2014 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)
September 11-13, 2014. Evanston, Illinois, USA

Tracking of Multiple Humans Using Subtraction Stereo and Particle Filter*

Takehiro KAWASHITA,[1] Masatoshi SHIBATA,[1] Gakuto MASUYAMA,[2] and Kazunori UMEDA[2]

***Abstract*— This paper proposes a method for the automatic tracking of multiple humans in various scenes using a stereo camera. The proposed method detects candidate regions of humans using "subtraction stereo," which restricts stereo matching to foreground regions extracted by subtraction and obtains distance information for those regions. Tracking of humans is carried out for the extracted regions using a particle filter. The particle filter consists of four steps, prediction, calculation of likelihood, data association, and resampling. Three features: distance, color, and direction of motion are used in the proposed method to achieve robust human tracking. When humans with similar clothing colors or human direction pass each other, occlusion occurs, and tracking often fails. The proposed method adds robustness to occlusions by explicitly considering the distance, color, and direction of human motion. The proposed method has been evaluated through experiments using a stereo camera that simulates a surveillance camera in real scenes. Tracking accuracy of more than 90% has been achieved in three different scenes, which shows the effectiveness of the proposed method for tracking humans.**

I. INTRODUCTION

Many surveillance cameras are installed in towns because of an increasing consciousness of crime prevention and a reduction in costs of the cameras recently. The role of a surveillance camera is to detect and track suspicious people or to measure a stream of people. Surveillance cameras currently in use can only verify details after an incident. Therefore, a prompt response cannot be made. However, if real-time supervision were possible, prompt response could be taken. For this reason, many researchers have focused on real-time human detection and tracking from a camera image [1][2][3][4][5].

Recording human movements in real-time is an important consideration in using surveillance cameras. However, pedestrians are often occluded by each other in a video surveillance system. Such occlusion is a difficult problem. Many studies deal with such occlusion problems, and there have been several approaches. The first approach is to reduce the occluded regions by using multiple sensors[6][7]. For operating a surveillance camera, this approach is unsuitable because of the installation costs. The second approach is to prevent the occlusion of humans by installing the camera on the ceiling with a bird's-eye view from just above the people. However, the problem with this approach is the restricted installation location. The third approach is to use a stereo camera that can obtain 3D information simply[8][9]. Because a stereo camera can obtain 3D information, it can robustly estimate human motion. In addition, the positioning of a stereo camera is not limited. Therefore, we decided to use a stereo camera.

In this paper, we will present a practical system without limitations of placement that uses one stereo camera. Human tracking is implemented by particle filter. Distance, color and direction features are utilized to improve robustness against occlusion.

Figure 1 shows the configuration of the system. First, humans are detected and distance information is obtained using subtraction stereo[10]. Second, human tracking is performed by applying a particle filter to the detected humans. In the proposed method, particles and humans are associated by using positions in the world coordinate system X, Y and directions that are obtained from a stereo camera, and the colors of the humans. When the target has been discriminated, we can track it robustly by using these three evaluations; distance, direction, and color.

This paper is organized as follows. In Section 2, we explain the acquisition of foreground information, which is obtained by subtraction stereo and shadow detection. In Section 3, we present a multi-human tracking method using a particle filter. In Section 4, we show the experimental result for the proposed method. Section 5 is the conclusion and future works.

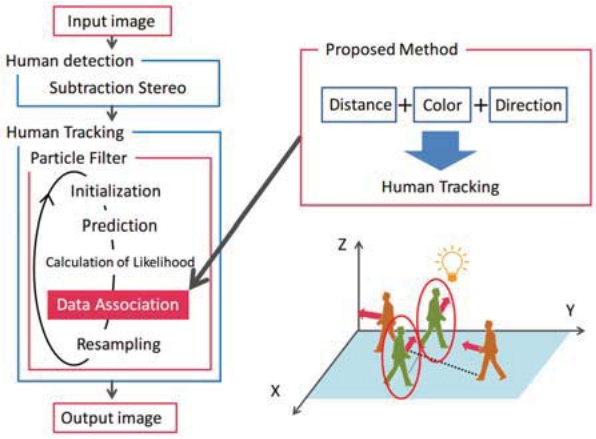


Fig. 1. Flow of proposed human tracking method

[1]Takehiro KAWASHITA and Masatoshi SHIBATA are with the Course of Precision Engineering, School of Science and Engineering, Chuo University {kawashita, shibata}@sensor.mech.chuo-u.ac.jp

