Pedestrian Detection using a LRF and a Small Omni-view Camera for Outdoor Personal Mobility Robot

Manabu SAITO, Kimitoshi YAMAZAKI, Naotaka HATAO, Ryo HANAI, Kei OKADA and Masayuki INABA

Abstract— **This paper describes a pedestrian detection method using a LRF and a small omni-view camera. In outdoor environment, the resolutions of LRFs are too low to recognize human reliably, and high resolution image requires high calculation cost for detecting walking persons. We propose a combination approach using these data. Particle filter based tracking and HOG (Histogram of Oriented Gradients) feature based identification enables to detect pedestrians with high reliability and effectiveness. Although the pedestrian identification from omni-view image needs computational effort for searching large area, LRF based tracking provides the identification process with limited interest regions in advance. We also reports how to construct a discriminative function which is able to cope with various resolution images. Finally, the proposed method was combined with gesture recognition and other robot system, and then an application to taxi service is introduced. The robot could find a person standing 15 meter away from the robot, recognized his swinging hand, and moved to the front of him.**

I. Introduction

Personal mobility is a robot for one person with autonomous navigation function. The robot is able to actively interact with or offer services to people if it has the ability to recognize the environment and humans around it. However, detecting humans in outdoor environments is not easy because finding from low-resolution images of distant pedestrians is often required. This paper proposes a robust pedestrian detection method suitable for personal mobility using a laser range finder(LRF) and a small omni-view camera.

Our method confines the areas of possible pedestrian to small regions of images using LRF and then apply image-based pedestrian detection process to the regions. For the former process, mixture particle filter is applied. By using probabilistic framework, multiple targets can be tracked even in the condition that the sum number of targets are changed in online. On the other hand, pedestrian identification is performed based on HOG (Histogram of Oriented Gradients) features. This allows to find pedestrians captured in pretty low resolution images. One of the advantages of this approach is that it reduces false detections, even when the resolution of images acquired by the robot moving around outdoor environment is fairly low.

HOG based pedestrian detection needs a discriminant function which should be generated from a training dataset by means of machine learning. That is, the detection performance depends almost entirely on the dataset. We investigate how to select dataset from plenty of images captured at outdoor environments. The result shows that a training dataset including various resolution images is available to generate useful discriminant function.

This paper also mentions the realization of taxi-like robot system as a concrete application using the human detection method described above. The system includes finding of passengers and recognition of their hand-waving gesture while the robot is moving.

This paper is organized as follows: Section II describes related works, and section III explains our approach. Section VI, V and VI explain the detail of our method and introduces the evaluation about dataset selection. Section VI describes experimental results, and section VII concludes this paper.

Department of Mechano–Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan {saito, yamazaki, hatao, hanai, k-okada, inaba}@jsk.t.u-tokyo.ac.jp

II. Related works

The purpose of the research is to detect humans as far as 10 to 20[m] distant, from a wide field of view, and there have been a lot of methods related to the human detection proposed.

Some approaches use contours of the whole bodies of humans, and apply techniques in the area of general object recognition. For example, [1] use Edgelet feature. They recognize humans by finding characteristic shapes that appear in human contours. Several works have been reporting that HOG feature [2] is effective for human detection. However, detecting specific objects from the whole images of scenes using only HOG feature requires a lot of computation.

Meanwhile, sensors other than ordinary cameras are also used for human detection. For example, a method of [3] are based on images captured with an infrared camera. [4] uses several LRFs set up at different heights.

Hao et al.[5] propose a method of calibrating a LRF with a camera and detect humans with the both sensors. They clip regions of interest using the LRF, and then determine whether each region is a human or a column by analyzing the curvatures of the edges. This method can be implemented as a compact system and suffer from less false positives, but has not been extended to objects other than columns.

III. Pedestrians detection using a LRF and a camera

A. Issues and approach

In pedestrians detection with wide field of view, a combination of a LRF (Laser Range Finder) with a omni-view camera has an important role. Because, a LRF is able to be used to crop interest regions which are regarded as candidates of pedestrians with accurate distance information. On the

Fig. 1. Human size in the panoramic image

other hand, omnidirectional images captured from a camera is useful to determine whether or not the cropped candidates is truly a pedestrian.

