Intelligent Human Detection Based on Depth Information

Tzu-Wei Chen, Ku-Ying Lin* and Yon-Ping Chen
Institute of Electrical Control Engineering, National Chiao-Tung University
1001, University Rd., Hsihchu City, Taiwan
jim0954.ece01g@nctu.edu.tw

Abstract

This paper proposes an intelligent human detection system based on the depth information generated by Kinect to find out humans from a sequence of images and resolve occlusion problems. The system is divided into three parts, including region-of-interest (ROI) selection, feature extraction and human recognition. First, the histogram projection and connected component labeling are applied to select the ROIs according to the property that human would present vertically in general. Then, normalize the ROIs based on the distances between objects and camera and extract the human shape feature by the edge detection and distance transformation to obtain the distance image. Finally, the chamfer matching is used to search possible parts of the human body under component-based concept, and then shape recognition is implemented according to the combination of parts of the human body. From the experimental results, the system could detect humans with high-accuracy rate and resolve occlusion problems.

1. Introduction

Recently, the techniques for human detection in images or videos have been widely and intensively studied because they have a variety of applications in intelligent vehicles, video surveillance and advanced robotics. For example, it aims to develop a robot which could take care of children and interact with them. In order to have a better interaction between robot and children, the robots have to judge whether the object in front of robots in the image is child or not. Therefore, human detection is an important and essential tool for the development of robots.

However, detecting humans is still a difficult task because of the variation and occlusion. In order to deal with the problems, many human detection methods [1-9] have been proposed. In general, the overall process could be roughly separated into three main steps: foreground segmentation, feature extraction and human recognition. Foreground segmentation is implemented to filter out background regions which are impossible to contain a human. Further, the appropriate features, like edges [7-9], skeletons [10], etc., would be extracted to detect human efficiently and correctly. Finally, the set of features would be delivered into the human recognition system to obtain the result.

The remainder of this paper is organized as follows. Section 2 introduces the proposed human detection system in detail. Section 3 shows the experimental results, and Section 4 is the conclusions.

2. Intelligent Human Detection System

The proposed intelligent human detection system uses depth images generated by Kinect as input and follows three steps shown in Fig.1. First, the system selects the ROIs based on the histogram projection and connected component labeling. In the second step, the ROI is normalized and then processed by edge detection and distance transformation to extract necessary features. Finally, all the extracted features would be further processed by the human recognition.

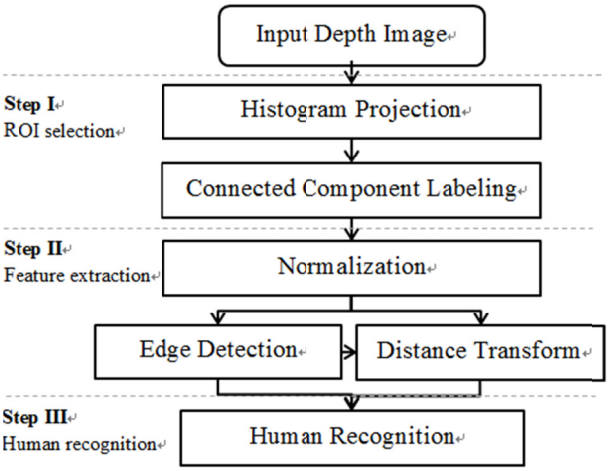


Figure 1. Flowchart of the human detection system.

2.1. Step I: ROI Selection

In general, a human being stands or walks in a vertical attitude, which implies the height of a human in the depth image should exceed a certain value, given as a threshold. Based on the threshold, the system implements the ROI selection from the histogram projection and connected component labeling (CCL).

The system computes the histogram in a depth image with 320×240 pixels according to intensity levels in the range [0, 255]. In this paper, the detection distance is from 1m to 4m and the corresponding intensity range is [40, 240]. Fig.2 shows the histogram image of the depth image. Obviously, there are four objects viewing from Fig.2. This demonstrates that an object would have a clear shape in the histogram image if it height is above a threshold.

