Panoramic Vision and LRF Sensor Fusion Based Human Identification and Tracking for Autonomous Luggage Cart

Mehrez Kristou, Akihisa Ohya and Shin'ichi Yuta
Intelligent Robot Laboratory, University of Tsukuba, Japan.
{anjin,ohya,yuta}@roboken.esys.tsukuba.ac.jp

Abstract— **In this paper, we propose a solution for human identification and localization with a mobile robot problem that implements multi-sensor data fusion techniques. This solution is designed for an autonomous luggage cart. The system utilizes a new approach based on identifying the target human visually from an omni directional camera then localizing and tracking him using LRF. This approach is composed of "Registration Stage" and "Identification and Localization Stage". The registration stage extracts all necessary information needed including patches from the clothes. The identification is made using a modified pattern-matching algorithm to fit to a real time application. The tracking is implemented using a positions history structure to keep record of all positions of surrounding objects and the identified human. We implemented the proposed approach in fixed configuration to test its effectiveness.**

I. Introduction

In the last years, the service robots got an increasing interest because of their advanced abilities which exceeds the navigation within the environment they have been placed in. Now service robots have more challenging perspectives. They have to interact with people to provide useful services and show good communication skills and to deal with the environment difficulties. In general, a service robot has to focus its attention on humans and be aware of their presence. Moreover, for some specific application like an autonomous luggage cart, it is necessary to distinguish the target human.

Therefore, It is necessary to have a tracking system that returns the current position, with respect to the robot, of the target person. This is a challenging task because of the unpredictable human behavior. Researchers have been using different methods to deal with this problem in many cases, with solutions subjected to strong limitations, such as tracking in rather simple situations with a static robot or using some additional distributed sensors in the environment.

Human tracking can help service robots plan and adapt their movements according to the motion of the adjacent people or follow an instructor across different areas of a building. For example, the tour-guide robot of Burgard et al. [1] adopts only laser sensor to implement people tracking both for interacting with users and for mapping the environment, discarding human occlusions. Another field of application is automatic or remote surveillance with security robots, which can be used to monitor wide areas of interest that are otherwise difficult to cover with fixed sensors. An example is the system implemented by Liu et al. [2], where a mobile robot tracks possible intruders in a restricted area and signals their presence to the security personnel. While these researches use LRF as a basic sensor other research consider the usage of camera. For example, the developed system of Jianpeng Zhou and al. [3] in which they present a real time robust human detection and tracking system for video surveillance which can be used in varying environments. Another example of human detection using cascade of histogram, we can find the work of Qiang Zhu et al. [4] in which they integrate the cascade-of-rejectors approach with Histograms of Oriented Gradients (HoG) features to achieve fast and accurate human detection system.

To ensure more accuracy, it is common to use sensor fusion and especially a combination of camera and LRF is adopted. For instance, The robot described by Luo et al. [5] uses a tilting laser to extract body features, which are fused then with the face detected by a camera. The solution is useful for pursuing a person in front of the robot; however, the complexity of feature extraction limits its application to multiple people tracking.

The solution presented in this paper adopts multi-sensor data fusion techniques for tracking people from a mobile robot using a laser scanner and an Omni directional camera. Different of the related works in many aspect, our system is a mobile configuration which differs from the work of [3] and based on multi-sensor fusion which also differs from [1], [2] and [4]. A new detection algorithm has been implemented to find the target human using patches of his clothes extracted from the image. This technique is to solve the limitation of the face detection when the human is facing back the robot in the work of [5]. The localization is based on the accuracy of the laser scanner. Different from other approaches, our system is able to recognize the target human visually and then localize him. The tracking system labels all objects around the robot through the time and keeps track of their positions.

This paper is organized as follows. Section II introduces the human registration stage. This stage is the initialization part of the system. Section III explains, in detail, the algorithm for human detection and localization and also introduces the general overview. It includes the tracking system, human position history, sensor fusion, and data association. Then, Section IV presents the conducted experiment and analyzes the results. Finally, conclusions and future work are illustrated in Section V.

