Vertical View Human Action Recognition from Range Images

Akinobu Watanabe, Keiichi Mitani

Hitachi, Ltd.

Keywords: TOF, Posture, Tracking

ABSTRACT

We developed the human joints' position estimation technique and the person tracking technique from upward view range image of TOF sensor, and confirmed the correct prediction ratio of hands' position is 97%, and confirmed the person tracking error is reduced to 1/7.

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. It is predicted that the market size of IoT grows up to approximately 1,300 billion dollars from approximately 700 billion dollars of 2015 in 2019 [1].

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 [2][3].

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. Microsoft announced Kinect for a game in the consumer market [4]. Kinect can not only capture the movement of the player as an animation, but also acquire three dimensions of joint positions of the subject as coordinate data in the space. In addition, there is movement to apply for industry because available SDK is provided on a PC [5]. For example, Kinect is used for the study of the system for fields of industry to detect the deviation action of the shop floor worker [6].

On the occasion of the use of Kinect, one PC equipped with GPU is necessary for one Kinect sensor. Furthermore, it is assumed the viewpoint of Kinect is the front view. Then in the production line floor, it is hard to keep a field of vision because of shielding, and it is an issue that there is much limitation for sensor setting.

1.2 Purpose

It was aimed for the establishment of the posture detection technology suitable for an available industrial use with the general-purpose 3D sensor including the TOF sensor in order to solve the problem mentioned above.

1.3 Target

In this study, we intend for processing to extract a joint position (skeleton) of the human body from the range image which photographed the human body. Particularly, high precision of the technique that can extract a skeleton with the viewpoint looked down from upper position at intends for suggestion of making it and an evaluation of the precision is our target.

2 EXPERIMENT

As shown in the prior publication, we developed vertical view human skeleton recognition method [7]. And we improved a correct rate of sequential joint searching method, and developed human model selection method to choose the most appropriate model from various models for various target bodies. In this study, we improve the accuracy of detection ratio of human joints and person position.

2.1 Previous method

Previous method is first to detect the head as the starting point and search the joints from a shoulder to a hand in a human body model sequentially [7].

Regarding people tracking, there are many open SDKs which have various algorithms that detect the person position from 2d images or 3d range images, point cloud. In which, a commonly procedure of that person detections is to classify person as move body after clustering point cloud., e.g. [8].

2.2 Issue of correct ratio

The correct ratio of human joints position by the previous method is shown in Table 1., which is reported in [7].

In previous method, average correct ratio for detected frames is 93 [%]. Hands' correct ratio for detected frames are 88[%] and 83[%].

Table 1 Correct Ratio

[%]	Shoulder		Elbow		Wrist		Hand	
	R	L	R	L	R	L	R	L
Detected	99	88	95	97	94	98	88	83

Then we decided the issue is the correct ratio of the hands.

2.3 Issue of people tracking

The current algorithms have several issues: (1) coalescence with other clusters, (2) occlusion caused by a person hiding behind the others, (3) undetected human pillages an ID of an already been detected and so on. In this work, we count the number of people tracking errors related to (3), shown in Fig. 14.

Developed method

We developed the hybrid approach using (1) image creation from IR image and depth data, (2) image recognition method and (3) 3D position estimation method. The flow of this approach is shown in Fig. 1.

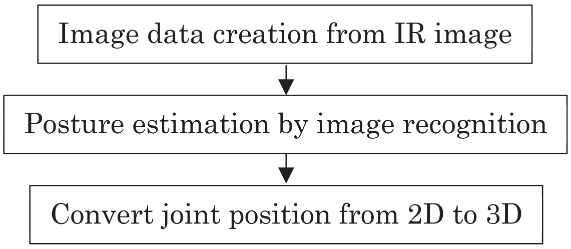
Image data creation from IR image
Posture estimation by image recognition
Convert joint position from 2D to 3D

Fig. 1 Hand position detection flow

2.4 Image creation

From Fig. 2, to 6 are the examples of evalution images of this work.

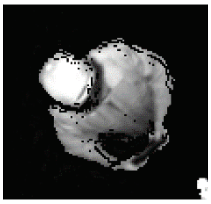

Fig. 2 2D IR

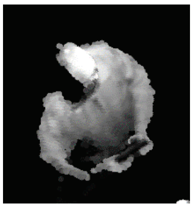

Fig. 3 Large pixel

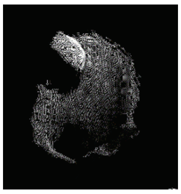

Fig. 4 Mesh

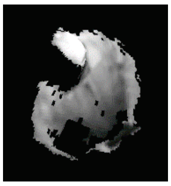

Fig. 5 Polygon

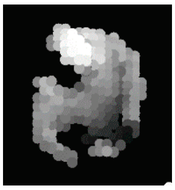

Fig. 6 Voxel

In order to improve detection ratio, we assumed the displaying method is important. Then we compared such cases as (1) the size of pixel is enlarged, (2) adding the line next to each other (to say, "mesh" or "wire frame"), (3) painting the triangle plane of 3 neibering pixels as 3 vertexs of it (to say, "polygon") and (4) voxel spheres instead of pixels.

