Figure from Background

Systematic processing of human activity analysis usually starts with human detection. In applications of indoor/outdoor surveillance and home care of elderly, targeting human figure could be disturbed by cluttered background and occasionally movable furniture in the scene. To cope with this issue, neural circuit array based on pulse coupled neural network model [10] is employed to perform the scene segmentation. The architecture of PCNN is a 2D array of interlinked neurons. Each neuron maps one pixel of the corresponding image, where the pixel feature is the depth quantity of that point in depth image. To perform image segmentation, PCNN neurons are supposed to link to each other according to the similarity of corresponding mapped pixels. Homogenous regions mapped by corresponding neurons will be segmented in the result. The PCNN circuit in [10] is a hardware-oriented variant of original PCNN. This neural circuit array can be running in parallel on a FPGA chip to perform the segmentation in real-time. Therefore it is viable for analyzing cases with varying backgrounds.

In this research, the region of interest (ROI) is certainly the human figure. To determine the ROI from the segmented image frames, difference of two sequential depth images is utilized based on the assumption that people is the moving object in the scene. Let set $\boldsymbol{D}(t)$ denote the depth image of frame t. The temporal difference can be computed as follows,

$$\boldsymbol{E}(t) = \widetilde{\boldsymbol{D}}(t) \oplus \widetilde{\boldsymbol{D}}(t-1) \tag{1}$$

where $\boldsymbol{E}(t)$ is the binary map for temporal difference (after denoise operations); $\widetilde{\boldsymbol{D}}(t)$ is obtained from $\boldsymbol{D}(t)$ with truncated quantization precision; ' $\oplus$ ' denotes pixel-wise exclusive OR. Now we denotes the segmentation result of PCNN as

$$\boldsymbol{D}(t) = \bigcup_{j=1}^{N_j} \boldsymbol{G}_j \quad \text{with } \boldsymbol{G}_j \cap \boldsymbol{G}_i = \Phi \quad (j \neq i) \tag{2}$$

where $\boldsymbol{G}_j$ denotes a resultant segment; j and i are segment indices; N_j is the total number of resultant segments. The segment of the human figure can be determined by considering the following two maps ('•' denotes and' operation),

$$\boldsymbol{S}_t = \boldsymbol{D}(t) \bullet \boldsymbol{E}(t) \tag{3}$$

$$\boldsymbol{S}_{t\text{-}1} = \boldsymbol{D}(t-1) \bullet \boldsymbol{E}(t)\,. \tag{4}$$

Consider the resultant nonzero pixels. If the depth values of a set of nonzero pixels in $\boldsymbol{S}_t$ is less than the corresponding values of map $\boldsymbol{S}_{t-1}$, $\boldsymbol{D}(t)$ is the foreground frame. Otherwise $\boldsymbol{D}(t-1)$ is the foreground frame. This is because the object nearer to the camera is the default foreground object. Then the segment $\boldsymbol{G}_j$ of the foreground frame will be selected as the human figure if it has overlap with the nonzero pixels in map $\boldsymbol{E}(t)$. This operation overcomes the difficulty for determining foreground from a temporal difference map in case of using 2D intensity (or color) image. Then the corresponding frame is selected for further analysis. Fig 3 summarizes the flowchart for extracting the human figure of interest from depth image with cluttered background. In this phase of detection, the precision of our method is 93%, slightly higher than the recall (92%). This is because our infrared depth camera is confined in the indoor working environment, where disturbance from other moving objects could be reduced much more than outdoor environment. The less false positive rate results in higher precision rate.

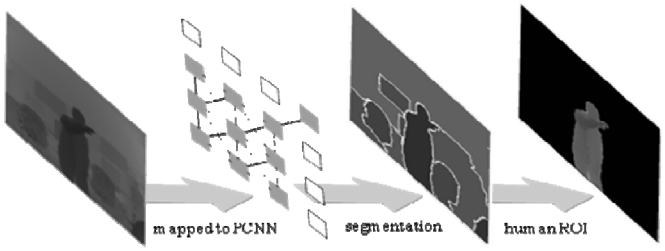


Fig. 3. Extracting human figure from background using PCNN.

III. LEARNING BASED HUMAN BODY PART RECOGNITION

A. Decision Trees for Body Part Classification based on Hierarchical Features

In conventional object recognition schemes, the feature vectors of the detected object are usually utilized for template matching [2, 11]. Time complexity of the matching procedure is proportional to the number of templates. For body-part recognition, the number of templates could be very large because of the versatility of human postures. This will result in a time consuming matching procedure. To circumvent this issue, we design a decision tree for body part classification based on hierarchical features. In the training procedure, the features will be extracted for different layers of the decision tree. These features are then chained as hierarchical searching indices for retrieval procedure, so that the efficiency of template matching is improved.