[2]Gakuto MASUYAMA and Kazunori UMEDA are with the Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University {masuyama, umeda}@mech.chuo-u.ac.jp

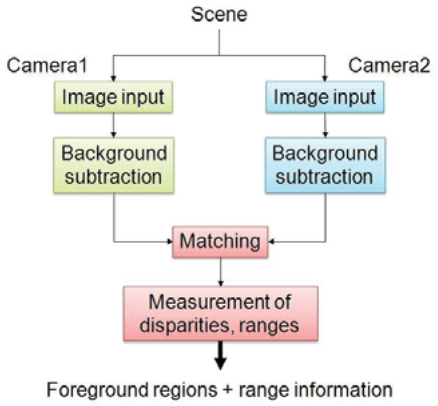
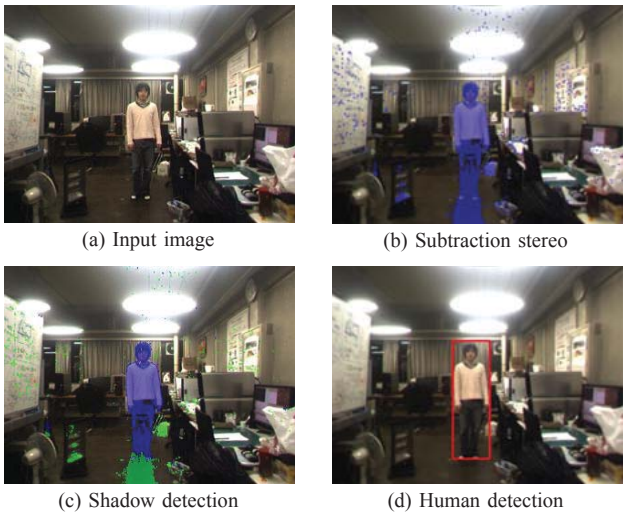


Fig. 2. Basic algorithm of the subtraction stereo



(a) Input image (b) Subtraction stereo

(c) Shadow detection (d) Human detection

Fig. 3. Flow of foreground detection

II. Human Detection

A. Subtraction stereo

The basic algorithm of the subtraction stereo is shown in Fig. 2. This algorithm is based on a background subtraction method and stereo matching. First, subtraction stereo extracts foreground objects by the background subtraction method. Next, stereo matching is applied to only the extracted regions. Therefore, the processing regions for stereo matching are restricted to the foreground, and we can obtain the foreground region and 3D information with less computational time. An example of an output image using subtraction stereo is shown in Fig. 3(b). The foreground region contains distance information.

B. Shadow detection

Shadow detection is used to refine the foreground. The image obtained using subtraction stereo includes non-human regions affected by the shadow. This non-human region seriously affects the projection plane. When we define $I_{(x,\ y)}$ as the intensity of the pixel located in the 2-D image position (x, y) and $I'_{(x,\ y)}$ as the intensity of the background pixel, the equation for shadow evaluation is described as

$$\theta_{(t+1,x,y)} = \begin{cases} \alpha\Psi_{(x,y)} + \beta\Lambda_{(x,y)} \\ \qquad +(1-\alpha-\beta)\theta_{(t,x,y)}, & \text{if } \frac{I_{(x,y)}}{\eta} < I'_{(x,y)} \\ \infty, & \text{otherwise.} \end{cases} \quad (1)$$

where $\theta_{(t+1,\ x,\ y)}$ corresponds to a shadow value. This value will be applied a threshold to determine if a pixel is a shadow. A small shadow value corresponds to a shadow point. The functions Ψ and Λ show the degree of difference in color between pixels and within a pixel. α, β, and η are constant weights of textures, colors, and intensity, respectively, which are determined empirically in our experiments. The details of this method are explained in [12]. The result of the shadow detection is shown in Fig. 3(c), and the result of the human detection is shown in Fig. 3(d). We define the central point of the detected rectangle of Fig. 3(d) as the center point of the human target.