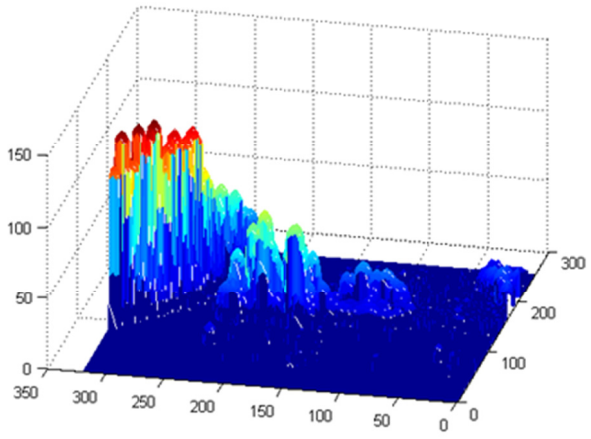
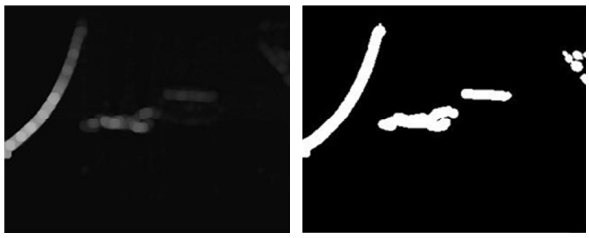


Figure 2. Histogram image of the depth image

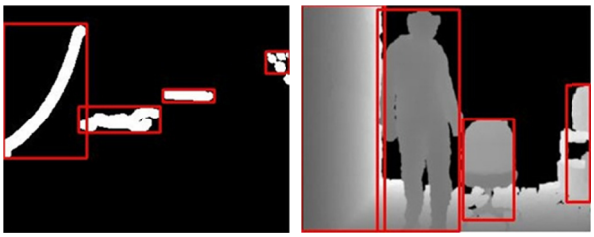
The top-view image of the 3-D distribution is shown in Figure 3(a), where a higher object would possess a larger intensity. Figure 3(b) shows the final result of histogram projection from a given threshold value.



(a) (b)

Figure 3. (a) Top-view image (b) Result of histogram projection

After the histogram projection, the CCL is used to extract the connected regions and compute their areas. If the area of a connected region is too small, i.e., less than a threshold, then the region would be filtered out because it is treated as a non-human object. The result of CCL is shown in Fig.4 (a), where four potential objects are marked by red rectangles and a small dot-like region is filtered out. Then, map the marked objects into the depth image, correspondingly shown in Fig.4 (b). Note that both Fig.4 (a) and Fig.4 (b) have the same horizontal coordinate, but the vertical coordinate of Fig.4 (a) represents the intensity value of Fig.4 (b).



(a) (b)

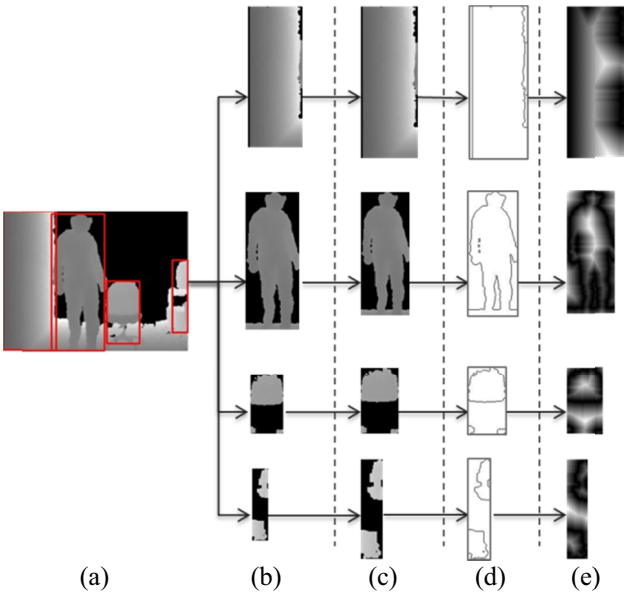
Figure 4. (a) Result of CCL (b) Result of ROI selection

Besides, Fig.4 (a) could also be used to judge the cases of occlusion, which could be roughly classified into four cases: non-occlusion, frontal-occlusion, left-occlusion and right-occlusion. After CCL, the selected ROI would be marked as shown in Fig.4 (a) and the system has to check whether the area contains other objects or not. If so, it is required to determine the case of occlusion from the overlapping region. The occlusion judgment is also a kind of feature and would be sent into the recognition system as a reference.

2.2. Step II: Feature Extraction

After ROI selection, the system has to extract related features, increase the detection rate and decrease the computational cost. The overall feature extraction could be separated into two parts. First, the size of the selected ROI would be normalized based on the distance between object and camera. Then, edge detection and distance transformation are implemented to extract the shape information.

Obviously, if the object is farther from the camera, the object would have smaller size in the image. Hence, normalization is required to reduce the influence by different distances. The concept of normalization is that no matter where the object is, the object would be transformed to the standard distance through normalization. In this paper, the standard distance is set to be 2.4m and the example is shown in Figure 5.



