Proceedings of the 2012 IEEE
International Conference on Robotics and Biomimetics
December 11-14, 2012, Guangzhou, China

A SIFT-Based Person Identification using a Distance-Dependent Appearance Model for a Person Following Robot

Junji Satake, Masaya Chiba, and Jun Miura

***Abstract*— This paper describes a person identification technique for a mobile robot which performs specific person following under dynamic complicated environments like a school canteen where many persons exist. We use the SIFT feature for identification of a person, and create the distance dependence appearance model which expects the number of SIFT feature matches based on the distance to a person. The person following experiment was conducted using an actual mobile robot, and the quality assessment of person identification was performed.**

I. Introduction

There is an increasing demand for service robots operating in public space like a shopping mall. An example of service task is to follow a person with carrying his/her items. This research develops a person identification method for such a robot aiming at realizing a mobile robot that can follow a specific user.

There have been a lot of works on person detection and tracking using various image features and classification methods. Beymer and Konolige [1] developed a stereo-based person detection based on background static obstacles subtraction. Howard et al. [2] proposed a visual person detection method whitch first converts a depth map into a polar-perspective map on the ground and then extracts regions with largely-accumulated pixels. Occlusions are not handled there. Ess et al. [3], [4] proposed to integrate various cues such as appearance-based object detection, depth estimation, visual odometry, and ground plane detection using a graphical model for pedestrian detection. Although their method exhibits a nice performance for complicated scenes, it is still costly to be used for controlling a real robot.

We built a mobile robot system with a stereo camera and a laser range finder [5], and realized specific person following in a complex environment with several walking people at a time. The method, however, did not have a sufficient performance to recognize people with similar clothing.

In this paper, we propose a method of identifying a person based on the pattern of clothing using the SIFT feature. We make the appearance model for various body directions, and set the distance-dependent threshold to cope with the decrease of the number of SIFT feature matches according to the increased distance. Finally, we implement the proposed method on an actual robot to perform person tracking experiments.

J. Satake, M. Chiba, and J. Miura are with Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

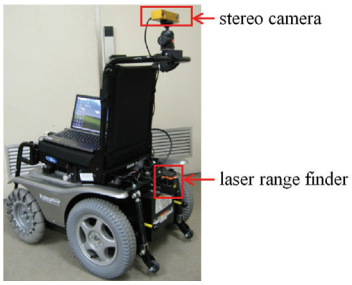


Fig. 1. A mobile robot with a laser range finder and a stereo camera.

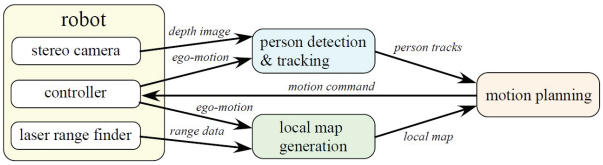


Fig. 2. Configuration of the system.

II. Person Following Robot

A. Configuration of our system [5]

Figure 1 shows our mobile robot system which is composed of

- a computer-controllable electric wheelchair (Patrafour by Kanto Auto Works Ltd.),
- a stereo camera (Bumblebee2 by Point Grey Research),
- a laser range finder (UTM-30LX by Hokuyo), and
- a Note PC (Core2Duo, 2.66GHz, 3GB memory).

Figure 2 shows the configuration of the software system. We deal with two kinds of objects in the environment: persons detected by stereo vision and static obstacles detected by a laser range finder (LRF). The functions of the three main modules are as follows:

1) Person detection and tracking module detects persons using stereo and tracks using Kalman filtering to cope with occasional occlusions among people. Details of the processing are described in Sec. II-B.

2) Local map generation module constructs and maintains an occupancy grid map, centered at the current robot position, using the data from the LRF. It performs a cell-wise Bayesian update of occupancy probabilities assuming that the odometry error can be ignored for a relatively short robot movement.

3) Motion planning module calculates a safe robot motion whitch follows a specified target person and avoids others, using a randomized kinodynamic motion planner.

978-1-4673-2126-6/12/$31.00 © 2012 IEEE

(a) positive samples

(b) negative samples

Fig. 3. Training samples for the SVM-based verifier.

Fig. 4. Snapshots of a specific person following at the cafeteria.

