International Journal of Automation and Computing

10(5), October 2013, 438-446
DOI: 10.1007/s11633-013-0740-y

Visual Person Identification Using a Distance-dependent Appearance Model for a Person Following Robot

Junji Satake Masaya Chiba Jun Miura

Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi 441-8580, Japan

Abstract: This paper describes a person identification method for a mobile robot which performs specific person following under dynamic complicated environments like a school canteen where many persons exist. We propose a distance-dependent appearance model which is based on scale-invariant feature transform (SIFT) feature. SIFT is a powerful image feature that is invariant to scale and rotation in the image plane and also robust to changes of lighting condition. However, the feature is weak against affine transformations and the identification power will thus 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 will decrease as the person becomes farther away from the camera. Therefore, we also use a distance-dependent threshold. The person following experiment was conducted using an actual mobile robot, and the quality assessment of person identification was performed.

Keywords: Mobile robots, image processing, intelligent systems, identification, scale-invariant feature transform (SIFT) feature.

1 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 who is carrying his/her items. This research develops a person identification method for such a robot that can follow a specific user among obstacles and other people.

There have been a lot of works on person detection and tracking using various image features[1, 2]. HOG[3] is currently one of the most widely used features for visual people detection. The detection methods in consideration of occlusion have also been proposed[4, 5], but emphasis is put on detection performance rather than the processing speed. Moreover, the person detection methods which combine HOG and the distance information acquired using an RGB-D camera such as Microsoft Kinect sensor were also proposed[6−8]. Spinello and Arras[6] performed an experiment using fixed cameras in the lobby of an university canteen. Munaro et al.[7] showed an example of tracking result using a mobile robot in an exhibition. Kinect sensor, however, cannot be used under sunlight. Ess et al.[9, 10] 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. Their method exhibits a nice performance for complicated scenes where many pedestrians exist. However, it is still costly to be used for controlling a real robot. Frintrop et al.[11] proposed a visual tracker for mobile platforms, but their experiments were performed in only laboratory environments.

We built a mobile robot system with a stereo camera and a laser range finder[12], 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 real environment where many ordinary people exist, it is important to distinguish the target person from ordinary people who wear various clothing.

Color histogram is widely used for person identification by a mobile robot[13−15]. Zajdel et al.[13] proposed a method in consideration of slow illumination changes by using a local trajectory of color feature. However, when the lighting conditions change, it is difficult to distinguish the person correctly by using the color-based method. The methods of person identification using not only clothes but also face images[14] or gait patterns[16] were proposed. However, the robot system which follows a person from behind cannot often work well with those methods.

In this paper, we propose a method of identifying a person based on the pattern of clothing using the scale-invariant feature transform (SIFT) feature. We make the appearance model from various body directions, and set a distance-dependent threshold to cope with the decrease of the number of SIFT feature matches due to the increased distance.

The organization of this paper is as follows. We describe our previous tracking system and its problems in Section 2. In Section 3, we propose a SIFT feature-based person identification method. In Section 4, we implement the proposed method on an actual robot to perform person tracking experiments. Finally, we conclude this paper and discuss future work in Section 5.

2 Person following robot

2.1 Stereo-based person tracking

2.1.1 Depth template-based person detection

To track persons stably with a moving camera, we use

Manuscript received March 29, 2013; revised July 18, 2013
Part of this work was supported by JSPS KAKENHI (No. 23700203) and NEDO Intelligent RT Software Project.

depth templates[17, 18], which are the templates for human upper bodies in depth images (see Fig. 1). We made the templates manually from the depth images where the target person was at 2 m away from the camera. A depth template is a binary template, the foreground and the background values are adjusted according to the status of tracking and input data.

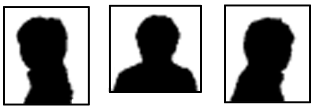
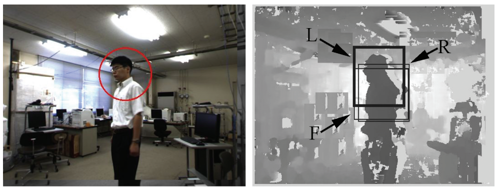


Fig. 1 Depth templates

For a person being tracked, his/her scene position is predicted using the Kalman filter. Thus, we set the foreground depth of the template to the predicted depth of the head of the person. Then, we calculate the dissimilarity between a depth template and the depth image using a sum of squared distances (SSD) criterion.

To detect a person in various orientations, we use the three templates simultaneously and take the one with the smallest dissimilarity as the matching result. An example of detection using the depth templates is shown in Fig. 2. We set a detection volume to search in the scene, its height range is 0.5 m– 2.0 m and the range of the depth from the camera is 0.5 m– 5.5 m.