III. Human Tracking

In the proposed method, a human is tracked by a particle filter that tracks a human center point detected by subtraction stereo. This method consists of four steps: prediction, calculation of likelihood, data association, and resampling. When the data association processing is performed, we can achieve robust human tracking by using 3D distance, human motions obtained from the stereo camera, and color features. The details of each process are described below.

A. Particle filter

(i) Initialization

When a human is detected, particles are spread around the human's center point.

(ii) Prediction

Prediction processing predicts the 3D positions of each particle from the past motions of the target. Prediction processing changes, depending on whether data association has succeeded or failed.

- if data association was succeeded

Prediction is performed by assuming a uniform linear motion, and a random number must be added to the prediction value in order to track a human's random motion.

- if data association was failed

Prediction is performed by assuming a uniform linear motion without adding a random number. This process prevents the particles from being diffused by a random number when the tracker did not find the target.

(iii) Calculation of likelihood

The likelihood of each particle is calculated between the human center point and each particle's point by the Euclidean distance on world coordinates X, Y.

$$L = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{d^2}{2\sigma^2}\right), \quad (2)$$

(a) Initialization (b) Prediction

(c) Calculation of likelihood & Data association (d) Resampling

Fig. 4. Flow of particle filter

where L corresponds to likelihood. The function d shows the Euclidean distance between the human center point and the tracker's weighted center point, and σ shows a standard deviation.

(iv) Data association

A tracker searches for the target on each frame. When searching for the target, we use three features: distance, direction of motion, and color. If the target is found, the tracker and the target are associated. We will explain these three features in section C.

(v) Resampling

The resampling process is performed by selecting reasonable particles for tracking. Selection is performed based on the likelihood of each particle. Through this processing, particles are concentrated around the human center point. Therefore, as tracking continues, we can track stably. If tracking is successful for a certain period of time, an identification number (ID) is assigned to the tracker.

B. Method of data association

We explain three evaluation values; distance, direction of motion, and human color, used in data association processing and how data association processing is carried out.

(i) feature of distance

The evaluation value of the distance is calculated by comparing the Euclidean distance on world coordinates, X, Y, between the human center point and the tracker's weighted center point. When we define D_d as the evaluation value of the distance, the equation for the evaluation of distance is described as

$$D_d = \sqrt{(X_p - X_h)^2 + (Y_p - Y_h)^2} \tag{3}$$

X_p and Y_p show the tracker's weighted center point, and X_h and Y_h show the human center point.

(ii) feature of direction

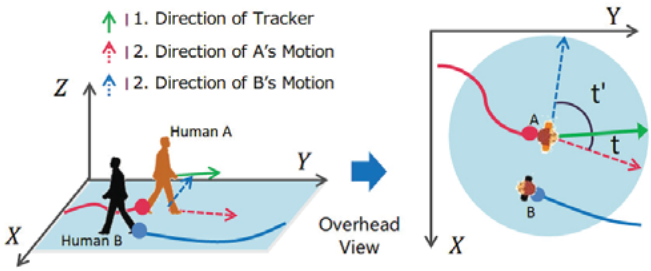


Fig. 5. Example of calculation human directions

The evaluation value of the direction is calculated by comparing the following two values:

1) Directions of the tracker
2) Directions of each human in the frame.

First, we explain the way to calculate the human directions. The direction of 1 is calculated by a tracker's position that is associated with the human's positions in previous frames.