To establish the decision tree, we need to define the features for each layer of the decision tree. In this research, a modified scheme of shape context is employed. In general, the shape feature of a pixel in ROI is defined as a set of the average depth of *feature regions* (also called *feature bins*) near the current pixel. Each layer of the decision tree has its own definition of the feature bin.

As shown in Fig. 4, *feature bins* about every pixel in ROI are defined in various heuristic shapes for each node of a layer. These layers of nodes can be cascaded to build a multilayer decision tree. For each foreground pixel, a local descriptor will be defined based on these local bins. Extracted features are similar to the shape context features [11], but efficiency of the retrieval process is drastically improved due to the hierarchical structure. In detail, average depth in each region is quantified as two or four levels as the feature of this layer of tree nodes, where the two-level quantification focuses on silhouette shape feature, and the four-level quantification focuses on depth information. Fig. 4(a) illustrates examples of pie-shape feature bins and circular feature bins with two-level descriptors. For a pixel $\boldsymbol{x}$, the node index of layer l selected for $\boldsymbol{x}$ can be quantified as follows,

$$h(\boldsymbol{x},l) = Q_l(r_1^l(\boldsymbol{x}), r_2^l(\boldsymbol{x})) \tag{5}$$

where l is the layer index; $r_1^l(\boldsymbol{x})$ and $r_2^l(\boldsymbol{x})$ are the functions for computing the average depth values of the two feature bins in layer l about $\boldsymbol{x}$; Q_l denotes the quantization function. About pixel $\boldsymbol{x}$, the quantization and branching through the layers of the tree can be illustrated as the example in Fig. 4(b). The node indices through the highlighted searching path are $h(\boldsymbol{x},1) = 01\text{B}$, $h(\boldsymbol{x},2) = 10\text{B}$, and $h(\boldsymbol{x},3) = 10\text{B}$, where 00B~11B are binary codes as index for node 1 ~ node 4. We can say $h(\boldsymbol{x},l)$ has local shape-context feature about $\boldsymbol{x}$, where the shape context is related to the body part containing $\boldsymbol{x}$. For a single layer of the tree, the feature characterized by $h(\boldsymbol{x},l)$ could be undistinguishable. Nonetheless, hierarchical local descriptor based layers of nodes will strengthen the capability of feature classification as $h(\boldsymbol{x},l)$ can be chained to vectors

$$\boldsymbol{H}(\boldsymbol{x}) = [h(\boldsymbol{x},1), h(\boldsymbol{x},2), \ldots, h(\boldsymbol{x},N_l)]^T , \tag{6}$$

to span a multi-dimensional feature space, where pixel $\boldsymbol{x}$ can be classified. Here N_l is the total layer number of the decision tree.

Based on the multilayer decision tree, the probability of various body parts, to which the current pixel belong, can be accumulated at each leaf node of the decision tree by training based on a set of labelled samples. The abovementioned probability can be computed as follows,

$$P(b \mid n_t) = \frac{N_b(n_t)}{\overline{N}_b} \tag{7}$$

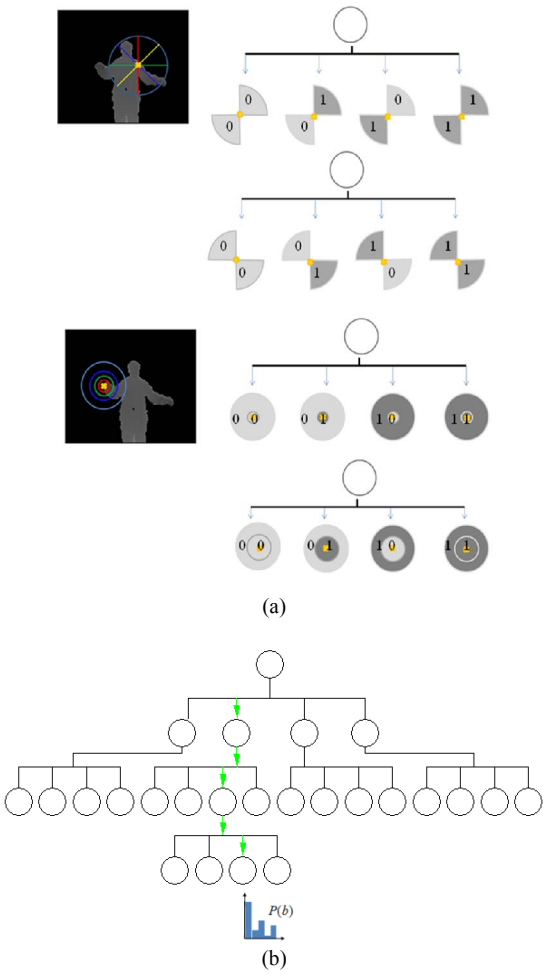
where b is the body part label; n_t is the index of the leaf node; $\overline{N}_b$ is the total number of pixels labeled as b; $N_b(n_t)$ is the number of pixels with label b, which is branched to leaf node n_t.