(a) Input image (b) Depth image

Fig. 2 Person detection result

2.1.2 SVM-based false rejection

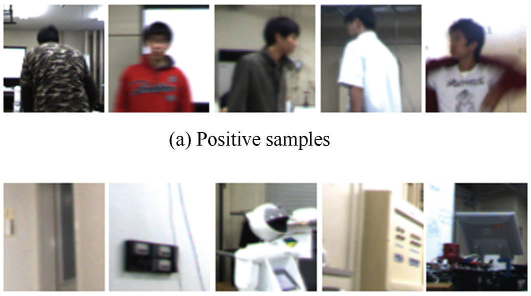
A simple template-based detection is effective in reducing the computational cost, but at the same time may produce many false detections for objects with similar silhouettes to the person. To cope with this, we use an SVM-based person verifier.

We collected many person candidate images detected by the depth templates, and manually examined if they are correct. 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[3] 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.

2.1.3 EKF-based tracking

We adopt the extended Kalman filter (EKF) for robust data association and occlusion handling[17]. The state vector $\boldsymbol{x} = [X\ Y\ Z\ \dot{X}\ \dot{Y}]^{\mathrm{T}}$ 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 each 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.



(a) Positive samples

(b) Negative samples

Fig. 3 Training samples for the SVM-based verifier

2.2 Configuration of our system

Fig. 4 shows our mobile robot system[12] which is composed of

1) A computer-controllable electric wheelchair (Patrafour by Kanto Auto Works Ltd.);

2) A stereo camera (Bumblebee2 by Point Grey Research);

3) A laser range finder (UTM-30LX by Hokuyo);

4) A Note PC (Core2Duo, 2.66 GHz, 3 GB memory).

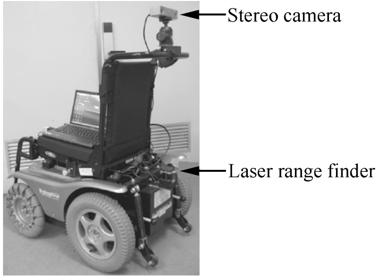


Fig. 4 A mobile robot with a laser range finder and a stereo camera

Fig. 5 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:

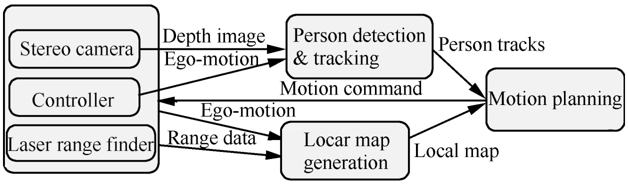


Fig. 5 Configuration of the system

1) The 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 Section 2.1.

2) The 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) The motion planning module calculates a safe robot motion which follows a specified target person and avoids others, using a randomized kinodynamic motion planner.

To develop and maintain the module-based software system, we use the RT-middleware[19] environment where each software module is realized as an robot technology (RT) component. The robot repeats the following steps: 1) Person detection and tracking and local map generation; 2) Motion planning; 3) Motion execution. The cycle time is set to 500 ms.

2.3 Problems of the previous system

Fig. 6 shows snapshots of a person following experiment at a cafeteria. Fig. 7 shows an example of the recognition and the planning result. From the stereo data (see Fig. 7 (a)), the robot detected two persons, the target on the left and the other on the right (see Fig. 7 (b)). Fig. 7 (c) shows the result of environment recognition and motion planning. We tested the system for the cases where three persons exist near the robot. Problems which became clear in the experiment are described below.

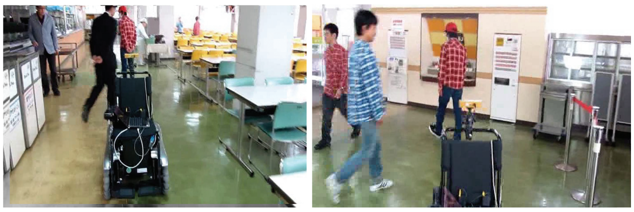
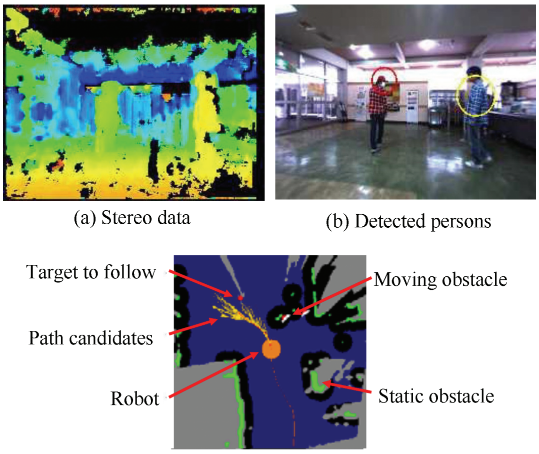