The direction of 2 is calculated before data association processing; therefore, each human's position is unidentified. For that reason, the direction of 2 is calculated by comparing a human center point with the weighted center point of the tracker within the fixed distance.

An example of the direction calculation is shown in Fig.5. Fig. 5 shows directions related only to the red tracker. The red arrow is calculated by the red tracker's position (direction of 1). The green dotted line arrow is calculated by assuming that the green human was associated with the red tracker in earlier frames (direction of 2). The green dotted line arrow is calculated by assuming that the black human was associated with the red tracker in previous frames (direction of 2). The figure on the right is the overhead view of the figure on the left, and the arrows of the right and left figures correspond. Second, we determine whether to use the calculated direction. For example, a human who is walking would not change direction significantly. On the other hand, it would be difficult to predict human directions when human is stopping. For that reason, calculated direction is used as the evaluation value only if the direction is higher than the threshold.

Last, we calculate the difference of the angles between the direction of 1 and the direction of 2 and define t as the calculated difference of the angles (Fig. 5 : t and t'). In addition, when we define D_a as the evaluation value of the distance, the equation for the evaluation of the direction is described as

$$D_a = \begin{cases} k^t - 1, & (t < t_{thr}) \\ k, & (otherwise), \end{cases} \tag{4}$$

where k shows the weight which is determined empirically in our experiments. t_{thr} shows the threshold. We set the t to 1.5 and the t_{thr} to 1.22[rad].

(iii) feature of color

The evaluation value of color is calculated by comparing the human's hue histogram with the tracker's hue histogram,

which is the associated human's histogram on the previous frame. The hue histogram is calculated from pixels included in the blue areas of Fig. 3(c) and whose saturation is more than the threshold. When we define D_c as the evaluation value of color, the equation for the evaluation of shadow is described as

$$D_c = \sqrt{1 - \sum_{u=1}^{m} \sqrt{p_u q_u}}, \tag{5}$$

where p and q show the normalized hue histogram, u shows the number of the hue, and m shows the amount of the hue.

Summing up these three evaluation values, we define D as the final evaluation value to the tracker. A tracker discriminates a target who has minimum D on each frame, and the tracker is associated with the target. After associating the tracker with the target, the tracker updates the 3D position and color information associated with the targeted human.

IV. Experimental result

A. Experimental condition

In this section, to verify the validity of our proposed method, we carried out experiments in three different scenes. Scene 1 and Scene 2 are staged experiments made to resemble real-life scenes. Scene 3 is an actual street scene. We see in Table I the camera conditions at the time of the experiments, and experimental sequences are shown in Fig.6.

- Scene 1 : Hall
- Scene 2 : Open space
- Scene 3 : AKIBA SQUARE (Event site)

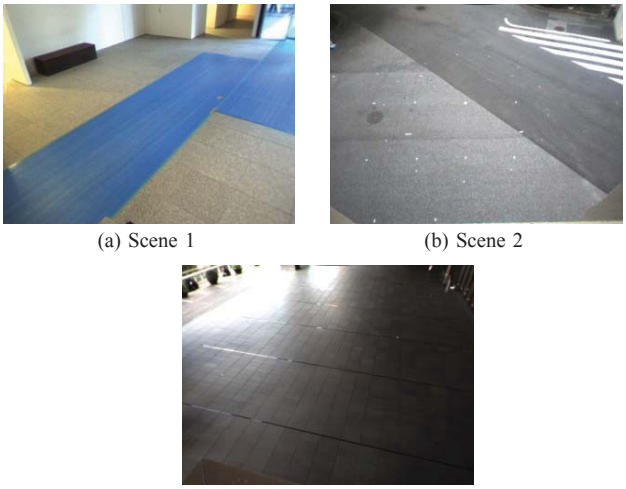
We used a Point Gray Research Bumblebee2 camera with 320×240 pixel resolution in these experiments. These videos were obtained at a rate of 20 frames per second. The results were obtained using an Intel Core2 Duo CPU, 2.93 GHz with 3 GB RAM. All experimental sequences were taken using a static stereo camera, and we used 500 particles for each person in our experiments.