II. Registration Stage

The system has to detect the human that it must follow, for this reason; the robot must have some features to look for it. Once found, it can decide the following behavior. To identify a human, the robot must first have a model of the searched human. This model is a collection of characteristics identifying the human and making his identification as unique as possible. This process is called human registration.

When the human stand in front of the robot and order the registration process, the robot starts to execute the first procedure: the registration step is shown in Fig.1.

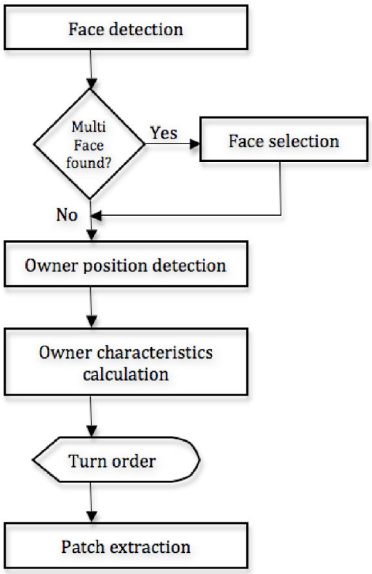

Fig. 1. Human registration steps flowchart

First, the robot must understand that the object that it precedes for registration is human. We use an omni directional camera, which delivers a 360deg image to recognize the human visually. The raw format of the image limits the possibilities of the image processing algorithms, so we transform it into panoramic image. To detect a face, we use a recent technique of face detection to find a face in the panoramic image. If more then one face is detected, the owner selects his face. The human face characterizes the human. The face detection algorithm[6] is implemented in OpenCV library[7] and it is based on using Haar-like features (Haarcascade). The idea here is to first train a classifier with number of sample views of object. Based on this concept, researchers have trained classifiers for detecting faces. Once the classifier is trained (or built), then based on certain Haar-like features it (classifier) can be used to search for desired object (face in our case) in an image or live video from a camera. Usually a number of simple classifiers are built, and then each of them is applied to the concerned image one by one (thus the name Haarcascade). If the classifiers output is 1, then it means that the detecting region is likely to include the object of interest. Otherwise, 0 is the output of the classifier. The classifier is designed so that it can be easily resized in order to be able to detect the objects of interest at different sizes, which is more efficient then resizing the image itself. So, to find an object of an unknown size in the image the scan procedure should be done several times at different scales.

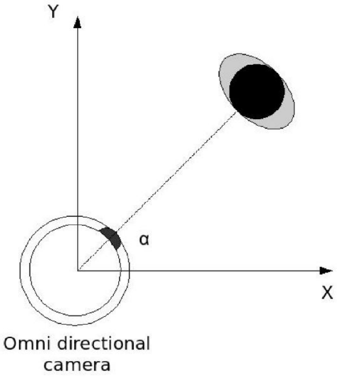

Fig. 2. Vertical view of the omni directional camera and human. the 360deg image delivered by the camera allows the localization of the human face using angular position

After locating the face in panoramic image, we use that position as a reference position. The registration of the characteristics of the standing human can start. The registration includes the detection of the human width, skin color, which can be recorded for future usage to improve the performance, and samples of the color pattern of the clothes. An early sensor fusion mechanism is to be used to calculate the position of the standing human and correct his relative distance in order to get the best and accurate measurements. The panoramic image is a conversion of the omni directional image. Using this characteristic, the position of every vertical line in the image can be transformed into angular coordinate as shown in Fig.2. The camera and the LRF are calibrated to share the same vertical axis, so the angular position of an object in the panoramic image is the same as the angular position of the same object in the LRF data. We consider the horizontal axis of the image as X-axis and the vertical axis as Y-axis.