Fig. 6 Snapshots of a specific person following at the cafeteria



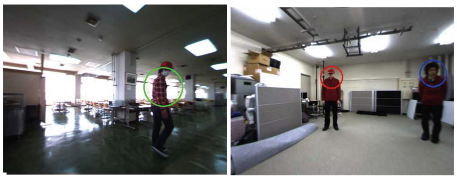
(a) Stereo data

(b) Detected persons

(c) Recognition result and path planning

Fig. 7 An example of environment recognition and motion planning

Fig. 8 (a) shows the failure of target identification using color due to the bad illumination. Fig. 8 (b) is an example 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.



(a) A bad illumination

(b) Cannot distinguish the target and another person with the same color

Fig. 8 Failure of target person identification

3 A SIFT feature-based person identification

Our previous person identification method using only color information is weak against 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[20] which uses the texture of clothing as a cue.

SIFT[21] 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 against affine transformations. Although a feature which increases the robustness to affine transformations, ASIFT[22] was proposed, the identification power will be degraded when the pose of the person changes largely. Therefore, we 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 will decrease as the person becomes farther away from the camera. Therefore, we use a distance-dependent threshold.

3.1 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, false corresponding points are also contained in matching. Therefore, false corresponding points are removed as follows using random sample consensus (RANSAC)[23]:

Step 1. Four pairs are randomly selected from the group of the corresponding points.

Step 2. A homography matrix is calculated based on the selected corresponding points.

Step 3. The number of corresponding points which satisfies the above homography matrix out of all pairs is counted.

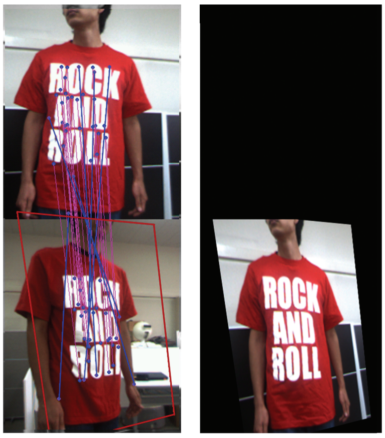
Step 4. By repeating Steps 1 to 3, the homography matrix with the maximum number of pairs is selected.

An example of homography estimated by using RANSAC is shown in Fig. 9. Fig. 9 (a) shows a correspondence between an model image (upper) and an input image (lower).

The brown quadrangle shows a form of the model image transformed by the estimated homography matrix. Fig. 9 (b) shows the transformed model image. Each pair of points connected by the pink line shows the correspondence judged as the inlier which satisfies the homography matrix, and the one connected by the blue line shows the outlier. By using RANSAC, 40 correspondences were divided into 34 correct ones (inliers) and 6 false ones (outliers). We use for person identification only the corresponding points judged as the inlier.

Fig. 10 shows the results of matching in different situations. Even when the lighting conditions were changed (Figs. 10 (a) and (b)), the feature points were able to be matched by using the SIFT feature. Furthermore, some of the features were able to be matched even when the patterns were deformed by the wrinkles of clothing (Fig. 10 (c)).

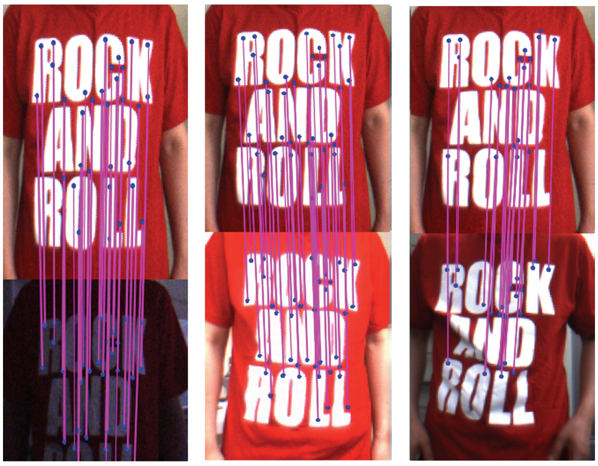
Since we used a wide angle camera, we were not sometimes able to discern the stripe/check pattern of clothing when the distance is large. Therefore, we used the clothing on which large characters were printed.



