Depth-image Based 3D Point Cloud on Mobile Carrier-Based Personnel Tracking System

Chia-Ming Liu, Chun-Yao Lin, Kuan-Wei Zeng, and Yen-Lin Chen*, *IEEE Member*

***Abstract*-- The mobile robots with low computing power are equipped with only depth sensors in a lower-view way, using 3D spatial point cloud information for pre-processing and object cutting, and analyzing the spatial changes of the candidate objects to extract 2D color information for pedestrian recognition, and calculating object similarity and movement through HOG features for tracking to realize a real-time personnel tracking system.**

I. INTRODUCTION

The rise of autonomous mobile robots, in addition to interacting with general users indoors, can quickly respond to a large number of customized needs in the era of Industry 4.0, making automated production lines more flexible and flexible to save manpower. The degree of intelligence has been improved by using the information generated by different imaging modules to create machine vision, so that the service-oriented robot can obtain both eyes and gain more awareness of the environment. Most of the current image processing technology is based on RGB images for implementation[1][2], all aspects have been mature, but RGB images still have their limitations, such as weak characteristics of the wall, the spatial coordinates of the object information, changes in light intensity and other external factors lead to a decline in recognition, three-dimensional image processing technology has flourished, the application of 3D scene model is also increasing, such as indoor service robots[3], unmanned self-driving vehicles, AR space model. The application of 3D scene modeling is increasing, such as indoor service robot, unmanned self-driving car, and AR spatial positioning. Due to the awareness of privacy and information security, there are often some information security risks and threats when using RGB color images as the source of robot visual information in environments that require a higher level of information security or privacy protection, such as factories or homes.

Therefore, based on the above application, in order to strengthen the function of the service-oriented robot, this paper only uses the 3D point cloud data of the depth camera of the depth image information, and implements the personnel tracking system by the 3D information generated from the depth information, and integrates the system with the mobile robot in series and realizes its related applications.

II. RESEARCH METHODS

A. Object cutting

In the 2D image plane, there are often overlapping areas between objects. Therefore, this paper uses Split histogram[4][5] as a segmentation method to find the target cluster area by cutting the wave peaks and troughs. In this thesis, the X-Z plane is used as the input image for histogram statistics, which is converted into a distribution of values. State 1 is a normal plane, which means that the height of the waveform at that position does not change; state 2 is a falling waveform, which means that the position is from a certain high place or plane downward; state 3 is a rising waveform, which means that the position is from a certain low place or plane downward; Fig 1 shows the flow of the state machine.

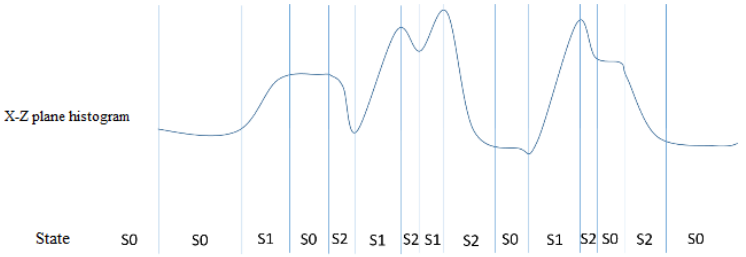


Figure 1.State Machine Flow Chart

In order to solve this problem, we apply the 3D spatial information provided by the depth camera to process the depth image as the input image to cut out the candidate objects at different distances, and the size of the object in the image will be proportional to the angle between the two ends of the object and the camera, and the size will be reduced as the distance increases. Therefore, in this paper, the threshold value is adjusted according to the depth value, and the new reference distance D_i is obtained by (1) with reference to the original depth value d_i of the wave height position, and the dynamic threshold value T_i is obtained by substituting into (2), where t is the original threshold value, and s is the range of the adjustment ratio. Then, the image of x-coordinate is normalized to the positive range, and finally the Split histogram is used to cut out the real candidate object.

$$D_i = \begin{cases} upper_{bound} & ,if\ d_i > upper_{bound} \\ lower_{bound} & ,if\ d_i < lower_{bound} \\ d_i & ,else \end{cases} \quad (1)$$

$$T_i = t * e^{-(s_{max} - s_{min}) * \frac{D_i - lower_{bound}}{upper_{bound} - lower_{bound}}} \quad (2)$$

In this case, we obtain the image of the candidate object segmented by x-plane and z-plane, and generate the information of the cut sub-point cloud according to the position of the image. However, there may still be drift noise in the image, so we use Find Contour to find the largest candidate object, and select the real candidate object according to the range box of the contour, and substitute the internal parameters into (3) and (4), (u, v) is the converted 2D coordinate value, convert the 3D coordinates to obtain the coordinates position in the 2D image plane, and use the candidate box as the candidate cloud selected by the ROI position frame is cut with the restriction.