B. Person detection and tracking using stereo

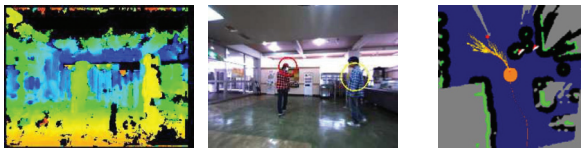
To track persons stably with a moving camera, we use *depth templates* [6], which are the templates for human upper body in depth images. A simple template-based detection is effective in reducing the computational cost but at the same time may produce many false detections for objects with similar silhouette to person. To cope with this, we use an SVM-based person verifier. Fig. 3 shows some of positive and negative samples. We used 356 positive and 147 negative images for training. A person candidate region in the image is resized to 40×40 [pixels] to generate a 1600-dimensional intensity vector. HOG features [7] for that region are summarized into a 2916-dimensional vector. These two vectors are concatenated to generate a 4516-dimensional feature vector, which is used for training and classification.

We adopt the Extended Kalman Filter (EKF) for robust data association and occlusion handling. The state vector includes the position and the velocity in the horizontal axes (X and Y) and the height (Z) of a person. The vector is represented in the robot local coordinates and a coordinate transformation is performed from the previous to the current robot's pose every time in the prediction step, using the robot's odometry information. Color information of the clothing is also used for identifying the target person to follow. The target person is shown with a red circle in the image.

C. Problems of the previous system

Figure 4 shows snapshots of a person following experiment at the cafeteria. And Figure 5 shows an example of person detection, local map generation, and motion planning. We tested the system for the cases where about three persons exist near the robot. Problems which became clear in the experiment are described below.

Figure 6 left shows the failure of target identification using color due to a bad illumination. Figure 6 right is an example



(a) stereo data (b) detected persons (c) recognition result and path planning

Fig. 5. An example of environment recognition and motion planning.

Fig. 6. Failure of target person identification.

which cannot distinguish the target person because there are two persons with same color of clothing. In order to realize stable specific person following, the person identification which used the color and other information together is required. In this paper, we describe how to solve the problem about identification of the target person.

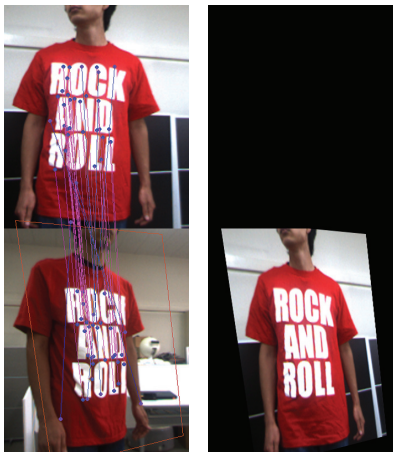
III. A SIFT FEATURE-BASED PERSON IDENTIFICATION

Our previous person identification method using only color information is weak to changes of lighting condition, and it is difficult to distinguish persons who wear the clothing of similar colors. Therefore, we propose a SIFT feature-based person identification method which uses the texture of clothing as a cue.

SIFT (scale-invariant feature transform) [8] is a powerful image feature that is invariant to scale and rotation in the image plane and also robust to changes of lighting condition. The feature is, however, weak to affine transformations. Although a feature which increases the robustness to affine transformations, ASIFT [9], is proposed, it is thought that the identification power will be degraded when the pose of a person changes largely. We therefore use a set of images taken from various directions to cope with pose changes. Moreover, the number of SIFT feature matches between the model and an input image decreases as the person becomes farther from the camera. Therefore, we use a distance-dependent threshold.

A. The number of SIFT feature matches

The number of SIFT feature matches is used for the judgment of whether the detected person is the following target. The person detected from each input image is matched with the appearance model learned beforehand. However, mistaken corresponding points are also contained in matching. Therefore, mistaken corresponding points are removed as follows using RANSAC (RANdom SAmple Consensus) [10] :



(a) estimation of homography (b) transformed model image

Fig. 7. Estimation of homography by using RANSAC.

(a) the number of matches = 30 (b) the number of matches = 52 (c) the number of matches = 30

Fig. 8. Relations between change of the body direction and the number of SIFT feature matches (upper: appearance model, lower: input images).

i) Four pairs are randomly selected from the group of corresponding points.
ii) A homography matrix is calculated based on selected corresponding points.
iii) The number of pairs which satisfy the above homography matrix out of all pairs is counted.
iv) By repeating i) to iii), the homography matrix with the maximum number of pairs is selected.