(a) Estimation of homography

(b) Transformed model image

Fig. 9 Estimation of homography by using RANSAC



(a) Dark situation (32 matches)

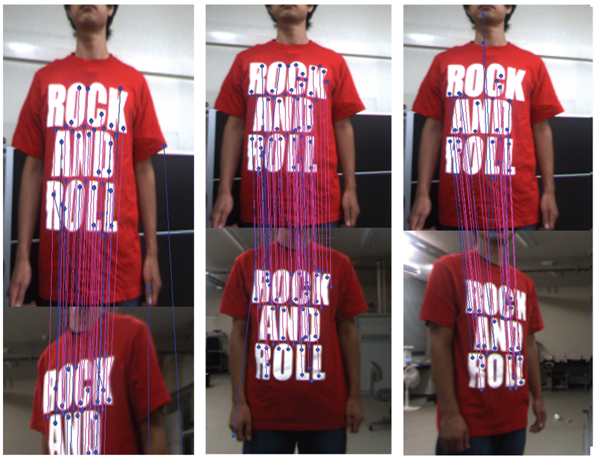
(b) Bright situation (38 matches)

(c) Wrinkles (22 matches)

Fig. 10 Results of matching in different situations

3.2 The appearance model

For person identification, we make the appearance model which is a set of SIFT features extracted from several model images. Fig. 11 shows the matching results of SIFT features between one of the model images and input images taken from different directions. For a frontal image, 52 matches were obtained (Fig. 11 (b)). On the other hand, the number of matches decreased for the different directions (Figs. 11 (a) and (c)). In order to cope with the pose changes, we make the appearance model with the following procedure (see Fig. 12) which uses several model images taken from various body directions:

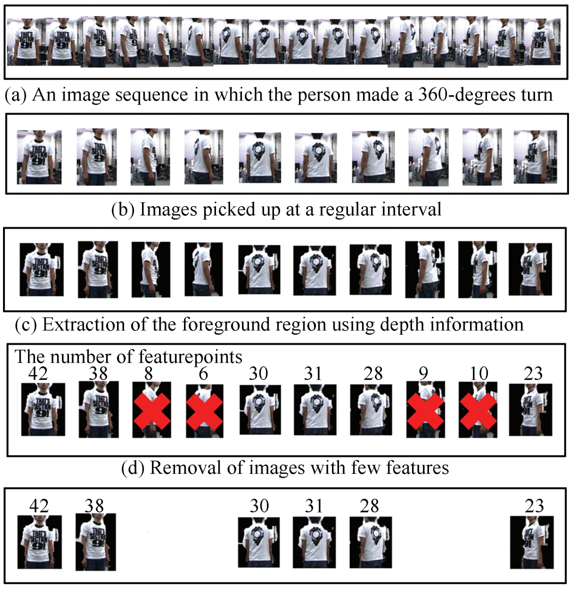


(a) The number of matches = 30

(b) The number of matches = 52

(c) The number of matches = 30

Fig. 11 Relations between change of the body direction and the number of SIFT feature matches (upper: model image, lower: input images)



(a) An image sequence in which the person made a 360-degrees turn

(b) Images picked up at a regular interval

(c) Extraction of the foreground region using depth information

(d) Removal of images with few features

(e) The appearance model

Fig. 12 The procedure of the appearance model generation for various body directions

Step 1. An image sequence is recorded, in which the person makes a 360-degrees turn at 1 m away from the camera.

Step 2. A certain number of images (in this paper, we set the number to 30) 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 the images in the sequence.

Step 3. In order to remove the feature points in the background region, each image is segmented into the foreground/background regions using depth information. We classified the pixels with a depth value of 1±0.5 m into the foreground region.

Step 4. SIFT features are 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 experiment.

Step 5. As the appearance model, we use the set of SIFT features extracted from the images selected by the above steps. The selected images are called model images.

3.3 A distance-dependent threshold

The number of SIFT feature matches will decrease as the distance from the camera to the person increases. The images in the upper right corners in Fig. 13 show the model images taken when the distance between the person and the camera is 1 m. The dashed 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. The appearance model with the threshold is called a distance-dependent appearance model.