Once the face is detected, we convert the x coordinate of the center of the face to the angular coordinate in radian. In the LRF data, we cluster it into cluster using maximum allowed Euclidean distance between successive points. We select the cluster corresponding to the angular position. This cluster is labeled as the face cluster. The face detection and LRF data fusion is explained in Fig.3. In order to detect human from all sides, the robot must get the patches of clothes from all sides. These patches must be from the chest area, so using the position of the detected face, we select the area from which the patch will be extracted. The position is 3/4 face down of the detected face and the size of the patch is the same as the face size. For this reason, the human must make a complete turn in front of the robot that it can get all needed patches (color pattern samples of the clothes). This information is stored and used as initialization of the system. All further process use this information as a reference data. After getting all information needed for human identification, the registered human can go and move normally without taking care of the robot.

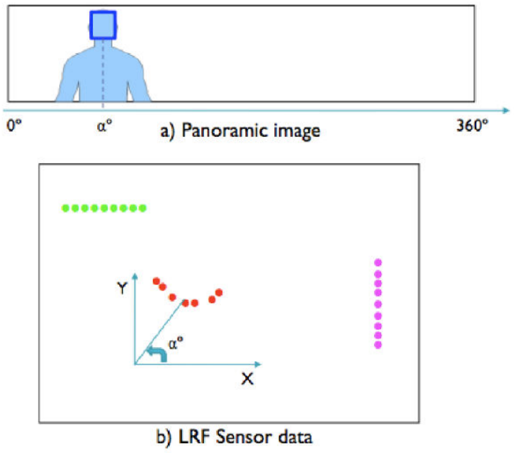

Fig. 3. a) Face localization in the panoramic image and its corresponding angular position calculation α, b) Standing humans cluster corresponding to the face angular position. The red cluster is the detected human from the face location

III. Identification and Localization Stage

The human-tracking algorithm adopts multi-sensor data fusion techniques to integrate the following two different sources of information: the first one is pattern matching, based on panoramic image adapted from an omni directional camera, and the other one is position history of all surrounding object based on the laser scans of the LRF. The first source of information needs an initialization step which prepare the patches to be applied by the pattern matching and gives extra discriminative features to eliminate false positives. The panoramic image is a 360deg image so we can transform the horizontal position into angular position. Using a configuration in which the omni-directional camera and LRF share the same vertical axis, we can insure the direct transformation of the calculated angle from one sensor to another. Next, we describe, in detail, the principles underlying these two detection algorithms.

A. *Panoramic vision based identification and localization*

Once the robot saved the initialization data (registration step), it can start finding the registered human. To implement this step, we use a pattern-matching algorithm implemented in OpenCV**??** that we modified to decrease the processing time. The basic idea of this algorithm is to slide the pattern or the patch over the source image and to calculate the correspondence coefficient then save this coefficient in an output image in the same position as the upper left corner of the patch. As a result, the output image has the dimension of the input image minus width of the patch in its width and minus height of the patch in it height. This output image is called disparity map. To know the best matching result, we can find the max intensity value of all pixels of the disparity map. The corresponding position is the upper left corner of the matched patch and its dimensions is equal to the original patch size. The pattern-matching algorithm detects only pattern having the same dimension as the extracted patches, so applying this method will result into identifying only human standing in the same distance as the registration step. For this reason, we used a patch preparation step, which consist of zooming out the initial patches using bilinear interpolation algorithm implemented in OpenCV**??** to make for each patch three sized down patches. This operation gives better flexibility to the algorithm to detect the same patch from different distances.

This pattern-matching algorithm is very time consuming due to the size of the searching image (panoramic image). In my case, the searching image size is 1125x178 pixels. As a consequence, the processing time for 4 patches is around 2 seconds for one image, which is very non real time processing. To improve this performance, we used a down sampling method as follow:

1) We performs down-sampling step of Gaussian pyramid decomposition[7] to down-sample the panoramic image two times. First this algorithm convolves source image with the Gaussian 5x5 filter and then down-samples the image by rejecting even rows and columns. So the destination image is four times smaller than the source image.
2) We perform a Template matching algorithm[7] , based on equation (1), on the down sampled images
3) We find the top N matches in the result image (this array is in order from best match to worst based on confidence value)
4) We coordinate transform the N matches back to the original image
5) We use the Template matching algorithm on the original images (panoramic image), but we set the ROI on the search image to the same size as the template image, centered at the N matches points.
6) We repeat 5 for remaining N matches until a suitable match is found

$$S(x,y) = \frac{\sum_{y'=0}^{h-1}\sum_{x'=0}^{w-1} T(x',y')I(x+x',y+y')}{\sqrt{\sum_{y'=0}^{h-1}\sum_{x'=0}^{w-1} T(x',y')^2 \sum_{y'=0}^{h-1}\sum_{x'=0}^{w-1} I(x+x',y+y')^2}} \qquad (1)$$

The result of this algorithm is a list of the best candidates with their respective positions. A threshold on minimum confidence value is used to limit the number of selected candidates. Because the original image is a panoramic image (converted from the omni directional image), we transformed the image coordinate positions into angular positions.

B. LRF sensor based localization

This step implements a cluster tracking system based on position history building. We consider the points received from the LRF as chain of data. We compare the Euclidean distance between two successive points, if the distance is more than a pre-defined constant, we break the chain into two and we continue up to the last point. The result will be cluster of data corresponding to the visible unconnected objects.

The step is realized basically following this order:

1) We segment the LRF data using Euclidean distance and a maximum in group distance to divide the original set of data into clusters.
2) We apply a linearity check algorithm based on the calculation of the variance of linear regression to eliminate clusters with high goodness-of-fit coefficient and keep only ones having the most likely a human shape.
3) We update the position history of the detected cluster.

The position history structure is a two dimensions table: one for candidates and other for the cycles and the corresponding intersection is the state. Using the minimum Euclidean distance between the current clusters and the last detected positions, the new measurement can be affected to the corresponding candidate row. If some candidates don't have a new measurement, the state will be affected as lost.

C. Vision and LRF fusion

Using the outputs of the last two steps: The angular position of detected candidates in the panoramic image and the position history from the LRF, this step enriches the position history with the candidate's information. As shown in Fig.4, the fusion of the output of the vision and LRF is list of positions and states of each candidates through the time. The positions history structure is a 2D table where rows are candidates, columns are time index and cells are candidates' position and state. Measurements are affected to the corresponding candidate's row depending on the shortest Euclidean distance between the last position of the candidate and the measurement.

The LRF and the omni directional camera are placed vertically one on top of the other and calibrated to have the same position in the horizontal coordinate. This configuration allows the two sensors to share the same Z-axis (vertical axis). Using this configuration, gives a common coordinate, which simplifies the sensor fusion in this method. The omni directional camera gives the angular position of the identified candidate and the LRF tracking gives the position history or the place of each detected clusters on the space and time. This step mark the state of the candidate with identified mark if the corresponding cluster exist in the corresponding angle of the identified candidate. The panoramic vision component, LRF and the fusion components are running independently. Each component sent its result to a shared memory and the fusion component update the enriched positions history only when there is available information.

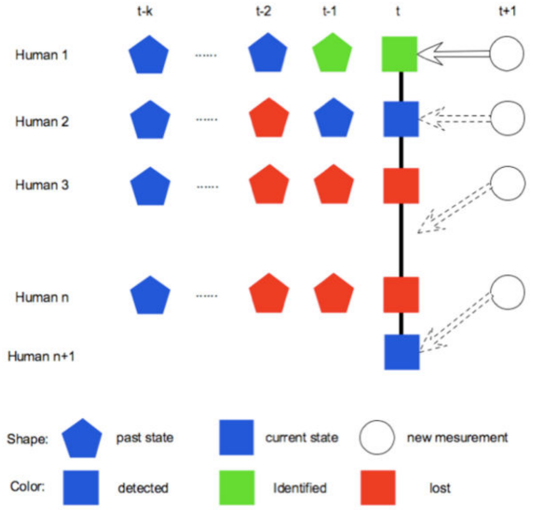

Fig. 4. Positions history structure where rows are candidates, columns are time index and cells are candidates' position and state. Measurements are affected to the corresponding candidate's row depending on the shortest Euclidean distance between the last position of the candidate and the measurement.