In this research, we use an omni-view camera NM33 conducted by Opt inc.. An image captured by this camera is shown in Fig.1, the width is 1280 pixels while covering 360 degrees of view. Although the advantage of using this camera is its compactness suitable for a mobile robot, the eyesight realized by the image resolution is much lower than that of human. For instance, red circles plotted in Fig.1 shows pedestrian candidates which are detected by using a LRF. These pedestrian only covers 10 pixels width in the image.

However, recent techniques proposed in the field of computer vision make it possible to detect pedestrians in spite of a severe situation described above. We take an approach to pedestrian detection based on HOG (Histogram of Oriented Gradient) features [2].

The proposed detection procedure has three steps as follows: (1) several interest regions are selected and tracked as pedestrian candidates. Next, (2) the regions are superimposed into images, and HOG features are calculated only about the regions on each frame. The size of an image area which is subject to feature calculation is decided from a distance between the camera and the candidates because the distance is known by using LRF. (3) After investigating the plausibility of the features by using discriminant function, the tracking results are improved.

The discriminant function is constructed by means of machine learning in advance. We apply SVM (support vector machine [6]) for this process with thousands of HOG features. Following subsections describe the steps (1) to (3) in order.

IV. Pedestrian Tracking by using LRF

Pedestrian candidates detection and tracking are achieved based on range data captured from LRF. The procedure is inspired from Nune'z approach[7]. First of all, scanned points are divided into some clusters. Segments which lengths are not more than 1000[mm] but not less than 200[mm] are selected as pedestrian candidates. In this process, some candidates extracted from background region such as buildings are removed because their segments is assumed to be occluded from other objects.

For tracking process, Mixture Particle Filter proposed by Vermmak et. al. [8] is applied. This probabilistic approach enables to track multiple targets even in the condition that the sum number of targets are changed in online. The equations is as follows:

$$Predict : p(x_t|y^{t-1}) = \int D(x_t|x_{t-1})p(x_{t-1}|y^{t-1})dx_{t-1}, \tag{1}$$

$$Update : p(x_t|y^t) = \frac{p(y_t|x_t)p(x_t|y^{t-1})}{\int p(y_t|s_t)p(s_t|y^{t-1})ds_t}, \tag{2}$$

where x_t indicates a state vector d at time t, $y^t = \{y_1, ..., y_t\}$ is an observation result since time t. A probabilistic function of the state vector x_t is represented as $D(x_t|x_{t-1})$.

Fig.2 shows an example of pedestrians tracking. Left column indicate some image frames. The center is set to direction of movement, and a field angle of the image is 135 degrees wide which is same as the measurement area of LRF. Right column shows tracking results. Green segments indicate tracking targets, white points show distributions of particles, and white lines show trajectories of tracked segments. A position of the LRF is shown by a white circle at the center of each image. According to the number of image frames increase, several pedestrians could be tracked at the same time. On the other hand, it sometimes happened that a tree and a part of a car could also be regarded to tracking targets.

V. Image cropping and HOG based Identification

Although pedestrian candidates are able to be detected by using a result of LRF scanning, some of the candidates could be wrong. For more reliable process, we take an approach to combine the LRF-based tracking with pedestrian detection from an omni-directional image. Because LRF-based tracking has a good effect of confining searching area, this is also advantage from the viewpoint of image processing. For instace, red frames described in Fig. 3 show the cropped areas which are specified from the LRF based tracking process. Fig. 4 shows extended pictures of them. In some cases, a pedestrian was captured in an accurate fashion, but other figures can include a different object such as a tree or a traffic sign.

After the candidates selection, it is investigated whether or not the image actually captures a pedestrian by means of HOG features. In this calculation, each image is first resized to 32 width by 64 height, and 30 width by 60 height area where is the center of the image is used for feature calculation. The size of each cell is 5×5 pixels. In order to calculate orientation histogram, The orientation range (0 to π) is divided into 9 bins, and block size which is used to normalization is set to 3×3. Finally, a 3240-dimensional vector is generated under these condition.