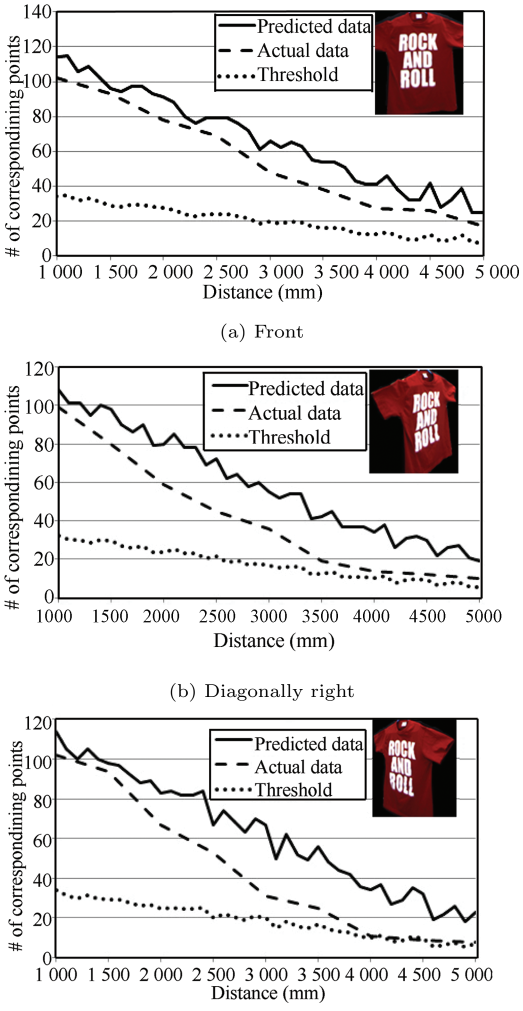
It is tedious to actually obtain the person images taken at various distances. Instead, we simulate the increasing distance by reducing the size of the model image for generating a simulated input image, and predict the effect of increasing distance. Considering the changes of lighting condition and wrinkles, we use 30 % of the predicted value as a threshold. Here, when the distance is 2 m, the error of estimated distance by using the stereo camera (f=2.5 mm, baseline 12 cm, image size 512 × 384) is about 2 cm. Since the influence of the error is small, we think that it can be disregarded.

The examples of three directions are shown in Fig. 13. The solid line shows the number of matches predicted by the simulation. It can read that the predicted value (solid line) and the actual value (dashed line) have a similar tendency. The dotted line shows a distance-dependent threshold. This threshold is calculated for each model image.

3.4 Identification of the target person

3.4.1 Representative images to estimate rough direction

When identifying the target person, matching an input image with all model images is costly. To reduce the calculation cost, we estimate a rough direction using a certain number of representative images (in this paper, the number is set to six with consideration of the processing speed). The representative images are chosen in advance from the model images. The best selection of the image set is the combination which can cover images of any body directions. Therefore, we choose an image set from which the largest number of corresponding points for each input image can be obtained.



(a) Front

(b) Diagonally right

(c) Diagonally left

Fig. 13 Distance-dependent appearance model

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

$$\max_{i} m_{ij}. \tag{1}$$

The set of representative images is denoted as S. The best selection of the set makes the following formulas the maximum:

$$\mathop{\mathrm{argmax}}_{S} \sum_{j} \max_{i \in S} m_{ij}. \tag{2}$$

Fig. 14 shows an example of six representative images selected using this method.

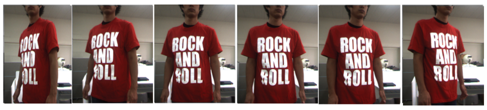


Fig. 14 An example of representative images

3.4.2 Processing of identification

Fig. 15 shows the relationship between the number of matches and the processing time. When the number of SIFT feature points in an input image is 80, our system needs about 20 ms for matching them to those in each model image. Because it is costly to compare all model images at each frame, the model images used for the comparison are selected according to the situation as follows:

1) If there is the model image matched with the previous frame, only three images (the same direction and neighbors) are used for matching.

2) Otherwise, after estimation of rough direction using the representative images described in Section 3.4.1, two images of neighbors of the estimated direction are used for matching.

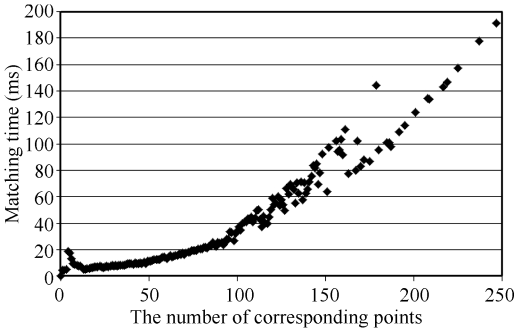


Fig. 15 Relationship between the number of matches and the processing time

In this paper, the orientation of person estimated by the EKF-based tracking has not been used for the identification. The orientation of upperbody may not accord with the direction of movement.

The person is judged to be the target to follow if the number of matches between input and model (Fig. 13 dashed line) is over the threshold (Fig. 13 dotted line). In other word, the person is judged as the target when the following evaluation value (matching score) is over 1.

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

When there are more than one target candidate, the person with the highest matching score is selected as the target to follow. In order to deal with the situation where another person with the same clothing exists, it will be necessary to use the trajectory information.

4 Experimental results

4.1 Verification of the robustness to direction change

We made the appearance models for five kinds of clothing. We selected similar clothes intentionally. The identification experiment was conducted on image sequences in which the person made a 360-degrees turn at 1.0 m, 1.5 m, 2.0 m, and 2.5 m away from the camera, respectively. The images without sufficient number of SIFT features were deleted from this evaluation.