IR images may include unnecessary information, for example, background. Then we create evaluation image not including other information than human appearance.

3D range image can be projected to any plane of any angle. Then we created 3D range images of X axis rotation chopping its rotation angle from top view to front view.

Image recognition method

We applied OSS to recognize human body joints from IR images and estimate joint position in 2D coordinates on IR images.

2.5 3D position estimation method

When we got 2D position on IR images, then it can be converted to 3D position in world coordinates.

2.6 People tracking method

We developed cluster-to-person classification method aiming to solve the problem (3) of subsection 2.3 in this work.

3 RESULTS

We developed techniques that detect human actions, such as posture and position, from 3D range image of TOF sensor. And we confirmed the improvement.

3.1 Reference data of joint position

Table 3 shows reference evaluation data to measure detection ratio and correct ratio in this study. This data is same as the one in previous study.

Table 2 Reference data

Model	Behavior	frame	Total Joint
1	A	231	1777
2	B	61	431
3	C	355	2284
4		321	2684
5	D	97	766
6		123	956
Total	-	1188	8988

3.2 Joint detection ratio

By applying OSS joint detection software, we got joint position in 2D coordinates on evaluation image.

The detection ratio of 2d IR image is shown in Fig. 7.

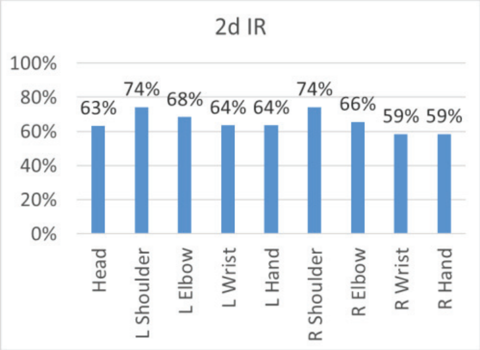

Fig. 7 Detection ratio

In this case, hands' detection ratios are 64% and 59%.

The correct ratio of 2d IR image is shown in Fig. 8 .

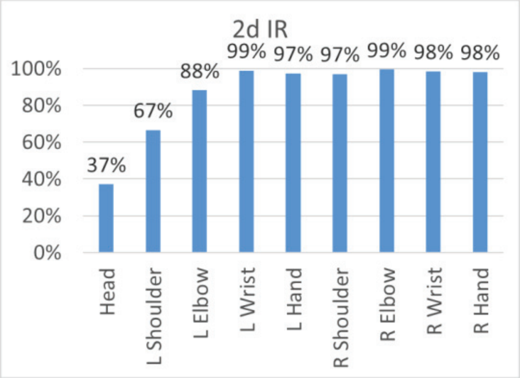

Fig. 8 Correct ratio

In this case, hands' detection ratios are 97% and 98%.

The detection ratio of 3d rotated image are shown in from Fig. 9 to Fig. 12.

Fig. 9 shows large pixel case. In this case, the 9 degrees rotaion marks the best result. And it is better than simple 2d IR image case.

Fig. 10 shows mesh case and Fig. 11 shows polygon case. In these cases, the detection ratio is lower than 2D IR image case.

Fig. 12 shows voxel sphere case. In this case, the detection ratio is almost zero in every rotation degree.

