Carried Objects Recognition from Pedestrians' Range Images

Akinobu Watanabe

Hitachi, Ltd.
Keywords: TOF, Object, Recognition

ABSTRACT

We developed the carried objects estimation technique from range image captured by TOF sensor, and confirmed the accuracy of single carried object's existence is about 98% and loss is 7% by the combination of the recognition method with CNN model trained by histogram images of point cloud and the rule-based pre-detection method.

1 INTRODUCTION

In the analyses of the operation of workers in the manufacturing premise and the customer action in the retail store, the inflection of provided range image data by TOF (Time of Flight) sensor is expected. Many techniques are suggested as a person posture estimate using range image data, but, as for the use case in the point of view looking down, examination does not advance enough.

1.1 Background

A sensor apparatus and an IT system have been developed and become low price. An IoT (Internet of Things) market using Information collected with them from a machine, a vehicle and a building, to use the information for analysis and control is spreading.

With progress of IoT, there is the movement that is going to feed back the result of future prediction by collecting the data of a machine and the facilities which are on-site physically, and reappearing as “digital twin” within the cyber world of the IT system, using an information processing technology. Siemens and GE stimulate research and development in conjunction with the digital twin, too and have begun to already send information in a general medium [1][2].

Furthermore, sensing object is being extended to "a Human being" from "a Thing". The detection, the reduction of the improvement, efficiency and work error that I included the movement of the person in is enabled by reproducing the spot that the Homo sapiens included as digital twin. In order to realize it, it is necessary to detect existence and the movement of the person, and to process it to convert it a fixed form as digital information. For example, OpenPose can digitalize human posture with 2d images [3]. However general 2D images of human being should be treated privacy sensitive information in public area, such as airports and stations, excluding security purpose.

1.2 Purpose

It was aimed for the establishment of the carried objects detection technology suitable for an available privacy sensitive use with the general-purpose 3D sensor including the TOF sensor in order to solve the restriction of RGB camera raised on privacy issues.

1.3 Target

In this study, we intend for processing to distinguish existence of carried objects with human being from the range image of pedestrians.

2 EXPERIMENT

In this study, we captured range images of pedestrians with and without carried object, removed background point cloud, tracked the position of the pedestrian and predict the existence of carried object around the pedestrian.

2.1 Previous method

As shown in the prior publication, we improved human tracking algorithms’ issue caused by undetected human pillages an ID of an already been detected with TOF sensor [4].

2.2 Issue of functionality

The previous method can only detect human position. As the next step, we need to detect more detailed human information, such as attributes, action and appearance.

On the other hand, methods such as PointNet are known to recognize objects from 3D point clouds. These techniques recognize objects directly from a 3D point cloud. Recognition accuracy is excellent, but processing cost is high because the number of point cloud is large.

In this study, we decided to consider a method that can process real-time at a lower processing cost because it was intended to be applied to a system with a cheaper and simpler configuration.

Then we decided to develop detection function carrying or wearing something or not with 3D point cloud in realtime on general environment.

2.3 Condition

According to the previous study[4], we are possible to obtain the target coordinates of the 3D point cloud to be recognized in real time by the method of detecting the position of the pedestrian.

In addition, as a method of separating the background 3D point cloud and the foreground 3D point cloud, a technique called motion extraction is known.

The 3D point cloud of the foreground, which is present in the vicinity of the target coordinates, is used as the source of the input data.

The objects to be detected are a suitcase and tool case of about the same size. The former has four wheels, the latter has two wheels. Therefore, the former includes the case of moving in a state where the four wheels are grounded, and the case of moving in a state of grounding only two wheels.

People carrying luggage are of children (short height) and adults (tall).

The directions of walking with carrying luggage are of three directions, the direction approaching the ranging sensor, and the direction of moving away, and in the direction of crossing. And the person carried out the following 2 kinds of movements, the movement of walking without luggage and walking for transporting luggage.

As the installation location of the ranging sensor was at a height of 2.3 meters, and at an angle raised 65 degrees from the vertical axis downward.

Table 1 shows the experimental environment.

Table 1 Experimental environment

Carried Object	Person's Height[m]	View	Sensor Height[m]	Sensor Angle
Suitcase Toolcase None	1.2 1.4 1.7 1.7	Front Side Back	2.3	X=65°

2.4 Normalization

In order to determine whether or not carrying luggage, information of excessively detailed 3D shape is not required. Further, the target, if it is in a distant position from the ranging sensor, the density of the point cloud becomes coarse, then it is not possible to obtain a detailed 3D shape. This problem occurs when obtaining a 3D point cloud using a ranging sensor such as a TOF sensor or LiDAR.

Therefore, by generating the representative information of a plurality of point cloud present in the vicinity in the spatial coordinates, we are able to reduce the amount of data.