TABLE I
Camera installation conditions

	Height [m]	Tilt angle [°]
Scene 1	2.4	30
Scene 2	6.4	60
Scene 3	5.9	40

B. Evaluation

In order to verify the effectiveness of the proposed method for human tracking, we experimented in three different scenes and compared using *"distance (D) and color (C) features"* with using *"distance (D), direction (A), and color (C)"*. The output images using distance, direction, and color features in each scene are shown in Fig. 7 (a)〜(c). The top figures of Fig. 7 (a)〜(c) are input images, and the bottom figures are output images. Colored rectangles in the bottom figures show the results of human detection and are color coded according to the ID to improve visibility. Colored circles in the bottom figures show the trackers (weighted center point of particles). Color coding is carried out for the same reason as with the rectangles. White circles show that the ID is not yet a distributed ID.



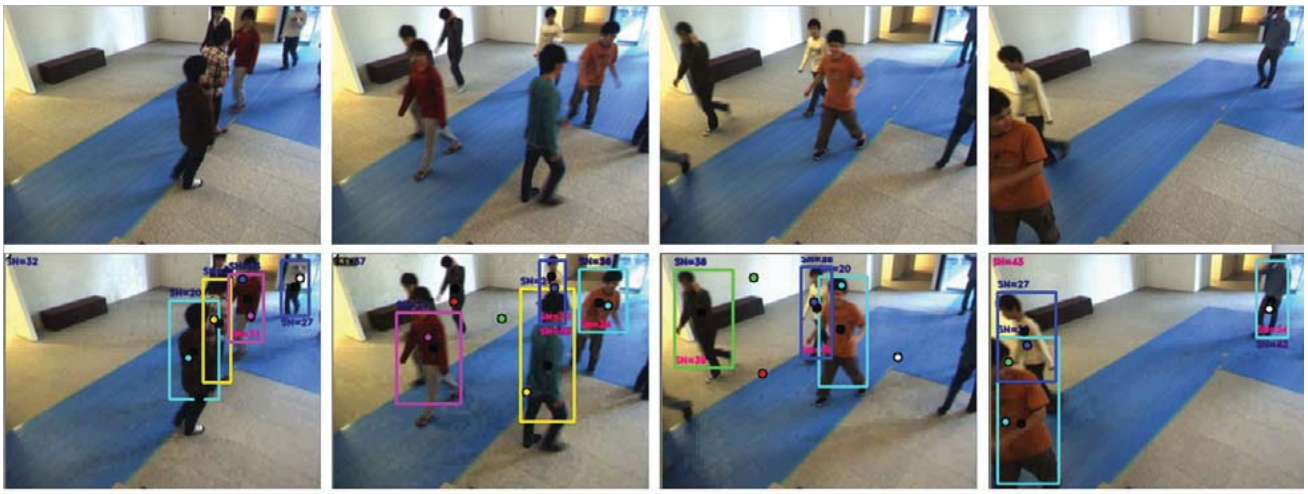
(a) Scene 1 (b) Scene 2

(c) Scene 3

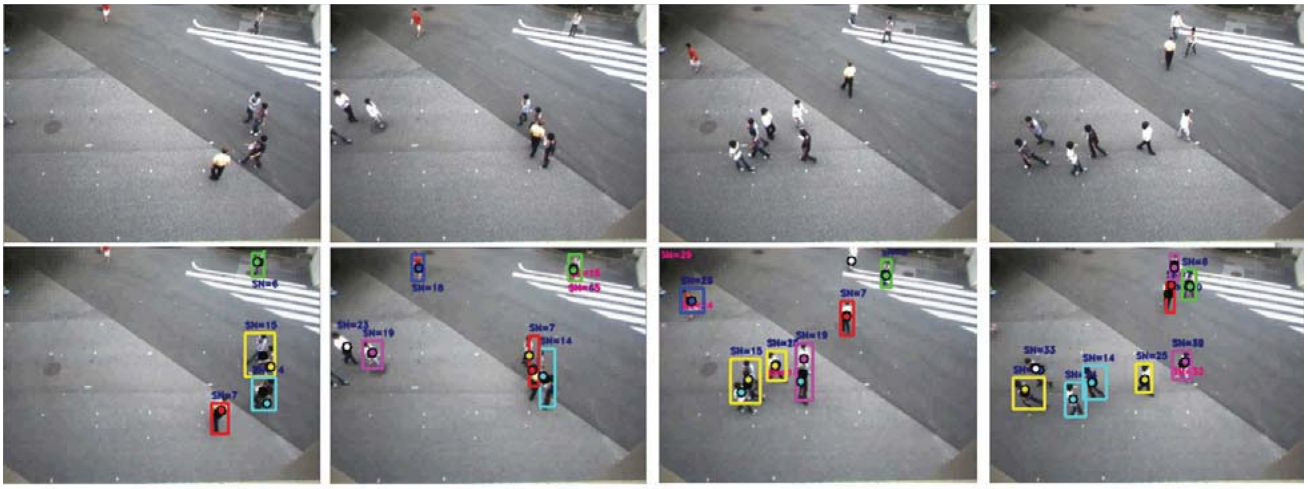
Fig. 6. Experimental Scenes

Table II (a)〜(c) shows the results of our analysis. As an evaluation of tracking accuracy, if the ID does not change from the ID given to the human until beyond the screen, the tracking is deemed successful. As is evident in Table II, our method can track robustly by using the distance, direction, and color features. Distance feature is reliable for tracking unless the target is occluded by other human. When the target is not detected due