(a) (b) (c) (d) (e)

Figure 5. Scheme of feature extraction. (a) Result of ROI selection (b) Extracted regions of ROIs (a) (c) Result of normalization (d) Result of edge detection (e) Result of distance transformation

After normalization, edge detection and distance transformation are implemented to extract human shape information. Distance transformation (DT) [11-13] is a technique to transform a binary edge image into a distance image. In general, the size of distance image is the same as edge image. The edge pixels in distance image are all set to be 0 and the other pixels contain the distance to the closest edge pixel. In this paper, the 4-neighbor distance is used to compute the distance between a pixel and the closest edge pixel, and the edge detection is executed based on Sobel operators. The edge image is presented in Figure 5(d) and the result of distance transformation is shown in Figure 5(e).

2.3. Step III: Human Recognition

In this step, the system has to judge whether the ROI contains human or not from the extracted features. To achieve higher detection rate and resolve occlusion problems, this paper adopts component-based concept which considers a whole human shape formed by different parts of body. There are two sub-steps, including chamfer matching and shape recognition.

Chamfer matching [13] is a matching algorithm to evaluate the similarity between test image and template image. First, the shape of the target object, such as head, leg, etc., is captured by a binary template. The test image is pre-processed by edge detection and distance transformation. After implementing the DT, the distance image would be scanned by the template image at all locations. Figure 6 is an example of chamfer matching. Figure 6(a) and Figure 6(c) are the test image and template image, respectively, and Figure 6(b) is the distance image of Figure 6(a). The template scans the distance image at all the locations, and the result of chamfer matching would be marked by yellow dots as shown in Figure 6(d).

(a) (b) (c) (d)

Figure 6. Example of chamfer matching

In this paper, chamfer matching is implemented to detect different parts of body, including head, torso and legs. However, when a human is occluded by objects or other humans, there might be only left-side body or right-side body in the image. In order to deal with the occlusion problem, separating the template image into left-side one and right-side one. Therefore, the template set is proposed in Figure 7 which contains six template images, named as left-head (LH), right-head (RH), left-torso (LT), right-torso (RT), left-leg (LL) and right-leg (RL) from left to right..

Figure 7. The template sets.

After chamfer matching, the system has to judge whether the ROIs contain human or not from the coordinates of matched regions of different parts of body. Because of the ability of variation tolerance, chamfer matching has a high true positive rate to correctly detect most of real parts of body, but also has an unwanted high false positive rate to misjudge other objects as parts of body. To cut down the false positive rate, the concept of shape recognition is used in the following process. Since the relations between different parts of body are fixed, these parts could be combined based on their geometric relation.

In the paper, there are two recognition approaches, including voting-based and neural-network-based approaches. These two approaches would be introduced below and their performances would be compared in the next section.

Approach 1: Voting-based recognition

The scheme of voting-based recognition is shown in Figure 8. In order to deal with occlusion problems, a whole human shape is separated into four groups, which are left-, right-, upper-, and lower-group. If a part of body could be combined with an adjacent part, the ROI would have more possibility to contain a human. Take left-head as an example, if left-head could be combined with left-torso, the left-group and upper-group could get one vote. Similarly, if left-head could be combined with right-head, the upper-group would have one more vote. If their relation is reasonable, the corresponding body group would have one more vote. After finding the votes of four groups, the occlusion judgment discovered previously would also be added as a kind of feature, and it is used to adjust the proportions of four groups. After adjusting the proportions, the system sums up these four votes to get the final vote. If the final vote exceeds a threshold, e.g. half of total votes, the ROI would be regarded as human, and vice versa. The concept of voting-based approach is straight and easy to implement. However, the relations between adjacent parts of body and the threshold have to be determined manually.

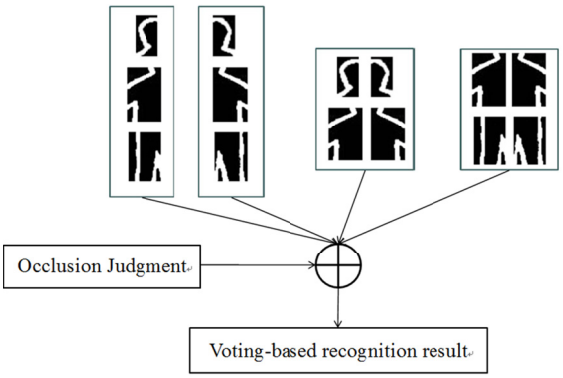


Figure 8. Scheme of voting-based recognition

Approach 2: Neural-network-based recognition

The second approach is using a neural network to combine different parts of body. The concept of neural-network-based recognition is similar to voting-based recognition. There are totally 1500 training data, including 500 positive data and 1000 negative data. The weights of neural network would be adjusted through the process of back-propagation learning. After learning, the human can be recognized according to the output value of neural network.