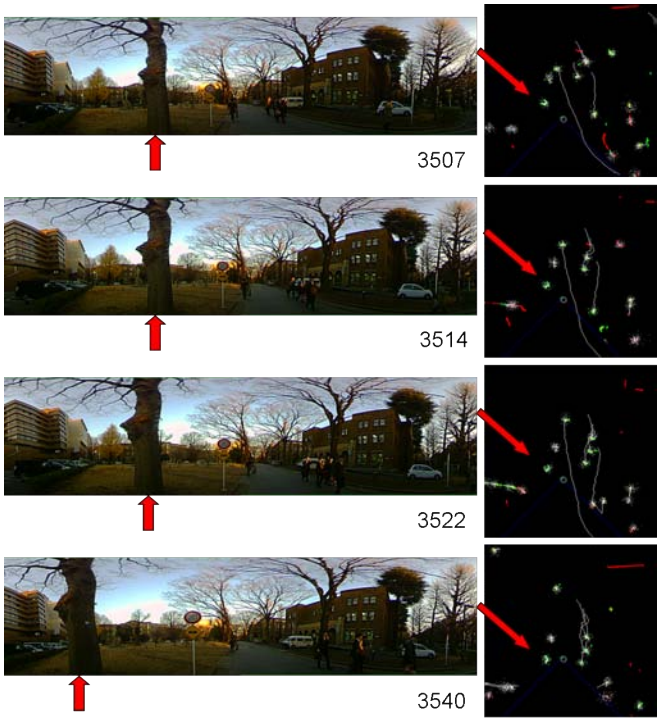


Fig. 2. Tracking segments using mixture particle filter

Fig. 3. Panoramic image and rectangles to clip

Fig. 4. Automatically clipped images from panoramic images

Images depicted in Fig.5(a) show part of dataset used to learning process. The dataset consists of selected images which are captured while driving tests of the robot. it includes 510 images capturing a pedestrian, and other 515 images. The latter images consist of 300 images randomly cropped from background, and the rest of it captures an object which easily causes error recognition.

Because it is assumed that various size of pedestrians in an image should be targeted, the detection process has to be able to cope with various resolution. In order to investigate the influence of resolution changes, INRIA Pedestrian Dataset[9] were used. 920 pedestrian images and 1000 background images were selected (Fig.5(b) shows some examples).

(a) Clipped images from camera images(distances:5m,7.5m and 10m)

(b) INRIA Person Dataset

Fig. 5. Part of our and INRIA person dataset

For constructing SVM classifier, C-SVM with RBF kernel provided by LIBSVM [10] was used. Cross validation method is applied for the evaluation. The method uses a dataset that all of samples are labeled. The basic evaluation procedure is as follows; (i) the dataset is divided into n subsets, and then (ii) a discriminant function is generated from other $n-1$ subsets. (iii) remained one subset is input to the function and a discrimination success rate is calculated. These (i) to (iii) is tried n times with changing the removed subset, and final result can be provided by averaging the success rates. In our case, the highest success rates were 99.8% at INRIA dataset, and 94.5% at our own dataset.

VI. Discrimination success rate against multi-resolution images

A. Performance when the resolution of testing data is differ from that of training data

If a person stands at a distance of 7 meters from a camera, an image region captures the person has almost the same size as 32×64 pixels which is an input size of HOG feature calculation. The more the pedestrian gets away from the camera, the more the image region is contracted. Because our aim is to detect a pedestrian who stands 20 meters away, it is important that we bring out discrimination success rate by using HOG feature against such low resolution target.

Yamauchi et. al [11] performed discrimination process by using various spatial and temporal patterns of HOG features. They generated a set of multi-resolution images from an image and investigated discrimination success rate. The result shows that the success rate varied among body parts.