IV. Experiment and Results

To test the performance of the proposed approach, the system has been implemented on a mobile robot provided, shown in Fig.5, with a SOKUIKI sensor[8] (URG-04LX, Hokuyo Automatic Co., Ltd.) as a LRF and an omni directional camera. The two sensors are mounted on a special support at approximately 1.64 m from the floor in order to facilitate the face detection. The on-board PC is a Core 2 Duo 2.5 GHz with 1 GB of RAM. The whole software has been written in C++ and runs in real time on the robot PC. The resolution of the laser device is Âś1% of the distance, with a scan every 0.36deg at 10 Hz, whereas the cameras provide images with a resolution of 640 x 480 pixels at 30 fps.

The experiments have been conducted in our laboratory. During the experiments, the robot was in fix configuration.

A. Registration stage

To evaluate the performance of the registration stage, a scenario was prepared as follows:

1) The robot is initially in the charging dock waiting for a human to stand in front of it.
2) A human stands in front of the robot and order the identification.
3) The robot detects the face in the omni-directional image then calculates the face angle, Fig.6.
4) Using the special angle of the face, the robot locates him using LRF data.
5) If the owner is very near or too far the robot, the robot order him to correct his distance.
6) 6. The robot orders the detected human to turn 360deg while the camera extracts patches of the human clothes to be used in further process.

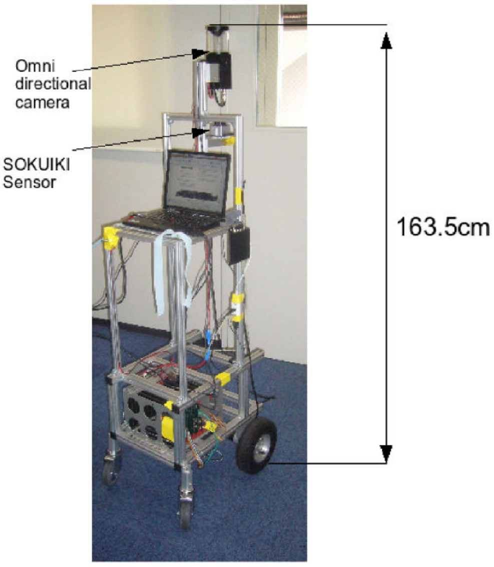

Fig. 5. Developed mobile robot equipped with an omni directional camera and LRF sensor

The characteristics of the camera allow seeing the upper part of human if he stands if front of it in distance between 1m and 1.5m. Over 1.5m the quality of the extracted patch will be dramatically affected, because the detected face will be too small therefore the patch will also small.

The face detection algorithm is time consuming for a big image like the panoramic image, so we limited the detection area to the front side of the robot. The result of the face detection is shown in Fig.6. The red square in the panoramic image is the position of the detected face. The blue square is the position of the patch; it is placed relatively to the position of the face. The number on the side of the square is an ordinal number to identify the pairs in case of multiple detected faces. The result of the early sensor fusion is shown in Fig.7. After segmenting the LRF data, the human cluster is selected using the angular position of the face. We applied and ellipse fitting algorithm to the selected cluster.