As a result, compared with the case of using the 3D point cloud as it is, since the amount of data to be processed is small, it was considered that high-speed processing can be expected. Furthermore, because the target present in a distant position from the ranging sensor can be handled by the same amount of information as of the target present in a close position, it was assumed that it is possible to recognize independent of the distance of the target.

2.5 Visualization

As a method for generating representative information from the target 3D point cloud, we proposed a method to visualize the distribution of the point cloud as a two-dimensional histogram mapping to the vertical axis Z and radius axis R, based on the position of the person obtained as the target coordinates.

The 3D information of the person carrying the luggage are generated as an image of a two-dimensional histogram, which has a vertical axis Z and the radial axis R of the cylindrical coordinate system around the vertical axis Z, and mapping the number of point cloud present in the Z and R coordinates. This image has no dependence on the rotation around the vertical axis Z, and the person does not depend on whether walking in either direction, then it can be expected to be robust with regard to the traveling direction.

As a method of comparison, we evaluated a method using a two-dimensional image projected on the X-Y plane, which is a water plane, with respect to the 3D point cloud centered on the coordinates of the person, the 3D point cloud colored according to the z-coordinate value.

These images projected on the X-Y plane are different images according to the position where the luggage is present around the person by the walking direction. Then we rotated these images in 30 degree increments, by generating an image of 12 times the number, aiming to eliminate variations of the recognition accuracy due to walking direction.

2.6 Colorization

As the Z-R Histogram, we generated the image which was colored in heatmap scale, and the image colored in greyscale, response to the number of point cloud. This comparison is to confirm that there is no difference in recognition accuracy in the case of representing the number of point cloud in three channels of RGB and the number of points in greyscale of one channel.

Similarly, in the case of projection to X-Y Height, we created the image which was colored in greyscale coloration in response to the z-coordinate value, as compared to the colored image in heatmap scale, to be confirmed if there is any difference in recognition.

If there is no difference in recognition accuracy due to these coloring, it can be expected that the amount of greyscale data is small because the number of channels is small, and is faster in recognition execution.

2.7 Modeling

We adopted a model using CNN's net configuration, which provides a relatively good score in MNIST character recognition, because it has same features of the MNIST character data set as the greyscale and low resolution two-dimensional images.

The final output, to determine the presence or absence of carrying luggage, is two-class classification. And we used TensorFlow[5] as the framework for deep learning.

Figure 1 shows the net configuration.

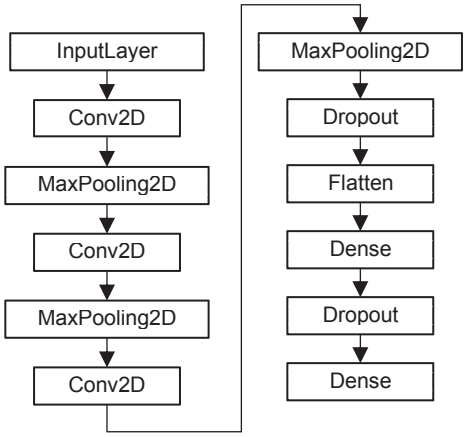

Figure 1 Net configuration

2.8 Pre-Detection

At the stage of annotation to the Z-R Histogram image, because it was blocked by the person transporting luggage, we found that there are some frames where luggage is not observed by the ranging sensor.

If we annotate these frames as “with luggage”, it is assumed that the inhibitor of recognition accuracy. Then we compared the case of adding a pre-processing to select whether the luggage is observed sufficiently and the case of not performing pre-processing.

This pre-processing is the following. If the ratio of the number of point cloud number present in the rectangular area is part of the Z-R Histogram image is more than a threshold value, the luggage is considered to be existing, otherwise, regarded as not existing. By performing pre-processing, the frames the ratio is more than the threshold value are labeled as “with luggage”, and other frames labeled as “ no luggage”.

2.9 Dataset

The data set created for this experiment is shown in Table 2.

Table 2 Dataset for carried baggage recognition

#	Pre-Detection	Visualization	Color	Normalization
1	Yes	Z-R Histogram	Color	Yes
2	↑	↑	↑	No
3	↑	↑	Grey	Yes
4	↑	↑	↑	No
5	↑	X-Y Height	Color	Yes
6	↑	↑	↑	No
7	↑	↑	Grey	Yes
8	↑	↑	↑	No
9	No	Z-R Histogram	Color	Yes
10	↑	↑	↑	No
11	↑	↑	Grey	Yes
12	↑	↑	↑	No
13	↑	X-Y Height	Color	Yes
14	↑	↑	↑	No
15	↑	↑	Grey	Yes
16	↑	↑	↑	No

Figure 2 shows sample images for each dataset variant.