A example of homography estimated by using RANSAC is shown in Fig. 7. Figure 7(a) shows a correspondence between the appearance model (upper) and an input image (lower). The brown quadrangle shows a form of the model image transformed by the estimated homography matrix. Figure 7(b) shows the transformed model image. The each pair of points connected by pink line shows the correspondence judged as the inlier which satisfies the homography matrix, and the pair connected by blue line shows the outlier. We use for person identification only the corresponding points judged as the inlier.

B. The appearance model

Figure 8 shows matching results of the SIFT features between one image in the appearance model and input images taken from different directions. In matching of the images in which the direction of the person is front, 52 matches were obtained (Fig. 8(b)). On the other hand, the number of matches decreased for the different directions (Fig. 8(a),(c)). Therefore, we make the appearance model for various body directions in the following procedure in order to cope with the pose changes (see Fig. 9).

i) An image sequence in which the person makes a 360-degree turn at 1 [m] away from the camera is recorded.
ii) A certain number of images are picked up at regular intervals from the image sequence. This is because the sequence contains many similar images with a small change of direction and identification would be very costly if an input image is compared with all images in the sequence. In the experiment, we set the number of the picked-up images to 30.
iii) In order to remove the feature points in the background region, each image is segmented into the foreground/background region using depth information. We classified the pixels with depth value of 1 ± 0.5 [m] into the foreground region.
iv) The SIFT feature is extracted from each image in the sequence, and the image whose number of features is less than a threshold is removed. This is for removing the image in which a sufficient number of features are not observed. We set the threshold to 20 in the

(a) an image sequence in which the person made a 360-degree turn
(b) images picked up at regular intervals
(c) extraction of the foreground region using depth information
(d) removal of images with few features
(e) the appearance model

Fig. 9. The creation procedure of the appearance model for various body directions.

experiment.

v) We use the images selected by the above steps for person identification as the appearance model.

C. A distance-dependent threshold

The number of SIFT feature matches decreases when the distance from the camera to the person increases. The image of the upper right in Fig. 10 shows the model image taken when the distance between the person and the camera is 1 [m]. The blue line shows the actual number of corresponding points when the direction of the body is the same and only distance changes. We use a distance-dependent threshold to cope with this decrease of the number of SIFT feature matches.

It is tedious to actually obtain the person images taken at various distances. Instead, we simulate the increasing distance by reducing the size of the model image for generating a simulated input image, and predict the effect of increasing distance. Considering the changes of lighting condition and wrinkles, we use 30% of the predicted value as a threshold.

The examples of three directions are shown in Fig. 10. The red line shows the number of matches predicted by the simulation. It can read that the predicted value (red) and the actual value (blue) have the similar tendency. The green line shows a distance-dependent threshold. This threshold is calculated about each image in the appearance model.

D. Identification of the target person

1) Representative images to recognize rough direction: When identifying the target person, computation cost for all images in appearance model is large. Therefore, in order to recognize rough direction, a certain number of representative images (in this paper, the number is set to 6 with consideration of processing speed) are chosen in advance from the images in appearance model. The best selection of the image set is the combination which can cover image of any body directions. Therefore, we choose an image set from which largest number of corresponding points is obtained about every input image.

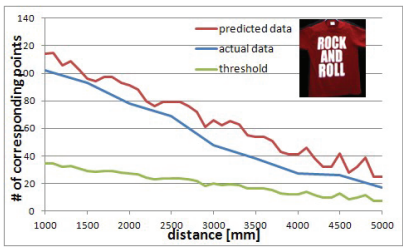
We select the representative images as follows. First, we calculate the number of SIFT feature matches m_{ij} between each image i in appearance model and each image j in image sequence in which the person made a 360-degree turn. On the image j in the sequence, the maximum of corresponding points with every model image is obtained as follows

$$\max_i m_{ij}. \tag{1}$$

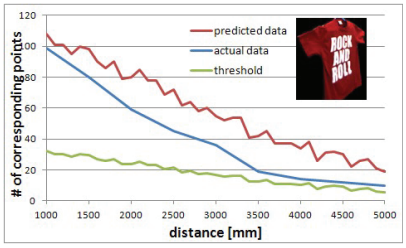
The set of chosen representative images from appearance model is denoted as S. Combination S which makes the following formulas the maximum out of all the combination is chosen

$$\operatorname*{argmax}_S \sum_j \max_{i \in S} m_{ij}. \tag{2}$$

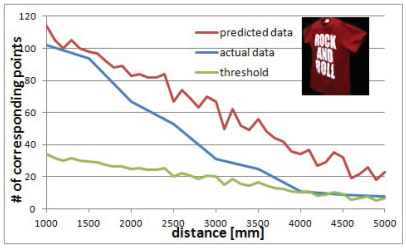
Figure 11 shows an example of combination of 6 representative images selected using this method.