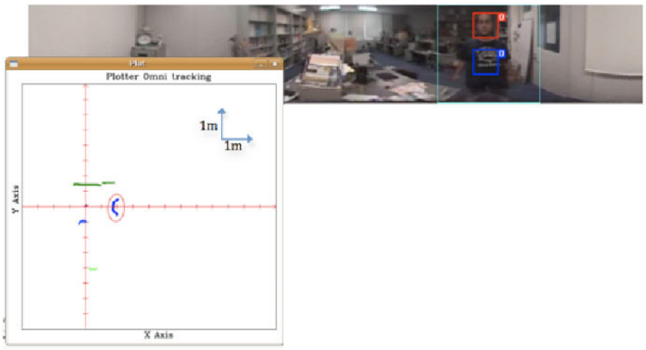

Fig. 7. Result of the face localization, the red ellipse is the result of the ellipse fitting applied to the selected cluster. The coordinate unit is 0.5m

After 5sec of detecting one face, the robot orders the human to start turning. While the human is turning, the ellipse fitting[9] applied to his cluster gives the approximate angle. Because the selected directions are extreme, the seen width of the human very from 75cm to 20 cm. This difference of width combined with the ellipse fitting method allows better human angle detection. During this process, 4 patches are extracted. The extracted patches are squares having the same size as the detected face area. The ellipse fitting result is facing some difficulties due the few number of points in the cluster. So we adopted a simple state machine detecting a successive state of the seen width of the turning human. The combination of these two methods improved the human turning detection.

B. Panoramic vision based identification and localization

During this stage, a continuous process take the 4 patches extracted in the registration stage, zoom out each one 2 times. The total number of patches is 3 sized sets of 4 patches. Then for each set of patches, we apply the modified pattern-matching algorithm.

The result of this algorithm is a confidence map. Because the bigger patch has better probability to be a correct matching, we multiply the resulting confidence map of each patch by 1 for the 1:1, 0.75 for 3:4 and 0.5 for 1:2.

The result of this stage is an array of all matching with their calculated confidence and their angular position. Fig.8 shows the pattern matching result with the position of the detected patches and their confidence.

The processing time of this algorithm for 12 patches, including 4 directions and each direction 3 sizes, is 460ms, which allows 2fps.

C. LRF based localization

This task is relatively easy but has a big impact on the accuracy of the localization and tracking in further steps. The raw data are segmented into clusters and a line eliminate is conducted, the result is shown in Fig.9.

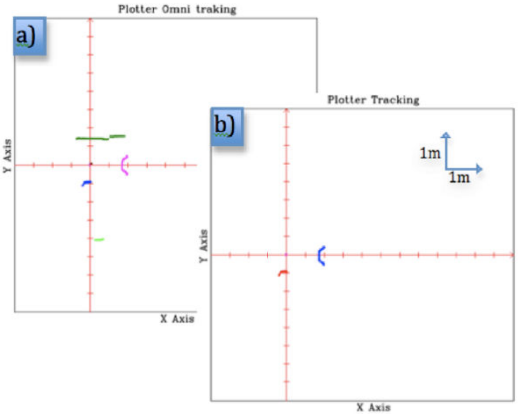

Fig. 9. Result of the line elimination. a) before elimination, b) after elimination. The coordinate unit is 0.5m

The processing time of this stage is 125ms including the delay of the LRF scanning time which is 100ms.

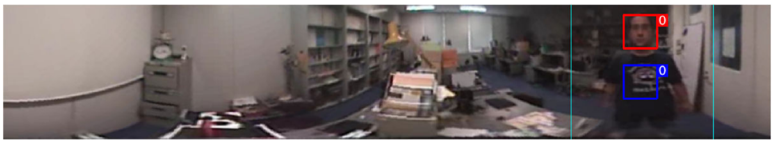

Fig. 6. Result of the face detection algorithm (red) and the position of the patch extraction (blue), the number is the index of the detected face

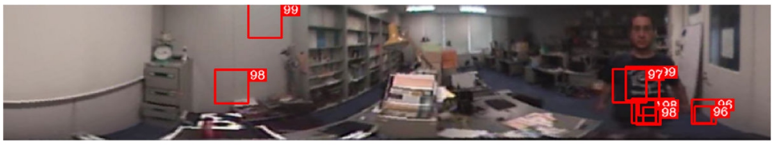

Fig. 8. Result of the pattern matching algorithm (red squares), the number is the confidence value (percentage) of the corresponding patch position

D. Vision and LRF fusion

This stage is the key component if the experiment in which the result of the two processes is merged to get a real time state of all object seen by the sensors and the position of the target human. As shown in Fig.10, the same object keeps its number through the time and shows that it is successfully tracked. The green color square is the identified human to be tracked. The blue square labeled 0 and 1 are desks, which are parts of the experiment's environment.