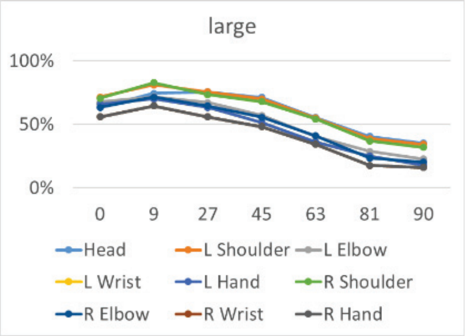

Fig. 9 Large pixel

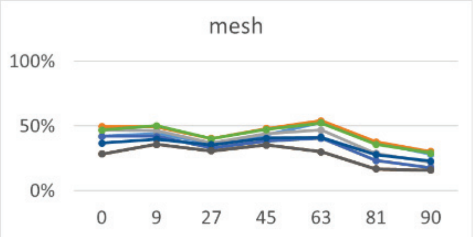

Fig. 10 Mesh

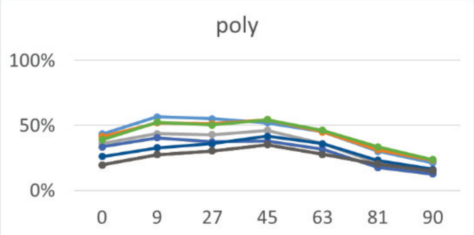

Fig. 11 Polygon

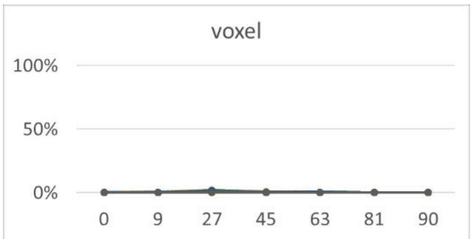

Fig. 12 Voxel sphere

Comparing among these displaying methods, joint detection ratio improvement from 2d IR of both left and right hands is shown in Fig. 13.

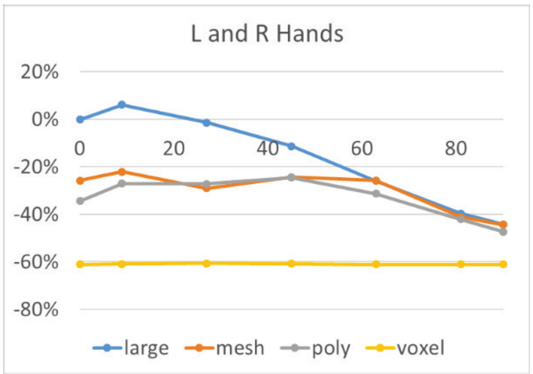

Fig. 13 Hands detection improvement

3.3 Person tracking ratio

We performed the people tracking method of this work to evaluate the improvement under the situations that cause the error undetected human pillages an ID of an already been detected; the results of the above are shown in Fig. 14. The x-axis means the scene pattern and y-axis means number of people tracking error.

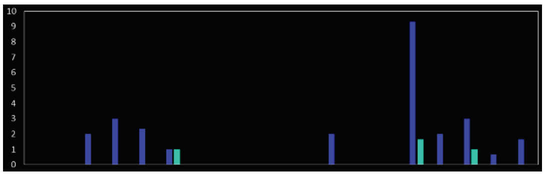

Fig. 14 Person tracking improvement

Each blue and green bar indicates previous and this work.

4 DISCUSSION

4.1 Joint detection

In Fig. 13, we confirmed that the 3D rotation method have the possibility to improve hands' detection ratio. We will evaluate other rotation variations and displaying methods.

IR images may include personal information, for example, face. We must avoid to use such image in privacy sensitive situation. Then we will try to create evaluation image from 3D depth data only.

4.2 Person tracking

In Fig. 14, the number of errors at the previous (blue bar) reduced to 1/7 at the error undetected human pillages an ID of an already been detected.

5 CONCLUSIONS

Hand position estimation is confirmed the correct prediction ratio of hands is 97%, and people tracking is improved that misdetection count is reduced to 1/7.

REFERENCES

[1] Complete bibliographic information (names of “all” authors, “fully descriptive article titles”) in standard format such as AIP or IEEE style, for all cited references is required. See the examples below.

[2] S. Iijima, “Toward Industrial Application of Carbon Nanotube,” Proc. IDW ’03, pp. 3-4 (2003).

[3] K. Maeda and Y. Nakao, “Mechanical properties and fracture analysis of glass substrate for PDPs,” J. SID, Vol. 11, No. 3, pp. 481-484 (2003).

[4] Kinect (https://developer.microsoft.com/ja-jp/windows/kinect/hardware)

[5] Kinect for Windows SDK (https://msdn.microsoft.com/ja-jp/library/dn799271.aspx)

[6] Hitachi and Daicel Develop Image Analysis System to Detect Signs of Facilities Failures and Deviations in Front-line Worker Activities (http://www.hitachi.com/New/cnews/month/2016/07/160713.html)

[7] A. Watanabe, et al. ‘Optimization of Vertical View Human Skeleton Recognition from Range Images’, Proceedings of IDW ‘18, pp. 1182 (PRJ5-3) (2018/12).

[8] M. Munaro, F. Basso and E. Menegatti, “Tracking people within groups with RGB-D data”, IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems, 2101-2107 (2012).