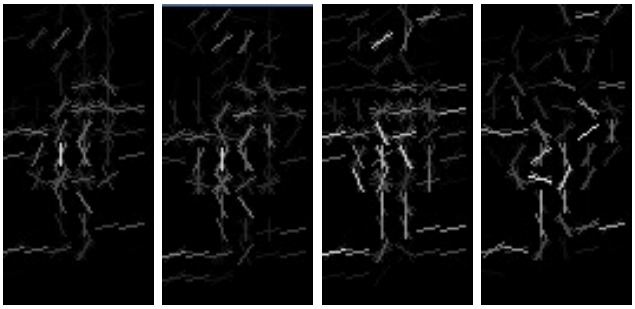
From this fact, we also tried to clear what happens if there are resolution difference between training data and test data. In this experiment, INRIA database was used. At the beginning, each original image was reduced the size to 30, 20, 12 and 6 pixels in width. Next the images were expanded to original size again (See examples at Fig.6). In the expanding process, linear interpolation was applied. Finally, these images divided into training and testing dataset,

Fig.8 shows that the transitions of success rate about the test data under the condition that the resolution of the training data is varied. The horizontal axis indicates the resolution changes.

Upper graph shows the result about false positives. The error rate should be reduced because this directly affects

(a) 30pixel (b) 20pixel (c) 12pixel (d) 6pixel

Fig. 6. Scaled-up images after being reduced once



(a) 30pixel (b) 20pixel (c) 12pixel (d) 6pixel

Fig. 7. Orientation and magnitude of cells

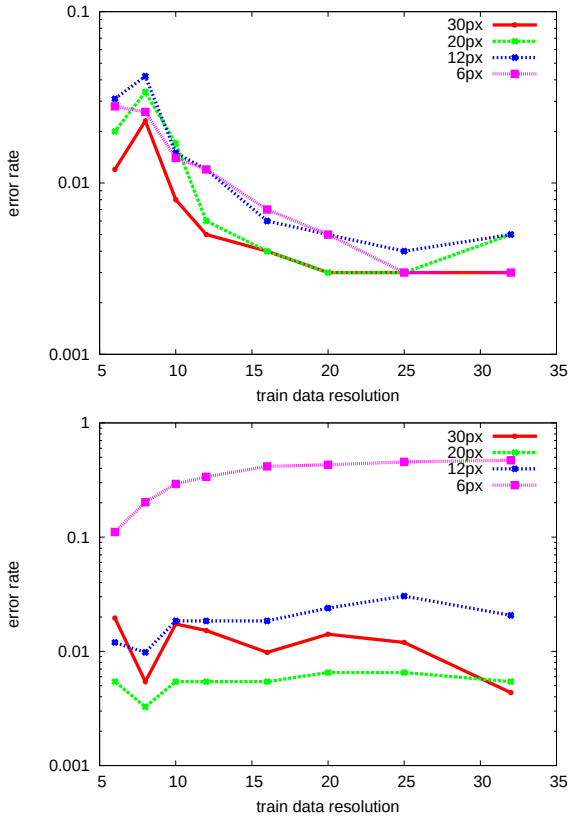


Fig. 8. The performance of SVM varying the resolution of training data and test (INRIA Person Dataset). The top shows the error rate of false positive, and the bottom shows that of false negative)

discrimination success rate. As large changes cannot be found when the resolution is higher than 15 pixels in width, it is possible to say that a certain level of low resolution is allowed. In our case, a person standing 15 meter away from a camera is captured with about 15 pixels in width.

In the case of false negative, there is not large changes by varying the resolution of training data. Instead the performance grows in the case of 6 pixels. This is because high resolution images belonged to positive data are characteristic images which have strong edges in its center. On the other hand, although some of images belonged to negative data may have edges derived from an object or background, they become less distinguishable when the images are smoothed.

B. Experiments using various resolution images

The former subsection reported about the performance if resolution of images is differ training data and testing data. However, this evaluation is insufficient because our training dataset itself includes various resolution images. So we also investigate the discrimination performance with a training dataset including various resolution images. In this experiment, one training dataset consists of 4 series of resolutions, and all of them are used for learning at once.