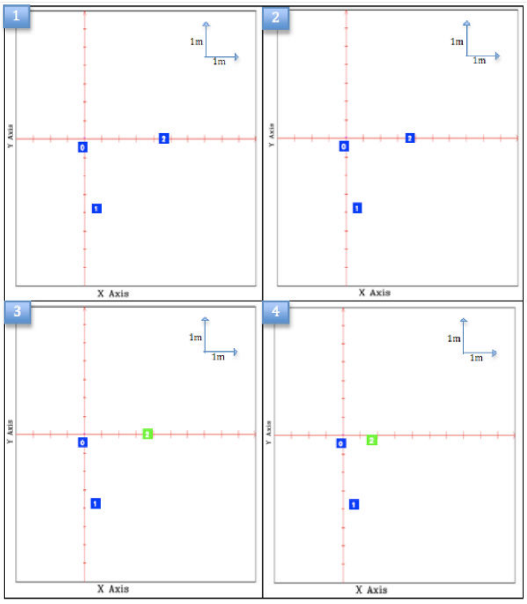

Fig. 10. Result of the sensors fusion through the time. 1),2),3) and 4) are successive states. Blue color squares are object detected only in the LRF. Green square is the identified human from the Panoramic vision. The coordinate unit is 0.5m

V. Conclusions and Future Works

In this paper, we have presented a multi-sensor based human identification and tracking system for an autonomous luggage cart robot. The proposed approach deals mostly with the owner identification using patches extracted from the color pattern on his clothes. The identified human in the panoramic image has been fused to the LRF based positions history. To verify the proposed approach, we implemented it on mobile robot using fixed configuration and pre-defined scenario. The fixed configuration experiment showed the efficiency of this approach to detect human and identify him.

The current solution could be further improved using better human turning detection method in the registration stage, in which other geometric features or pattern recognition techniques should be investigated and possibly integrated. Aside from this, our future research will focus on the crowdedness part of the environment for a more robust tracking of single person or multiple people, particularly when they gather around the robot.

References

[1] W. Burgard, P. Trahanias, D. HÃd’hnel, M. Moors, D. Schulz, H. Baltzakis, and A. Argyros, “Tourbot and webfair: Web-operated mobile robots for tele-presence in populated exhibitions,” in *Proc. IROS Workshop Robots Exhib.*, 2002, pp. 1–10.

[2] J. N. K. Liu, M. Wang, , and B. Feng, “ibotguard: An internet-based intelligent robot security system using invariant face recognition against intruder,” in *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, Feb. 2005, pp. 97–105.

[3] R. C. Luo, Y. J. Chen, C. T. Liao, and A. C. Tsai, “Mobile robot based human detection and tracking using range and intensity data fusion,” in *Proc. IEEE Workshop Adv. Robot. Social Impacts*, 2007, pp. 1–6.

[4] J. Zhou and J. Hoang, “Real time robust human detection and tracking system,” in *13DVR International Inc.*

[5] Q. Zhu, S. Avidan, and K. T. C. M. Yeh, “Fast human detection using a cascade of histograms of oriented gradients,” in *TR2006-068*, June 2006.

[6] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Dec. 2001.

[7] I. Corporation. (2006) Open source computer vision library. [Online]. Available: http://www.intel.com/technology/computing/opencv/index.html

[8] H. Kawata, A. Ohya, S. Yuta, W. Santosh, and T. Mori, “Development of ultra-small lightweight optical range sensor system,” in *Proc. of IROS’05*, 2005, pp. 3277–3282.

[9] V. Pratt, “Direct least squares fitting of algebraic surfaces,” in *ACMJ. Computer Graphics*, vol. 21, July 1987.