The identification result about each model is shown in Table 1. Acceptance rates in Table 1 is the number of the images identified as the clothing of the appearance model among the number of images in the test sequence.

Table 1 Acceptance rates for various body directions

(a) When the distance is 1.0 m

Appearance model (made at 1 m)	Test data set (turned at 1.0 m)				
	0.933	0	0	0	0
	0	**0.900**	0	0	0
	0	0	**0.943**	0	0
	0	0	0	**0.957**	0
	0	0	0	0	**0.926**

(b) 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**

(c) 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**

(d) 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**

The detail of an identification result when the same clothing is tested as a model is shown in Fig. 16. When the matching score is 1 or more, the person in the input image is the same as the registered person. In this case, input images at #11 and #32 were rejected and the acceptance rate is 25/27 = 0.926. Note that input images at #12–21 and #31–42 were not used for the evaluation because the body became sideways mostly and a sufficient number of features was not detected. The target person, however, was identified almost correctly when the pattern of clothing was observed. We think that motion blur and wrinkle of clothing caused failures of identification.

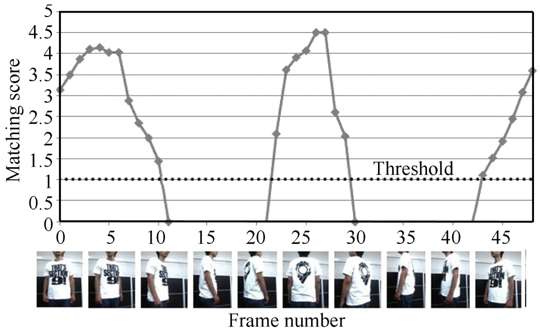


Fig. 16 The detail of an identification result

4.2 Matching results when occlusion occurs

Matching results when the target person is occluded by other person is shown in Fig. 17. The target person was standing at 1.5 m away from the camera ($X_1 = 1500$, $Y_1 = 0$). The other person was 1 m away from the camera nd moved from the left to the right every 100 mm ($X_2 = 1000$, $Y_2 = [-1000, 1000]$). Note that the person is judged as the target when the number of corresponding points is more than 28.8 (see Fig. 13 (a)).

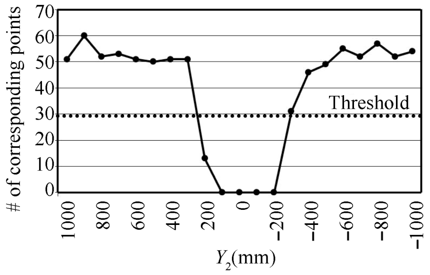
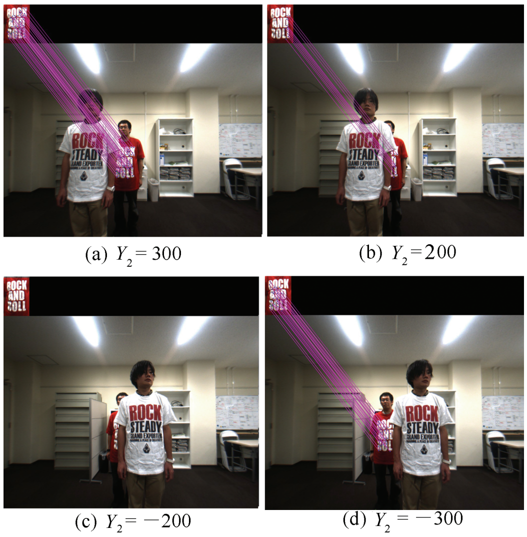


Fig. 17 Matching results when the target parson is occluded ($X_1 = 1500$, $Y_1 = 0$, $X_2 = 1000$, $Y_2 = [-1000, 1000]$)

The examples of the matching results are shown in Fig. 18. When the target person was not occluded yet ((a) $Y_2 = 300$), sufficient feature points were matched. When the clothing was almost occluded ((b) $Y_2 = 200$), only 13 features were matched. Since the number of corresponding points is smaller than the threshold value, the person was not judged as the target. When the person was occluded completely, no feature points were matched. When the person appeared again ((d) $Y_2 = -300$), sufficient corresponding points were obtained. Even when the clothing was occluded partially, it was able to identify the person as the target.


(a) $Y_2 = 300$ (b) $Y_2 = 200$

(c) $Y_2 = -200$ (d) $Y_2 = -300$

Fig. 18 Examples of the matching results

Fig. 19 shows the example of matching when multiple persons exist and the target person is occluded partially. The target person was matched correctly out of the multiple persons. We present the specific person following in the situation where the target person is occluded by the other person in the following subsection.