to occlusion, distance feature's reliability for tracking is decreased. When humans with similar clothing colors pass each other, the direction feature is especially effective. When humans are walking in a similar direction, the color feature is especially effective. We can tell each evaluation value worked effectively in each scene. In scene 1, using distance and color only, people were tracked incorrectly when they passed each other on the right side of the camera. However, when using direction, change of ID did not occur. In scenes 2 and 3, evaluation using color was ineffective due to the influence of shadows and humans with similar clothing colors. However, we can use direction to track correctly. The group of people walking to the left and right passed each other between t = 52 s and t = 63 s in Fig. 7(b). In this scene, human tracking failed when using only distance and color. In contrast, by using distance, direction, and color, human tracking was successful. Additionally, information can be processed at 9〜10 fps; therefore, we can process online.

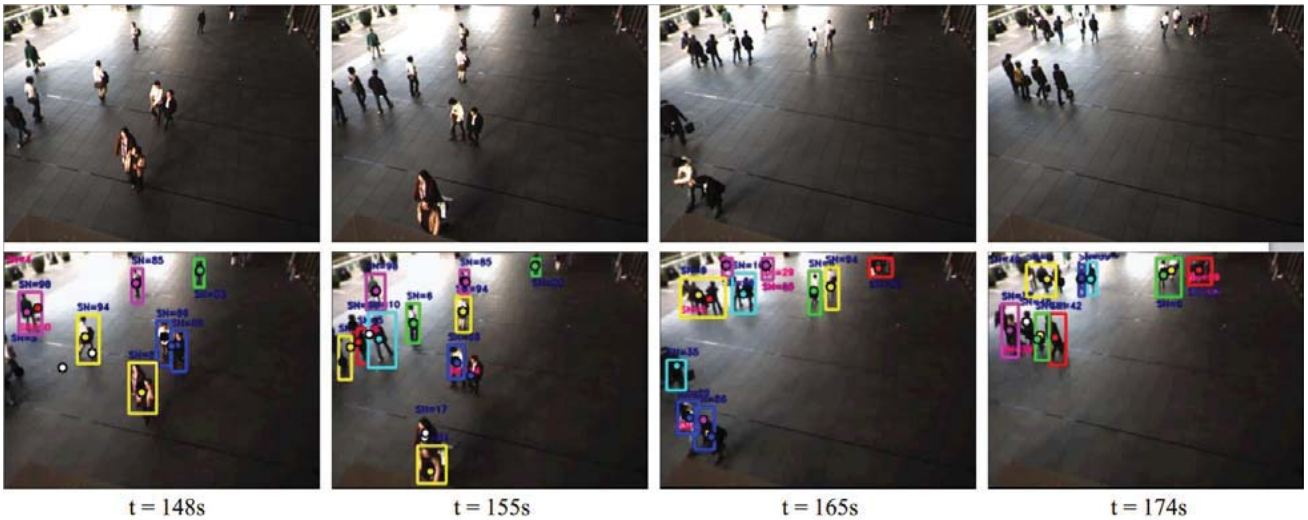
On the other hand, failure tracking occurred. When a human is near the camera, as in scene 1, the duration in which the human cannot be detected grows longer, and the distance between the positions of the tracker and the target grows larger while the human is not detected. Therefore, even if using distance, direction, and color, tracking failures can occur. The tracking failure is shown in Fig. 8. It shows