(a) front



(b) diagonally right



(c) diagonally left

Fig. 10. The distance dependence appearance models in the each direction.

Fig. 11. Combination of six models selected using the valuation function.

2) Processing of identification: When the number of the SIFT feature points in an input image is 80, our system needs about 20 [ms] for matching of the feature to each model image, respectively (see Fig. 12). Because it is difficult to compare all model images at each frame, the model images used for the comparison are selected according to the situation as follows:

- If there is the model image matched with the previous frame, only images of direction near it are used for matching.
- In the other case, after recognizing rough direction using the representative direction images described in Section III-B, images of direction near it are used similarly.

Identification of the target person is judged whether the number of matches between input and model (Fig. 10 blue) is over the threshold (Fig. 10 green). That is, the person

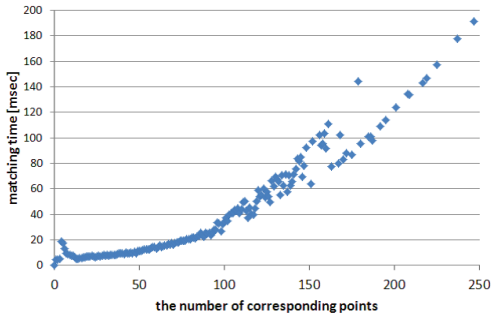


Fig. 12. Relations between the number of matches and process time.

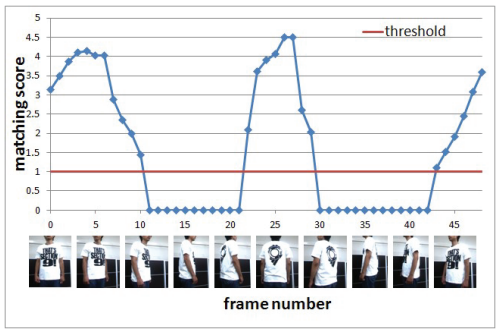


Fig. 13. Tolerance to direction change.

is judged as the target when the following evaluation value (matching score) is over 1.

$$\text{matching score} = \frac{\text{the number of matches}}{\text{threshold according to distance}} \quad (3)$$

Where there are more than one target candidates, the person with the highest matching score is selected as the person to follow.

IV. Experimental Result

A. Verification to direction change

The appearance model was created about five kinds of clothing, respectively, and the identification experiment was conducted on image sequences in which the person made a 360-degree turn at 1.5, 2.0, 2.5 [m] away from the camera. The identification result about each model is shown in table I. The acceptance rate in the table is the number of the images identified as the clothing of the appearance model among the number of images in the test sequence. Note that the model images without sufficient number of SIFT features are deleted from the statistics. The detail of an identification result when the same clothing as a model is tested is shown in Fig. 13. When the matching score is 1 or more, it judges that the person in an input image is same as the registered person. In Fig. 13, input images at #11–21 and #30–42 are rejected. It is because the feature points were not obtained since the body became sideways mostly. The target person, however, was identified almost correctly when the texture of clothing was observed. We think that the cause of failure of identification is transformation of the texture by wrinkles of clothing.

TABLE I

THE ACCEPTANCE RATE ABOUT VARIOUS BODY DIRECTIONS

(a) when the distance is 1.5 [m]

appearance model (made at 1 [m])	test data set (turned at 1.5 [m])				
	**0.961**	0	0	0	0
	0	**0.909**	0	0	0
	0	0	**0.969**	0	0
	0	0	0	**0.968**	0
	0.035	0	0	0	**0.939**

(b) when the distance is 2.0 [m]

appearance model (made at 1 [m])	test data set (turned at 2.0 [m])				
	**0.929**	0	0	0	0
	0	**0.888**	0	0	0
	0	0	**0.945**	0	0
	0	0	0	**0.939**	0
	0	0	0	0.022	**0.952**

(c) when the distance is 2.5 [m]

appearance model (made at 1 [m])	test data set (turned at 2.5 [m])				
	**0.944**	0	0.090	0	0
	0	**0.897**	0	0	0
	0	0	**0.976**	0	0
	0	0	0	**0.843**	0
	0.035	0.054	0.044	0.022	**0.908**

B. a specific person following

We implemented the proposed method on an actual robot to perform person tracking experiments. The detail of our system is described in Section II-A. The robot's speed and acceleration are restricted, and the actual average of speed was about 0.3 [m/s]. The target person whom the robot follows wears the clothes shown in Fig. 11.