For the retrieval process, a pixel will be classified to the body part with the maximum probability of the resultant leaf nodes.

$$z = \arg\max_{b}(P(b \mid n_t)) \tag{8}$$

where z denotes the resultant body part. The retrieval process is a hierarchical search using the index of the nodes selected by (5). Taking Fig. 4(b) as an example again, the node indices through the highlighted searching path are 01B for layer 1, 10B for layer 2, and 10B for layer 3 respectively. All are 1 over 4 selections. Thus total number of index searches is within 4+4+4=12. The linear complexity gives a significant improvement in contrast to the total number of leaf nodes, 4×4×4=64, as number of template matching in conventional shape context methods.

Once all the pixels in ROI are classified, the key points of body parts and joints will be computed as the weighted centroid of each individual class. Then the 3D coordinates of these key points can be determined.



(a)

(b)

Fig. 4. Decision trees. (a) various feature bins for building the tree; (b) indexed searching path in retrieval phase.

B. Translation, Scaling and Rotation Invariance

The translation invariance is obvious as the feature bins are local regions about the current pixel. For the scaling invariance, the fact that scaling ratio is proportional to the depth from the camera head can be considered. In the scaling invariance design, the radial range of the feature regions and the thresholds for 2-level/4-level quantization are scaled according to the average depth of the ROI. The rotation invariance is not required as human postures are orientation-dependent in case special postures such as bending and lying down are supposed to be recognized.

Fig. 5. Examples of training data. Upper row: intensity images; lower row: depth images.

IV. Experimental Results and Data Analysis

In experiments, 1500 depth image samples extracted from 150 clips of depth video (series of continuous depth images) are acquired for the research, where 1200 for training and the other 300 for retrieval test. For training data, the different model peoples perform formulated actions so that various postures are sampled. Fig. 5 shows some examples of training images, where a model people performs some postures of Taiji in the first 5 images, and another people poses 2 arms in the last 5 images.

To perform the training of the decision trees, all sampled depth images are manually labeled to highlight the ground truth of body parts. Then the training procedure described in Section III is conducted so that the probability distribution in (7) can be obtained for each leaf node. Using the trained tree, retrieval process can be performed as (8). Fig. 6 visualizes some results of body part recognition.

The hierarchical decision tree increases the dimensions of the feature space with the increasing layers of nodes. In general, this can result in improved classification accuracy with the cost of more computations. Fortunately, the increment of computing has linear relationship to the number of tree layers due to the hierarchical structure. Fig. 7 summarizes the trend of average classification error and the retrieval computing complexity with the increasing number of tree layers. The result verifies the estimation.

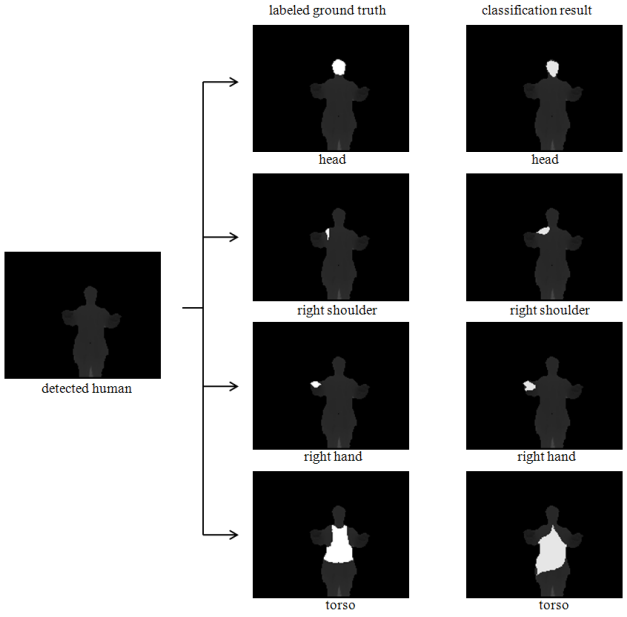


Fig. 6. Examples of retrieval results.