(a) Scene 1



(b) Scene 2



(c) Scene 3

Fig. 7. Snapshots of human tracking

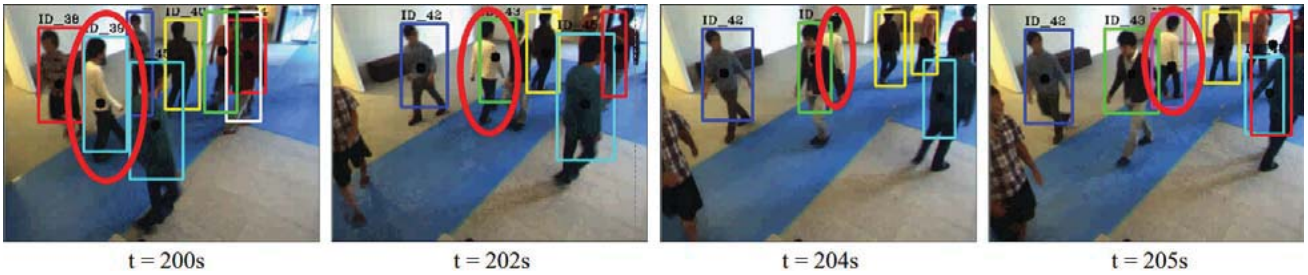


Fig. 8. Failure of human tracking

the system's failure to track the human who is enclosed in a red oval. This human is not detected after t = 202 s. By t = 205 s, this human is detected again; however, the ID has changed. The human center point, the target, deviates from the correct point because another human is detected together with the first human; therefore tracking fails. At this point, we must improve the system's decision as to whether human detection is performed correctly.

TABLE II
EXPERIMENTAL RESULTS

	Number of detected people	Tracking accuracy (D & C)[%]	Tracking accuracy (D & C & A) [%]
Scene 1	40	88	90
Scene 2	35	88	94
Scene 3	51	86	92

V. CONCLUSION

In this paper, we have presented a multiple-human tracking system that uses subtraction stereo for human detection and a particle filter using distance, human direction, and color for human tracking. We performed experiments in three scenes to demonstrate its robustness to occlusions. One scene is an actual environment. The others are staged to mirror an actual environment. With these experiments, we achieved tracking accuracy of 90% or more in all scenes, which demonstrates the effectiveness of the proposed method for human tracking.

Tracking fails when predicted positions deviate from a target's position after failing to continue detecting the human. Our method can track correctly even if the prediction position deviates slightly from the target's position. However, the longer human detection fails, the more difficult it becomes to carry out human tracking correctly. In the future, we will continue to work to improve tracking accuracy. First, we will improve prediction processing when human detection is missed. Second, we will improve human detection to introduce a feature-based object classification[13] using subtraction stereo. Third, we will weight each evaluation value to depend on each situation rather than simply adding all evaluation values. Last, we will experiment with a greater variety of situations and verify the effectiveness of the proposed method for human tracking.