Fig.9 shows the evaluation results using INRIA dataset. There were three patterns of training datasets which included 4 series of resolutions during (i) 21 to 30 pixels, (ii) 15 to 30 pixels and (iii) 6 to 30 pixels in width. Learning results were used to distinguish an input dataset which consisted various resolution during 10 to 30 pixels in width. Pink dots shows the result against an input dataset includeing images with 30 pixels in width.

The result of false positive tended to be improved. Especially if a low resolution images was input, the performance became good against testing dataset including low resolution images. On the other hand, almost no changes couldnot be found expect for the case that the dataset included low resolution images as 6 pixels in width. These facts tells us that the learning using various resolution images does not particularly influence the discrimination performance.

VII. Application to a taxi robot

The proposed method implemented on a personal mobility robot as shown in the left figure of Fig.10. As an application to the robot which moves outdoor, we tried to develop a taxi service. If a person swings his arm with facing to the robot, it finds the motion and moves to the front of him. Pedestrian detection and tracking has a role in the identification of target position.

The figures shown in Fig.10 indicate our experimental hardware. This was an inverted pendurum type mobile robot which was the development of PMR made by Toyota Motor Corp. Some sensors such as cameras, LRFs, and an IMU (inertial measurement unit) were equipped on. The LRFs were UTM-30LX made by HOKUYO Inc. and an omni-view camera mounted on the top of the LRF is NM33 made by

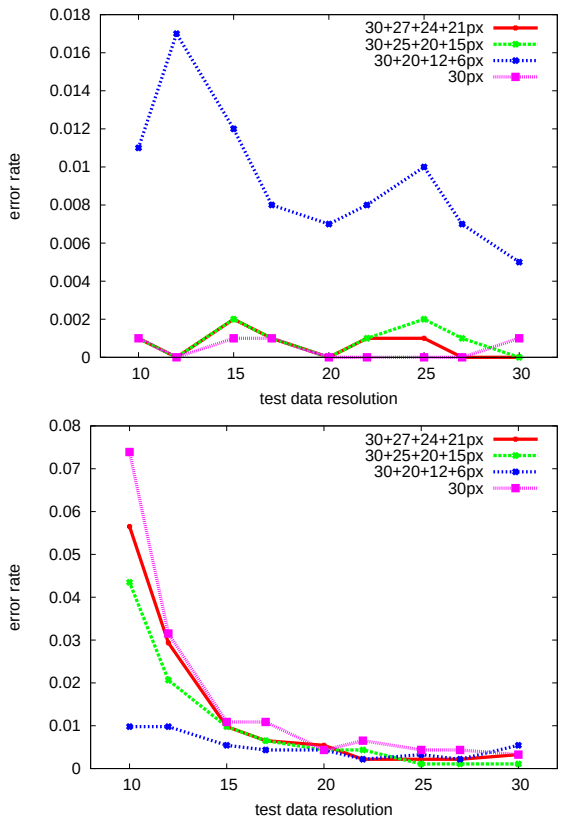


Fig. 9. The performance of SVM using mixed multi-resolution images as training data(INRIA), (top:false negative, bottom:false positive)

Opt Inc. This camera mounted fish-eye lens which had 17° depression angle. This configuration provided significantly wide viewangle. In our case, the camera mounted with 1300 [mm] height could capture a foot of human separated from 4000 [mm]. The IMU was used to compensate swinging motion of the robot.

Two PCs which had 2.33GHz and 2.60GHz CPU were mounted on. The former one was used for all of processing about pedestrian detection and tracking, another was used to plan and control the robot motion.

A. Pedestrian detection and tracking

Table. I shows a classification result. During an experiment, 1979 regions were cropped from image streams as a pedestrian candidates. In this result, images which were very low resolution were classified to negative result because it was difficult to judge whether or not a pedestrian was included in it. Training dataset for the making of a discrimination function had been collected at different environments in advance.