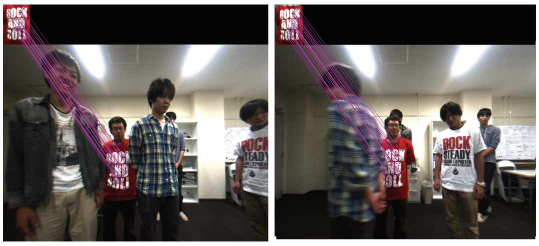


Fig. 19 Matching results when multiple persons exist

4.3 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 2.2. 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. 14.

Fig. 20 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 the red circle. A yellow square shows that a new person was detected at that frame, and a blue square shows that a candidate of the target person was rejected by using SVM. Fig. 21 shows snapshots of the experiment. The robot successfully followed the specific person even when other people with a similar color clothing (like a person shown with yellow/blue circles) exist near the target person. When the robot missed the target person temporarily

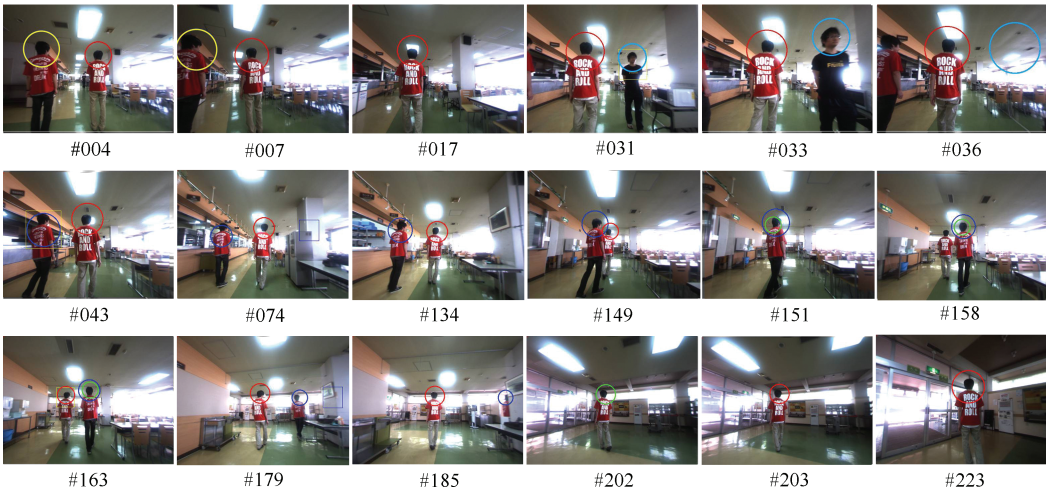


Fig. 20 Experimental result of a specific person following with a mobile robot

Fig. 21 Snapshots of experiment

because of occlusion (#151–158) or failure of identification (#202), the robot moved toward the target person′s position predicted by the EKF-based tracking. Since the person was again identified as the target to follow at #163 and #203, the robot was able to continue following the person.

The processing time of the identification (including SIFT feature extraction and matching) per frame was about 120 ms in the case where one person exists in the image, and about 230 ms in two persons′ case. In this experiment, the identification process was performed for all persons in each frame. However, identification is unnecessary when the target person is isolated from the others. In addition, we will implement tracking and identification process using multithreaded program, since the identification process is not necessary for all frames.

5 Conclusions

In this paper, we proposed a person identification method using 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 color 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.

References

[1] D. Beymer, K. Konolige. Tracking people from a mobile platform. In *Proceedings of the 8th International Symposium on Experimental Robotics*, Springer, Berlin, Heidelberg, Germany, 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 environments. In *Proceedings of the 13th International Symposium of Robotics Research*, Hiroshima, Japan, pp. 26–29, 2007.

[3] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, San Diego, USA, pp. 886–893, 2005.

[4] X. Wang, T. X. Han, S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Proceedings of IEEE 12th International Conference on Computer Vision*, IEEE, Kyoto, Japan, pp. 32–39, 2009.

[5] S. Tang, M. Andriluka, B. Schiele. Detection and tracking of occluded people. In *Proceedings of British Machine Vision Conference*, BMVC, Guildford, UK, pp. 9.1–9.11, 2012.

[6] L. Spinello, K. O. Arras. People detection in RGB-D data. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, San Francisco, USA, pp. 3838–3843, 2011.

[7] M. Munaro, F. Basso, E. Menegatti. Tracking people within groups with RGB-D data. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Algarve, Portugal, pp. 2101–2107, 2012.

[8] J. Salas, C. Tomasi. People detection using color and depth images. In *Proceedings of the 3rd Mexican Conference on Pattern Recognition*, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 127–135, 2011.