To evaluate the proposed scheme, the counterpart methods reported in [8] and [12] are considered for comparison. From Fig. 8, it is observed that the accuracies of the proposed scheme are competitive to the two counterparts for the classification of head and hands. For the classification of shoulders, the proposed scheme outperforms the method in [12] but uncompetitive to the approach in [8]. The strength of approach in [8] is their high number of tree layer, which is as high as 20. In contrast, the proposed decision tree has only up to 6 layers. Therefore it has much lower hardware requirements.

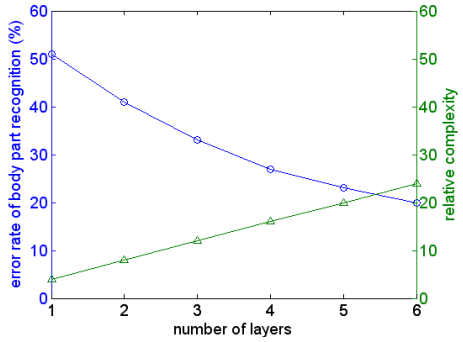


Fig. 7. Impact to error rate and computing complexity by various number of tree layers. The measure of computing complexity is the number of the nodes.

V. Conclusion

This paper proposes a new scheme for human body part based posture recognition. Utilizing a time-of-flight depth camera, a moving human can be detected by employing a PCNN circuit array. This overcomes the shortcoming of conventional model subtraction algorithm, which is unable to handle the case with varying background. To determine the 3D coordinates of the key points of the detected human body, a hierarchical decision tree is designed for classification of body parts. On the one hand, the hierarchical tree circumvents the computing burden of the conventional template matching approach. On the other hand, it gives competitive performance as compared with the state-of-the-art counterparts.

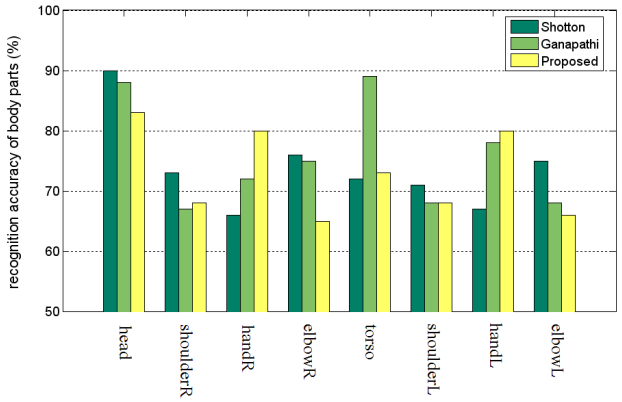


Fig. 8. Comparison with counterparts.

References

[1] C.F. Juang and C.M. Chang, "Human body posture classification by a neural fuzzy network and home care system application," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 37, pp. 984-994, 2007.

[2] S. Chen, P. Akselrod, B. Zhao, J. A. P. Carrasco, B. Linares-Barranco and E. Culurciello, "Efficient feedforward categorization of objects and human postures with address-event image sensors", *IEEE Trans Pattern Anal. Machine Intell*, vol. 34, pp. 302-314, 2012.

[3] J.W. Hsieh, C.H. Chuang, S.Y. Chen, C.C. Chen and K.C. Fan, "Segmentation of Human Body Parts Using Deformable Triangulation," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol.40, pp. 596-610, 2010.

[4] E. H-Jaraha, C. Urunuela, and J. Senar, "Detected motion classification with a doublebackground and a neighborhood-based difference," *Pattern Categorization Letters*, vol. 24, pp. 2079–2092, 2003.

[5] L. H. W. Aloysius, G. Dong, H. Zhiyong, and T. Tan, "Human posture categorization in video sequence using pseudo 2-d hidden markov models," in *8th Control, Automation, Robotics and Vision Conference*, Dec. 2004,pp. 712–716.

[6] L. Wang, W. Hu and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, pp. 585 – 601, 2003.

[7] J. Gu, X. Ding, S. Wang and Y. Wu, "Action and Gait Recognition From Recovered 3-D Human Joints," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol.40, pp. 1021-1033, 2010.

[8] J. Shotton *et al*, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Proc. CVPR*, pp. 1297-1304, 2011.

[9] C. Ye and M. Bruch, "A visual odometry method based on the SwissRanger SR4000", in *Proc. Of SPIE*, vol. 7692, pp. 76921I(1-9), 2010.

[10] H. L. Zhuang K. S. Low, W. Y. Yau, "Multi-chanel pulse coupled neural network based color image segmentation for object detection," *IEEE Trans. Industrial Electronics*, 2012, in press.

[11] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape context," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 24 no. 24, pp. 509–521, 2002.

[12] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. CVPR*, pp. 755-762, 2010.