Because one pedestrian walking around the robot was able to be tracked over several images, the classification result was calculated by a equation as $p^2/(p^2+n)$. p indicates the number of positive result, and n indicates the number of negative result. When this value became over 0.9, the target was judged as a pedestrian. The reason why p was squared that sufficient discrimination could not be realized

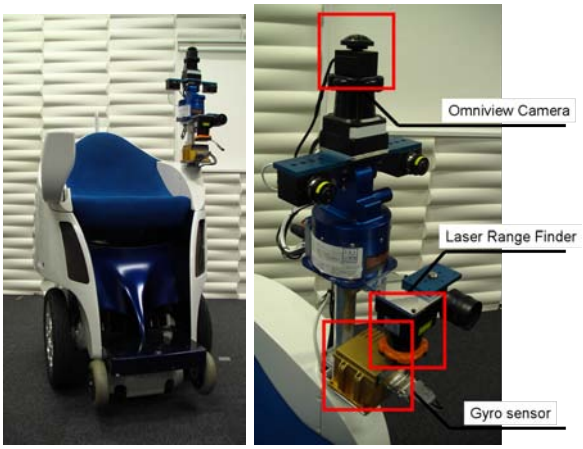


Fig. 10. Robot and sensor positions

TABLE I

CLASSIFICATION RESULT OF IMAGES COLLECTED BY MOVING ROBOT

dist [m]	< 6	< 8	< 10	< 12	< 14	< 16	< 18	< 20
true pos	146	129	72	45	21	12	5	4
false pos	10	15	24	59	25	22	28	26
false neg	61	75	47	36	38	18	6	2
true neg	72	64	132	235	96	210	110	96

with a ratio of positive results to negative results when a target scored a number of positive results.

Fig.11 shows the tracking result by combining LRF data with Image data. Some candidates were first detected and tracked by using LRF, and then pedestrians were identified from them by using omni-directional image. The number framed by '$< >$' indicate sequential serial number of tracking targets. Red numbers indicates extracted pedestrians. In this case, No.2017 and No.2037 shows right result, other two candidates are wrong.

To evaluate the pedestrian tracking method, 5000 frames captured at 10 [Hz] resolution were utilized. In this experiment, several pedestrians were walking, but other moving existences as cars and bicycles interfered stable observation. A total of 2156 clusters were extracted and tracked, and then 130 clusters of them captured a pedestrian. An average of survival time of each cluster was 36 frames. It was relatively short because these results included some situation that other moving object as a car cut across in front of a pedestrian. In such case, a new cluster was generated about the same pedestrian.

A number of clusters which were judged as a pedestrian was 116, these consisted of 38 true positives and 78 false positives (that is, error rate was $78/2026 = 3.8\%$). Because the number of positives was not much, few false positives were observed. This error rate was lower than the result using only HOG features, our approach had an effect at outdoor environment including significant noises.

B. Application to taxi service

As an application of daily assistive robot, we setup a robot system by integrating gesture recognition and navigation.

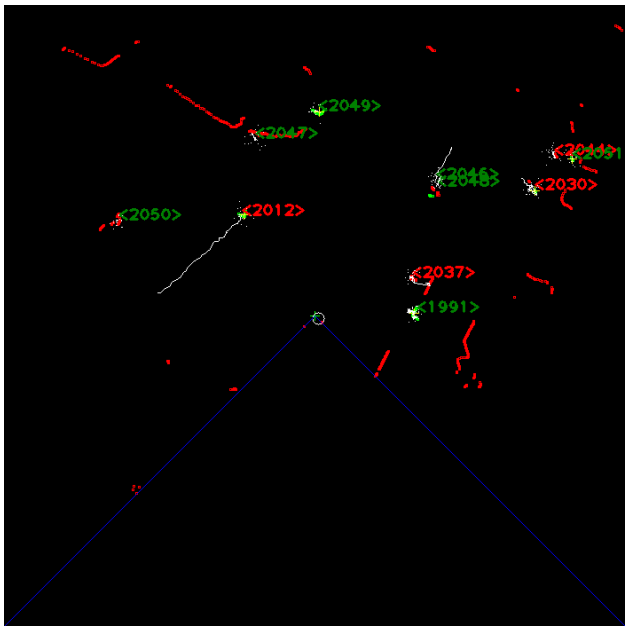


Fig. 11. The result of the proposed human detection method: red clusters are judged as human and green ones are non-human.