Acknowledgment. This work was in part supported by The Precise Measurement Technology Promotion Foundation, Japan.

REFERENCES

[1] W. Lu, K. Okuma, J. J. Little, Tracking and Recognizing Actions of Multiple Hockey Players Using the Boosted Particle Filter, Image and Vision Computing, vol. 27, no. 1-2, pp. 189-205, 2009.
[2] L. Huchuan, Z. Ruijuan, C. Yen-Wei, Head Detection and Tracking by Mean-shift and Kalman Filter, Innovative Computing Information and Control, p. 357, 2008.
[3] M. Liebens, T. Sakiyama, J. Miura, Visual Tracking of Multiple Persons in a Heavy Occluded Space Using Person Model and Joint Probabilistic Data Association, Multisensor Fusion and Integration for Intelligent Systems, pp. 547-552, 2006.
[4] M. Ronald, A Theoretical Foundation for the Stein-Winter "Probability Hypothesis Density (PHD)" Multitarget Tracking Approach, Proc. of the 2000 MSS National Symposium on Sensor and Data Fusion, 2002.
[5] R. Kurazume, H. Yamada, K. Murakami, Y. Iwashita, T. Hasegawa, Target Tracking Using SIR and MCMC Particle Filters by Multiple Cameras and Laser Range Finders, Proc. of the 2008 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp.3838-3844, Oct. 2008.
[6] S. Haraguchi, K. Hukushi, I. Kumazawa, Person Tracking Using Multiple Stereo Cameras with Color Information, IEICE Trans., vol. 109, no. 471, pp. 229-234, 2010.
[7] K. Nakamura, H. Zhao, R. Shibasaki, K. Sakamoto, T. Ohga, N. Suzukawa, Tracking Pedestrians Using Multiple Single-Row Laser Range Scanners and Its Reliability Evaluation, IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, vol. J88-D-II, no. 7, pp. 1143-1152, 2005.
[8] P. Kang-Il, P. Chan-Ik, L. Jangmyung, Vision Based People Tracking Using Stereo Camera and CamShift, Intelligent Robotics and Applications, vol. 8102, pp. 145-153, 2013.
[9] J. Kovacevic, S. Juric-Kavelj, I. Petrovic, An Improved CamShift Algorithm Using Stereo Vision for Object Tracking, MIPRO 2011 Proc. of the 34th International Convention, pp. 707-710, 2010.
[10] K. Umeda, Y. Hashimoto, T. Nakanishi, K. Irie, K. Terabayashi, Subtraction Stereo–A Stereo Camera System That Focuses on Moving Regions, Proc. of the 2009 SPIE-IS & T Electronic Imaging, vol. 7239, Three-Dimensional Imaging Metrology, 723908, 2009.
[11] M. -T. Yang, K. -H. Lo, C. C. Chiang, W. -K. Tai, Moving Cast Shadow Detection by Exploiting Multiple Cues, Image Proc. of the 2008, IET, vol. 2, pp. 95-104, 2008.
[12] A. Moro, K. Terabayashi, K. Umeda, E. Mumolo, Auto-adaptive Threshold and Shadow Detection Approaches for Pedestrians Detection, Proc. of the 2009 AWSVCI, pp. 9-12, 2009.
[13] T. Mitsui, Y. Yamauchi, H. Fujiyoshi, Object Detection by Two-Stage Boosting with Joint Features, IEICE Trans., vol. J92-D, no. 9, pp. 1591-1601, 2009.
[14] A. A. Gorji, R. Tharmarasa, T. Kirubarajan, Performance Measures for Multiple Target Tracking Problems, Information Fusion (FUSION), pp. 1-8, 2011.
[15] T. Ubukata, K. Terabayashi, A. Moro, K. Umeda, Multi-Object Segmentation in a Projection Plane Using Subtraction Stereo, Proc. of the 20th International Conference on Pattern Recognition, pp. 3296-3299, 2010.