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = K \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (3)$$

$$u_{2D} = f_x * \frac{X'}{Z'} + c_x$$
$$v_{2D} = f_y * \frac{Y'}{Z'} + c_y \quad (4)$$

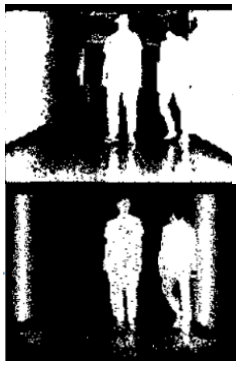


Figure 2. Z-aix schematic

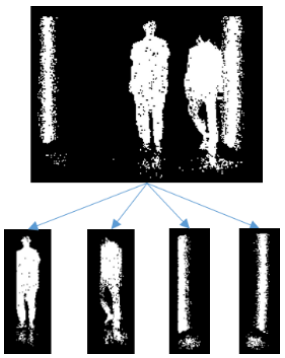
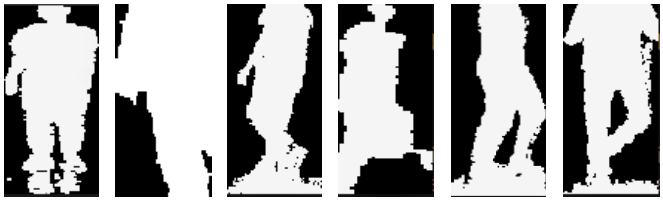


Figure 3. Object cutting result

B. *Personnel Detection*

Traditionally, human detection mostly relies on the complete human body information, such as head, hands and feet, etc., or uses the human body proportions for part estimation, but the application of this thesis often cannot see the complete human body due to the FOV of the depth camera and erection constraints as shown in Fig 4. Therefore, the human scale cannot be predicted by anatomy. In order to ensure the correctness of the detection, the limitation of human scale will be relaxed and the limitation of spatial distribution of point clouds will be added.



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

Figure 4.ROI images of candidate objects (a) full human (b) half human (c-f) walking incomplete human

The similarity between the wall and the person with feet together is very high in the depth image, and several methods are added to filter it out because the wall is actually flat and the variation of the depth distance value of the wall parallel to the camera is very low. Therefore, by using the object cutting method in this paper, the sub-point cloud information is cut out, and the variation of the point cloud is calculated by equation (5), and the variation of the z-coordinate is calculated for the initial filtering.

$$S^2 = \frac{\sum(x_i - \bar{x}^2)}{n-1} \quad (5)$$

Then, the center of gravity of the lower half of the ROI is calculated as shown in the red circle in Fig 5, and the center of gravity is extended down to the bottom of the ROI to cut out the lower half of the estimated target object, dividing the upper half and the lower half, and setting the filling level of the upper and lower half of the ROI within a certain threshold.

(a) (b) (c)

Figure 5.Human ROI cropping diagram (a) Center of gravity diagram (b) Upper body diagram (c) Lower body diagram

Finally, the confidence value calculated by HOG-SVM is used as the human recognition criterion, and this feature is used in the paper [5] for calculation. Therefore, this paper uses ROI images, rescales them to a target image size of 64x128, adds a black border to accommodate the operation of the SVM classifier, converts the color image information into floating point numbers based on the adjusted point cloud information, and performs the calculation to obtain the HOG features. The HOG feature is obtained by applying the pre-trained SVM classifier to the hyperplane mapping of the feature and calculating the confidence value, which is used as a criterion to identify the human object.

C. *Tracking Process*

In this thesis application, users can specify the person they want to track, so the tracking process is added to ensure the location of the target object, but because the movement of the mobile carrier and the person itself will make its characteristics change due to different movements, the personnel detector does not ensure 100% detection of each image of the person, so this thesis adds the tracking process, when the confidence value detected by the target person is low, the target movement changes and related characteristics are calculated for tracking to ensure that the correct detection of the person on the screen to achieve the tracking effect.