(a) Human detection result (red rectangles mean positive, black ones mean negative)

(b) Hand wave detection

(c) Approaching the user and go to the selected destination automatically

Fig. 12. An experiment of taxi service application

Although it is difficult to recognize gestures by a pedestrian who stands far away, a good recognition method which targets hand waving motion have been proposed [12]. Because the method extract intrinsic frequency by means of FFT, it is robust to noise and works at the situation of low resolution. Our gesture recognition was inspired by their method.

The recognition procedure was as follows: (i)after the pedestrian detection, the robot captured an image with zooming in the target. (ii)by using Haar-like feature, it detected the region of upper body and specify a face position. (iii)both side of the face were observed and the time series variation was checked. If large and regular changes could be observed, it was judged as calling gesture. Fig.12 shows an experimental result. As shown in Figures (a) and (b), the robot found its target person, and moved to the front of him, and carried him.

VIII. CONCLUSION

This paper represented about a pedestrian detection method by combining a LRF with an omni-view camera. Low-resolution images of distant pedestrians were used, We proposed a robust pedestrian detection method suitable for personal mobility using a laser range finder(LRF) and a small omni-view camera.

By tracking the possible human areas using the LRF, it could reduce false detections, even when the resolution of images acquired by the robot moving around outdoor environment was fairly low. In addition, the realization of taxi-like robot system as a concrete application was introduced. The system included finding of passengers and recognition of their hand-waving gesture while the robot was moving.

Future works, an unified approach which considers the distance to a target should be developed. As other issues, because HOG feature can also represent general objects as well as pedestrians, some applications to detect other outdoor objects will be useful for more robust pedestrians and other object detection.

REFERENCES

[1] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. ICCV '05, pp. 90–97. Citeseer, 2005.

[2] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *Lecture Notes in Computer Science*, Vol. 3952, p. 428, 2006.

[3] R. Miezianko and D. Pokrajac. People detection in low resolution infrared videos. In *In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops(CVPR Workshops 2008)*, pp. 1–6, 2008.

[4] O.M. Mozos, R. Kurazume, and T. Hasegawa. Multi-Layer People Detection using 2D Range Data. *In Proc. of IEEE International Conference on Robotics and Automation Workshop (ICRA Workshops 2009)*, 2009.

[5] LI Hao, Y. Ming, and Q. Huijia. Camera and Laser Scanner Co-detection of Pedestrians. *In Proc. of IEEE International Conference on Robotics and Automation Workshop (ICRA Workshops 2009)*, 2009.

[6] Vladimir N. Vapnik. Statistical learning theory. *Wiley-Interscience*, 1998.

[7] P. Nú nez, R.Vázquez-Martin, A. Bandera, and F. Sandoval. Feature extraction from laser scan data based on curvature estimation for mobile robotics. *In Proc. of IEEE International Conference on Robotics and Automation (ICRA 2006),*, Vol. 1, pp. 1166–1172, 2006.

[8] J.Vermaak, A.Doucet, and P.Perez. Maintaining multi-modality through mixture tracking. In *In Proc. of International Conference on Computer Vision(ICCV 2003)*, Vol. 2, pp. 1110–1116, 2003.

[9] N.Dalal. INRIA Person Dataset. http://pascal.inrialpes.fr/data/human/.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/˜cjlin/libsvm/.

[11] Y. Iwahori Y. Yamauchi, H. Fujiyoshi and T. Kanade. People Detection based on Co-occurrence of Appearance and Spatio-Temporal Features. *National Institute of Informatics Transactions on Progress in Informatics*, No. 7, pp. 33–42, 2010.

[12] N. Wakamura K. Irie and K. Umeda. Construction of an Intelligent Room Based on Gesture Recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 193–198, 2004.