There are one input layer with 18 neurons, one hidden layer with 30 neurons, and one output layer with 4 neurons as shown in Figure 9. The inputs are the differences between coordinates of different parts of body at x- and y-coordinate. Through the computation of neural network, the four output neurons would get different outputs. These four output neurons represent the performances in left-, right-, upper- and lower-group, respectively. Similar to voting-based recognition, the occlusion judgment is added to adjust the proportions of four groups and then these performances are summed up as final performance. If the final performance exceeds a threshold, the ROI would be regarded as human, and vice versa.

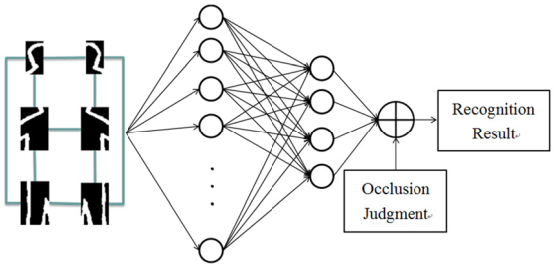


Figure 9. Structure of the neural networks

3. Experimental results

In order to examine the robustness of the human recognition system, many test images in different situations are collected. The possible situations could be roughly separated into three cases: different poses (DP), occlusion by other objects or humans (OC) and complex background (CB). The performances are shown in Table I, where the accuracy rate of overall test images is higher than 90% and the average execution time is about 0.1 sec/frame.

Obviously, with the help of depth image and component-based concept, the system could detect humans correctly even suffering from serious occlusion.

TABLE I. ACCURACY RATES IN DP-, OC-, CB-, TOTAL GROUP, AND AVERAGE EXECUTION TIME

	AR(DP)	AR(OC)	AR(CB)	AR(Total)	Execution Time
Voting	96.15%	93.36%	92.02%	93.74%	0.122s
NN	97.26%	95.03%	95.83%	95.80%	0.131s

4. Conclusions

This paper proposes an intelligent human detection system based on depth information generated by Kinect to find out humans from a sequence of images and resolve occlusion problems. From the experimental results, there are some conclusions listed as below:

- The proposed system could detect human with high accuracy rate even suffering from serious occlusion.
- The use of depth image to implement human detection would have some distinct advantages over conventional techniques. First, it is robust to illumination change and influence of distance. Second, it could deal with occlusion problems efficiently. Third, it is suitable for moving camera because no background modeling is required.
- The use of chamfer matching to achieve significant human features could highly reduce the dimension and size of the neural network. The conventional pattern recognition often directly applies a patch of image or the whole pixels of an ROI into the neural network. Consequently, the neural network requires hundreds and thousands of neurons in its input layer and a whale of training data for training. With pre-processing via chamfer matching, the number of neurons in the input layer could be reduced to less than 50.

ACKNOWLEDGEMENT

This work was supported in part by a grant provided by National Science Council, Taiwan, R.O.C. (NSC 101-2221-E-009-060).

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893 vol. 1.

[2] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Computer Vision - Eccv 2004, Pt 1*. vol. 3021, T. Pajdla and J. Matas, Eds., ed Berlin: Springer-Verlag Berlin, 2004, pp. 69-82.

[3] L. Zhe and L. S. Davis, "Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 604-618, 2010.

[4] D. M. Gavrila, "A Bayesian, Exemplar-Based Approach to Hierarchical Shape Matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 29, pp. 1408-1421, 2007.

[5] X. Lu, C. C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 2011, pp. 15-22.

[6] M. Bertozzi, E. Binelli, A. Broggi, and M. D. Rose, "Stereo Vision-based approaches for Pedestrian Detection," in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 2005, pp. 16-16.

[7] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International Journal of Computer Vision,* vol. 73, pp. 41-59, Jun 2007.

[8] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-Based Pedestrian Detection for Collision-Avoidance Applications," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 10, pp. 380-391, 2009.

[9] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 1, pp. 148-154, 2000.

[10] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Transactions on Information and Systems E Series D,* vol. 87, pp. 113-120, 2004.

[11] G. Borgefors, "Distance transformations in digital images," *Computer vision, graphics, and image processing,* vol. 34, pp. 344-371, 1986.

[12] A. Meijster, J. B. T. M. Roerdink, and W. H. Hesselink, "A general algorithm for computing distance transforms in linear time," *Mathematical Morphology and its applications to image and signal processing,* pp. 331-340, 2002.

[13] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," DTIC Document1977.