According to the confidence value calculated in Section II.B as the threshold for initializing the tracker, the algorithm is shown in Alg 1, and all the trackers are visited based on the candidate detection objects. Find the closest tracker and record the corresponding detection object and tracker IDs and the distance between them to build TDIdMap (Tracker Detection ID Map) as shown in Alg 1, the TDIdMap will record the combination of multiple detection objects and trackers found in the first visit that should be updated, and then delete the already updated If there are still detection objects in the distance range that have not been updated after the deletion, the TDIdMap will iterate repeatedly until the detection objects have been updated or there are no more trackers in the distance range, and then the TDIdMap will stop updating.

Algorithm 1 UpdateTracker(Detection)
TDIdMap: Tracker ID, map<Detection ID, distance>
uIDs: updated detection ID list

```
while Detection size > 0 and TDIdMap size > 0 do
    TDIdMap ← createTDIdMap(Detection)
    uIDs ← Ø
    for i ∈ TDIdMap do
        if checkHeightDiff(i) AND checkSimilarity(i) THEN
            updateTracker(i)
            uIDs ← uIDs ∪ DetectionID
        end if
    end for
    removeUpdatedDetection(uIDs)
end while
```

Through the distance to select the candidate to update the tracker object, but the target object movement process, there may be obscured, crossing and other conditions occur, so the closest distance tracker and the candidate detection object is not necessarily the same object. In this paper, we use the height restriction to make judgment, if we use the height of the object alone to make judgment, the overall Y-axis position of the object may be shifted upward. Therefore, we calculate the candidate object $\vec{h}$ as a new vector in (6), and calculate the European distance between them, taking into account the maximum, center and minimum values of the Y-axis of the candidate object to ensure their correlation.

$$\vec{h} = (y_{max}, y_{center}, y_{min}) \tag{6}$$

Then, the HOG feature descriptors of the human object information calculated in Section II.B are used to calculate the cosine similarity[6] between them. Since cosine similarity takes into account the spatial distribution information of the overall appearance, it is more useful for recovering tracking information of objects that have been obscured for a long time, and the above three information are used as the criteria for updating the tracker.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{7}$$

After the existing trackers are updated, the remaining unused objects are checked for confidence and height, and the objects that meet the requirements are added to the tracking target list as new trackers. Finally, since the Detector may split the same target object several times, which may cause the same object to be added twice, the distance judgment is added to remove the tracers that are too close and too similar to complete the overall tracking process.

III. Experimental Results and Analysis

A. Experimental Results and Analysis

In this section, we introduce the results and related data of the personnel tracking system in this thesis, and evaluate the validation method by binary confusion matrix, as shown in Table I, and add (8)、(9) and (10), Accuracy, Precision, and Recall to evaluate the system method in this thesis. Since the application context of this thesis is personnel tracking, there is no contextual item of TN defined, therefore, TN is removed from (8) and not counted.

TABLE I. Confusion Matrix Event Definition

Actual / Prediction	Target object presence	Object not present
Target object detected	TP (True Positive)	FP (False Positive)
Target object not detected	FN (False Negative)	TN (True Negative)

$$Accuracy = \frac{TP}{TP + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

In the application of this thesis, ideally only one person is on the screen for tracking, but in the process of tracking there may be other people entering or the tracking person is obscured. In the mobile carrier set up the camera, in addition to the movement of the personnel themselves, the carrier will also move, compared with the fixed camera set up environment, may cause the carrier to move, the camera is also receiving information, may cause some changes in the location of the transmitting and receiving, and then the problem of data floating. In order to prevent the carrier's speed or rotation speed is too fast, acceleration is too large and so on, the robot's movement will also add the speed smoothing limit, the carrier from stationary to the time of movement need to be smooth acceleration, the response time is longer, in this process is to receive the displacement information to inform the motor, drive the axle displacement, so it will also cause a small amount of transmission delay, even if the carrier does not consider the problem of moving too fast, but still will reduce the size of the speed limit on the movement of personnel.