Figure 14 shows an experimental result of a specific person following. Each circle shows a tracking result of each person, and the person identified as the following target is shown by red circle. A yellow square shows that a new person was detected at the frame, and a blue square shows that a candidate of person was rejected by using SVM. Figure 15 shows snapshots of the experiment. The robot successfully followed the specific person even when other people with a similar clothing (like a person shown with yellow/blue circle) exist near the target person. When the robot missed the target person because of occlusion (#151–158) or failure

Fig. 14. Experimental result of a specific person following with a mobile robot.

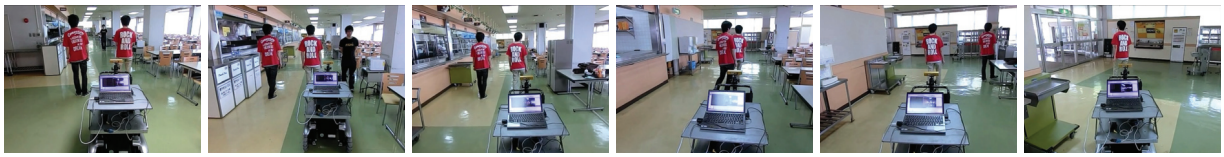


Fig. 15. Snapshots of experiment.

of identification (#202), the robot moves toward the target person's previous position. Since the target person was found again, the robot was able to continue following the person. The processing time of the identification per frame was about 120 [ms] in the case where one person exists in the image, and about 230 [ms] in two persons' case.

V. Conclusions

In this paper, we proposed a person identification technique using the SIFT feature for a mobile robot which performs specific person following. We made the appearance model for various body directions, and set the distance-dependent threshold to cope with the decrease of the number of SIFT feature matches according to the increased distance. Experimental results showed that the proposed method is able to identify the person even when other people with a similar clothing exist near the target person. Using the method, the robot successfully followed a specific person in the cafeteria.

For more robust identification, it is necessary to additionally use other sensors such as a laser range finder or other personal features such as the height or gait patterns.

Acknowledgments

A part of this research is supported by JSPS KAKENHI 23700203 and NEDO Intelligent RT Software Project.

References

[1] D. Beymer and K. Konolige, "Tracking People from a Moving Platform," *Proc. 2002 Int. Symp. on Experimental Robotics*, pp. 234–244, 2002.

[2] A. Howard, L. Mathies, A. Huertas, M. Bajracharya, A. Rankin, "Detecting Pedestrians with Stereo Vision: Safe Operation of Autonomous Ground Vehicles in Dynamic Enviroments," *Proc. 13th Int. Symp. of Robotics Research*, 2007.

[3] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, "Moving Obstacle Detection in Highly Dynamic Scene," *Int. Conf. on Robotics and Automation*, pp. 56–63, 2009.

[4] A. Ess, B. Leibe, K. Schindler, and L. V. Cool, "Object Detection and Tracking for Autonomous Navigation in Dynamic Environments," *Int. J. of Robotics Research*, vol. 29, no. 14, pp. 1707–1725, 2010.

[5] J. Miura, J. Satake, M. Chiba, Y. Ishikawa, K. Kitajima, and H. Masuzawa, "Development of a Person Following Robot and Its Experimental Evaluation," *Proc. 11th Int. Conf. on Intelligent Autonomous Systems*, pp. 89–98, 2010.

[6] J. Satake and J. Miura, "Robust Stereo-Based Person Detection and Tracking for a Person Following Robot," *Proc. IEEE ICRA-2009 Workshop on People Detection and Tracking*, 2009.

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[8] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] J.-M. Morel and G. Yu, "ASIFT: A New Framework for Fully Affine. Invariant Image Comparison," *SIAM J. Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[10] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, pp. 381-395, 1981.