#1 #2 #3 #4 #5 #6 #7 #8

Figure 2 Image Examples

3 RESULTS

Figure 3 shows the results of recognition accuracy and loss for 16 conditions combination. It shows the average of 10 attempts.

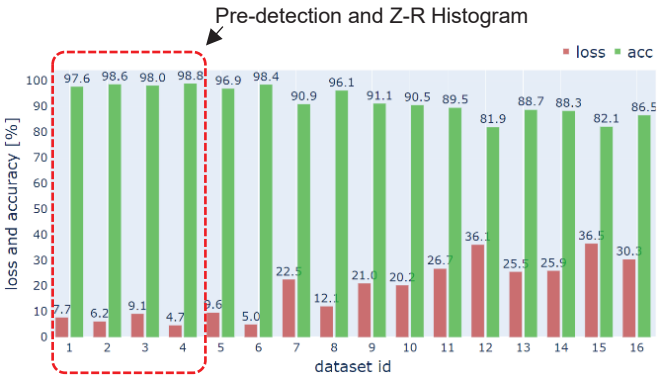

Figure 3 Result of recognition

The most accurate group is Pre-selected and Z-R Histogram, whose id is from #9 to #12. Its accuracy is 98.3% and its loss is 6.9% as its groups’ average. The best condition is the combination Pre-selected and Z-R Histogram.

Table 3, depending on the presence or absence of pre-treatment, shows a comparison of accuracy and loss.

Table 3 Pre-detection and Visualization

	Z-R Histogram		X-Y Height		Max. diff.
	Pre-detection	All-data	Pre-detection	All-data	
Accuracy	98.3	88.2	95.6	86.4	11.9
Loss	6.9	26.0	12.3	29.6	-22.7

In Table. 3, we confirmed that the combination of the Z-R Histogram and the pre-detection improved accuracy by 11.9% and loss by 22.7%.

The following shows the difference between pre-detection presence or absence, visualization method, coloring presence or absence, voxel presence or absence, and normalization presence or absence, in average score.

3.1 Pre-detection

Table 4, depending on the presence or absence of pre-treatment, shows a comparison of accuracy and loss.

Table 4 Pre-detection

	Pre-detection	All-data	Diff.
Accuracy	96.9	87.3	9.6
Loss	9.6	27.8	-18.2

In Table. 4, we confirmed that the pre-detection of images by rule-based method improved accuracy by 9.6% and loss by 18.2%.

3.2 Z-R Histogram and X-Y Height

Table 5 shows the comparison of Z-R Histogram and the X-Y Height projection of accuracy and loss.

Table 5 Visualization

	Z-R Histogram	X-Y Height	Diff.
Accuracy	93.3	91.0	2.3
Loss	16.5	20.9	-4.4

In Table. 5, we confirmed that the Z-R Histogram has been improved in accuracy by 2.3% and in loss by 4.4%.

3.3 Color and Greyscale

Table 6 shows a comparison of colored, greyscale, accuracy and loss.

Table 6 Colorization

	Color	Greyscale	Diff.
Accuracy	93.8	90.5	3.3
Loss	15.1	22.2	-7.1

In Table. 6, we confirmed that the Colored data has been improved in accuracy by 3.3% and in loss by 7.1%.

3.4 Pixel and Voxel

Table 7 shows a point cloud, voxels, a comparison of accuracy and loss.

Table 7 Normalization

	Pixel	Voxel	Diff.
Accuracy	92.4	91.8	0.6
Loss	17.6	19.8	-2.2

In Table. 7, we confirmed that the Pixel data has been improved in accuracy by 0.6% and in loss by 2.2%.

4 DISCUSSION

4.1 Discussion

To be obtained from Table 3 to Table 7, the factors whose influence on both of the recognition accuracy and loss is large is to use the Pre-detection, and especially the combination of Pre-detection and Z-R Histogram is considered as the best condition.

On the other hand, the visualization and the normalization is considered that they have not so large difference sololy.

5 CONCLUSIONS

The carried objects estimation technique from range image captured by TOF sensor achieved the accuracy of single carried object's existence is 98.3% and loss is 6.9%.

REFERENCES

[1] Digitalization in machine building - The digital twin (http://www.siemens.com/customer-magazine/en/home/industry/digitalization-in-machine-building/the-digital-twin.html)

[2] 'Digital Twin' Technology Changed Formula 1 and Online Ads. Planes, Trains and Power Are Next (http://www.gereports.com/digital-twin-technology-changed-formula-1-online-ads-planes-trains-power-next/)

[3] OpenPose (https://github.com/CMU-Perceptual-Computing-Lab/openpose)

[4] A. Watanabe, et al. 'Vertical View Human Action Recognition from Range Images', IDW '19

[5] TensorFlow (https://github.com/tensorflow/tensorflow)