[9] A. Ess, B. Leibe, K. Schindler, L. V. Gool. Moving obstacle detection in highly dynamic scenes. In *Proceedings of*

the 2009 IEEE International Conference on Robotics and Automation, IEEE, Kobe, Japan, pp. 56–63, 2009.

[10] A. Ess, B. Leibe, K. Schindler, L. V. Cool. Object detection and tracking for autonomous navigation in dynamic environments. *International Journal of Robotics Research*, vol. 29, no. 14, pp. 1707–1725, 2010.

[11] S. Frintrop, A. Konigs, F. Hoeller, D. Schulz. A component-based approach to visual person tracking from a mobile platform. *International Journal of Social Robotics*, vol. 2, no. 1, pp. 53–62, 2010.

[12] J. Miura, J. Satake, M. Chiba, Y. Ishikawa, K. Kitajima, H. Masuzawa. Development of a person following robot and its experimental evaluation. In *Proceedings of the 11th International Conference on Intelligent Autonomous Systems*, IAS, Ottawa, Canada, pp. 89–98, 2010.

[13] W. Zajdel, Z. Zivkovic, B. J. A. Krose. Keeping track of humans: Have I seen this person before? In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, IEEE, Barcelona, Spain, pp. 2081–2086, 2005.

[14] N. Bellotto, H. Hu. A bank of unscented Kalman filters for multimodal human perception with mobile service robots. *International Journal of Social Robotics*, vol. 2, no. 2, pp. 121–136, 2010.

[15] G. Cielniak, T. Duckett. Person identification by mobile robots in indoor environments. In *Proceedings of the 1st IEEE International Workshop on Robotic Sensing*, IEEE, Orebro, Sweden, 2003.

[16] D. Cunado, M. S. Nixon, J. N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, vol. 90, no. 1, pp. 1–41, 2003.

[17] J. Satake, J. Miura. Robust stereo-based person detection and tracking for a person following robot. In *Proceedings of IEEE ICRA-2009 Workshop on People Detection and Tracking*, IEEE, Kobe, Japan, 2009.

[18] J. Satake, J. Miura. Person following of a mobile robot using stereo vision. *Journal of Robotics Society of Japan*, vol. 28, no. 9, pp. 1091–1099, 2010. (in Japanese)

[19] N. Ando, T. Suehiro, K. Kitagaki, T. Kotoku, W. K. Yoon. RT-middleware: Distributed component middleware for RT (robot technology). In *Proceedings of 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Alberta, Canada, pp. 3933–3938, 2005.

[20] J. Satake, M. Chiba, J. Miura. A SIFT-based person identification using a distance-dependent appearance model for a person following robot. In *Proceedings of 2012 IEEE International Conference on Robotics and Biomimetics*, IEEE, Guangzhou, China, pp. 962–967, 2012.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] J. M. Morel, G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[23] M. A. Fischler, 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, no. 6, pp. 381–395, 1981.

Junji Satake received the B. Eng., M. Eng. and Ph. D. degrees in information engineering from Okayama University, Okayama, Japan in 1998, 2000, and 2003, respectively. From 2003 to 2004, he was a researcher at Okayama University, Okayama, Japan. From 2004 to 2008, he was an expert researcher at the National Institute of Information and Communications Technology, Japan. Since 2008, he has been with Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan, first as a research associate and later as an assistant professor. He is a member of IEEE, RSJ, IPSJ, and IEICE.

His research interests include pattern recognition, computer vision, and human computer interaction.

E-mail: satake@cs.tut.ac.jp (Corresponding author)

Masaya Chiba received the B. Eng. and M. Eng. degrees in information engineering from Toyohashi University of Technology, Toyohashi, Japan in 2010 and 2012, respectively. He is a member of RSJ.

His research interests include intelligent robotics and pattern recognition.

E-mail: chiba@aisl.cs.tut.ac.jp

Jun Miura received B. Eng. degree in mechanical engineering in 1984, M. Eng. and Ph. D. degrees in information engineering from the University of Tokyo, Tokyo, Japan in 1986 and 1989, respectively. From 1989 to 2007, he was with Department of Mechanical Engineering, Osaka University, Suita, Japan, first as a research associate and later as an associate professor. In 2007, he became a professor of Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan. From 1994 to 1995, he was a visiting scientist at Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA. He received the Best Paper Award from the Robotics Society of Japan in 1997. He was also selected as one of the six finalists for the Best Paper Award at the 1995 IEEE International Conference on Robotics and Automation. He is a member of IEEE, AAAI, RSJ, JSAI, IPSJ, IEICE, and JSME.

His research interests include intelligent robotics, computer vision, and artificial intelligence.

E-mail: jun@cs.tut.ac.jp