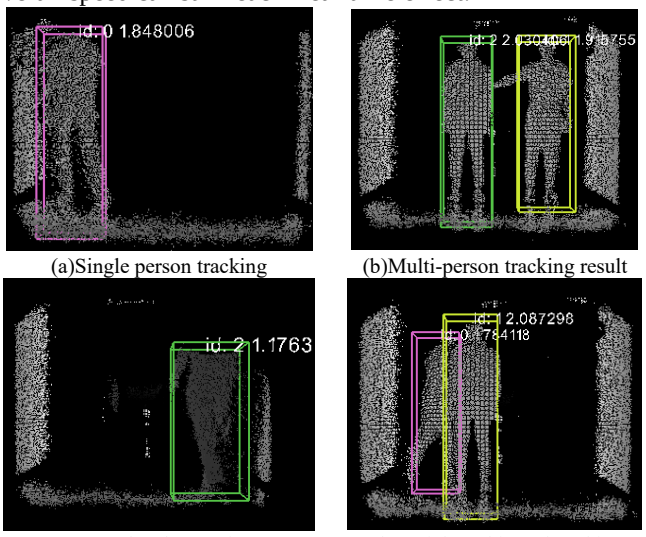
The person setting the target in this thesis will move from 2.5 m to 5.5 m from the mobile carrier. In this thesis, Ground truth is marked by manual judgment, and the criteria are that the person is marked as a person with complete foot features and more than half of the body, satisfying the aforementioned conditions as shown in Fig 6 (a). To ensure the correctness of object tracking, the distance between human and mobile carrier has to be increased to 2.6 meters before the object is officially added to the tracking range, and the continuous tracking range is set at 2.5 meters to 5.5 meters.

TABLE II. Personnel detection data

Situation Name	Number of people	TP	FP	FN	Precision	Recall	Accuracy
Static Single Person	1390	1389	1	1	99.93%	99.93%	99.86%
Static double People with cover	776	776	3	0	100.00%	100%	99.61%
Dynamic Single Person 1	410	400	3	10	99.26%	97.56%	96.85%
Dynamic Single Person 2	453	435	24	18	98.69%	96.03%	91.19%
Dynamic multi People with masking	390	384	0	6	100.00%	98.46%	98.46%
Total	3419	3384	31	35	99.09%	98.98%	98.09%

In this thesis, in order to ensure that the target object is a

person, the human recognition standard is raised, and a higher confidence value must be obtained before tracking can begin, and the confidence threshold is relaxed after tracking begins. When a person is obscured by other objects or objects in an application situation as shown in Fig 6 (a~d), and the duration is long or the distance after obscuring is too far, the object needs to meet a higher threshold before tracking can be restarted, thus increasing the number of cases where the person is on the screen but not being tracked. Since the initialized tracker uses the confidence value as the threshold, the distinction between human and non-human objects is stabilized and the number of False Positive cases is reduced more precisely. The tracking part compensates for the problem that the detector cannot detect the target object completely and correctly, improving the overall Accuracy, the average result can reach 98.09% and the overall speed can still reach Real-time effect.



(a)Single person tracking (b)Multi-person tracking result

(c)Incomplete human image tracking (d)Partial masking of tracking objects

Figure 6. Tracking results in different contexts

IV. CONCLUSION

Due to the hardware limitation of CPU only computing platform and the limitation of using depth image as the basis, it is not possible to use the mainstream fusion RGB image to segment the object in 2D color space and use some color features for recognition, and then combine it with 3D information. In addition, the narrow vertical field of view of the camera increases the difficulty of recognition, and the data shows that the accuracy of the proposed method can be maintained at about 97% on average in different contexts.

ACKNOWLEDGMENT

The study is supported by Ministry of Science and Technology (MOST) to National Taipei University of Technology under MOST 109-2628-E-027 -004 -MY3 and MOST 109-2622-E-027-034.

REFERENCES

[1] M. Munaro and E. Menegatti, "Fast RGB-D people tracking for service robots." *Autonomous Robots*, 2014.

[2] Z. Deng and L. Longin Jan, "Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[3] Linder, T., Breuers, S., Leibe, B., Arras, K.O., "On Multi-Modal People Tracking from Mobile Platforms in Very Crowded and Dynamic Environments", IEEE International Conference on Robotics and Automation (ICRA) 2016

[4] W. L. Shin, " Real-time Human Skeleton Recognition and Fitting System Based on Low Resolution Depth Data", Master's Thesis, National Taipei University of Technology，Taipei，2020，https://hdl.handle.net/11296/evvvn2。

[5] H. Jack, "Drawing the Head and Figure: A How-To Handbook That Makes Drawing Easy, " *TarcherPerigee: USA*, 1983, pp.39.

[6] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters 9.3* (1999): 293-300.