Masuyama *et al. Robomech J (2017) 4:24*
DOI 10.1186/s40648-017-0092-4

ROBOMECH Journal

RESEARCH ARTICLE

Open Access

Complementary human detection and multiple feature based tracking using a stereo camera

Gakuto Masuyama[1*], Takehiro Kawashita[2] and Kazunori Umeda[1]

Abstract

A human detection and tracking method using a stereo camera is presented in this paper. Two human detection methods are independently implemented, and the results are combined to reduce misdetection and nondetection. The tracking step is based on particle filtering. In the data association phase, we introduce three features: distance, traveling direction, and color. The color feature is obtained from every segmented detection window. Eligibility and co-occurrence of the blocks provide robustness to occlusion. The proposed method is tested by using measurement data at the entrance of a building, where occlusion is frequently observed. The experiments demonstrate improved tracking performance over standard particle filtering.

Keywords: Stereo camera, Human tracking, Human detection

Introduction

The ability to detect and track multiple humans has been of major interest to the computer vision community. Many applications are associated with human tracking systems, such as surveillance, marketing, and robotics use. One of the difficulties in human tracking is occlusion caused by pedestrians and other objects in the real world. They temporarily cause a target person to disappear within the scene, and as a consequence, complicate tracking.

Various approaches have been presented for tracking multiple humans. One work [1] integrated soft biometrics with an appearance model in a single-camera setting. Generating coarse representation of a target using soft biometrics, they demonstrated tracking that was robust to changing poses and illumination. However, performance in a heavily occluded environment is not provided. Satake et al. used a stereo camera [2]. The study utilizes silhouette templates to realize fast tracking; however, the number of overlapping persons can be no more than three. RGB-D data-based tracking using three Kinect sensors is presented by Luber et al. [3]. The method has advantages in that it does not require a ground plane assumption. Although Kinect would provide reliable measurements, its measurable range is shorter than that of cameras, and it is hard to apply in outdoor environments. Tseng also applies Kinect; however, the sensor is hung from the ceiling [4]. This approach can avoid the occlusion problem; therefore, it is computationally efficient. Such methods have limitation in that they require the large cost of setting up the measurement system. Another approach is to use multiple sensors to reduce the occluded areas [5]. This is promising but prone to requiring large computational costs and effort to implement.

In this paper, we propose a novel human tracking method using a single stereo camera, which has advantages over the previous studies in (1) robustness to occlusion and (2) ease of installation. Our motivation is to build human measurement system that requires less cost to its users. Therefore, fewer pieces of equipment (light weight and small size) are desirable to compose the system. However, a single sensor cannot avoid the problem of occlusion.

An overview of the proposed method is depicted in Fig. 1. The procedure can be separated into two steps:

*Correspondence: masuyama@mech.chuo-u.ac.jp
[1] Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
Full list of author information is available at the end of the article

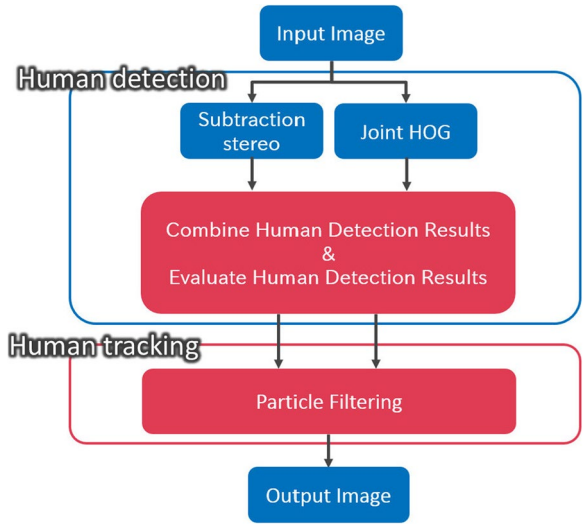

Fig. 1 Overview of proposed method. The figure represents flow of detection and tracking of the proposed method

human detection and tracking. In the human detection step, humans are detected using two methods independently. The two detection results are complementary based on their reliability in terms of color and distance information. In the human tracking step, humans are tracked by particle filtering. We introduce three features for data association: distance, traveling direction, and color features. Color features are obtained based on the relationships among the separated blocks of the detected window. Main contribution of this paper is a set of techniques for a stereo camera to combine two (potentially multiple) detection methods and particle filtering.

This paper is organized as follows. In "Detection" section, the human detection step is presented. The human tracking step is described in "Tracking" section. The validity of the proposed method is verified through experiments in "Experiments" section, and in "Conclusion" section, we conclude and discuss future works.

Detection

First, we briefly review subtraction stereo and Joint HOG-based human detection. For more detail about these methods, see [6]. After that, a combination of the two methods is presented in this section.

Subtraction stereo

We use subtraction stereo [7] to obtain the foreground region in input images and corresponding distance information (Fig. 2). First, background subtraction is applied to the two images captured by a stereo camera. Second, only the foreground regions are used for stereo matching. Third, shadows in the extracted images are removed [8]. Finally, we can obtain the foreground region and corresponding distance information. This method is useful in that it can reduce incorrect stereo matching and computational costs by limiting the region for stereo matching based on the color information of each camera.

Joint HOG-based detection

Another human detection method is based on Joint HOG features and boosting [9]. Joint HOG is represented by the co-occurrence of HOG features: effective combinations of features are determined using Real AdaBoost. Real AdaBoost is again applied to obtain classifiers using pooled joint features.

Integration

First, detection results obtained by the two methods are combined as shown in Fig. 3. By integrating the results obtained from two different approaches, nondetections are reduced. However, that would inevitably increase misdetections. Misdetections are then reduced by checking the validity of every detected window.

We find that misdetection occurs as (a) duplicate detections of one person, and (b) detection in non-human regions.

Duplicate detections

Using two different detection methods might cause duplications of detection results. In case a human is correctly detected by both methods, we have to integrate the two detection windows. Detection windows indicating an identical person must be remarkably similar in their positions. Therefore, we simply apply a thresholding process.

If the distance between the centers of two windows in an image coordinate is less than a threshold value, two windows are integrated into one detection window. First, the thresholding is independently conducted with respect to the two detection methods to reduce the duplicate detections with thresholding parameters thr_{SS} and thr_{JH}. After that the thresholding is again conducted for every pairs of remaining windows detected by the two methods with parameter thr_d. If the duplication occurs between the subtraction stereo and Joint HOG-based detection windows, we use the Joint HOG-based window in this paper.[1]

Detection in non-human regions

Detection windows are scored based on the degree of coincidence between the foreground and foreground

[1] It is not clear which method should be used to determine the representative window. We chose Joint HOG because the background adaptation for subtraction stereo is not implemented in the experiments. Interpolation of two windows considering features of the both method may be another research direction.

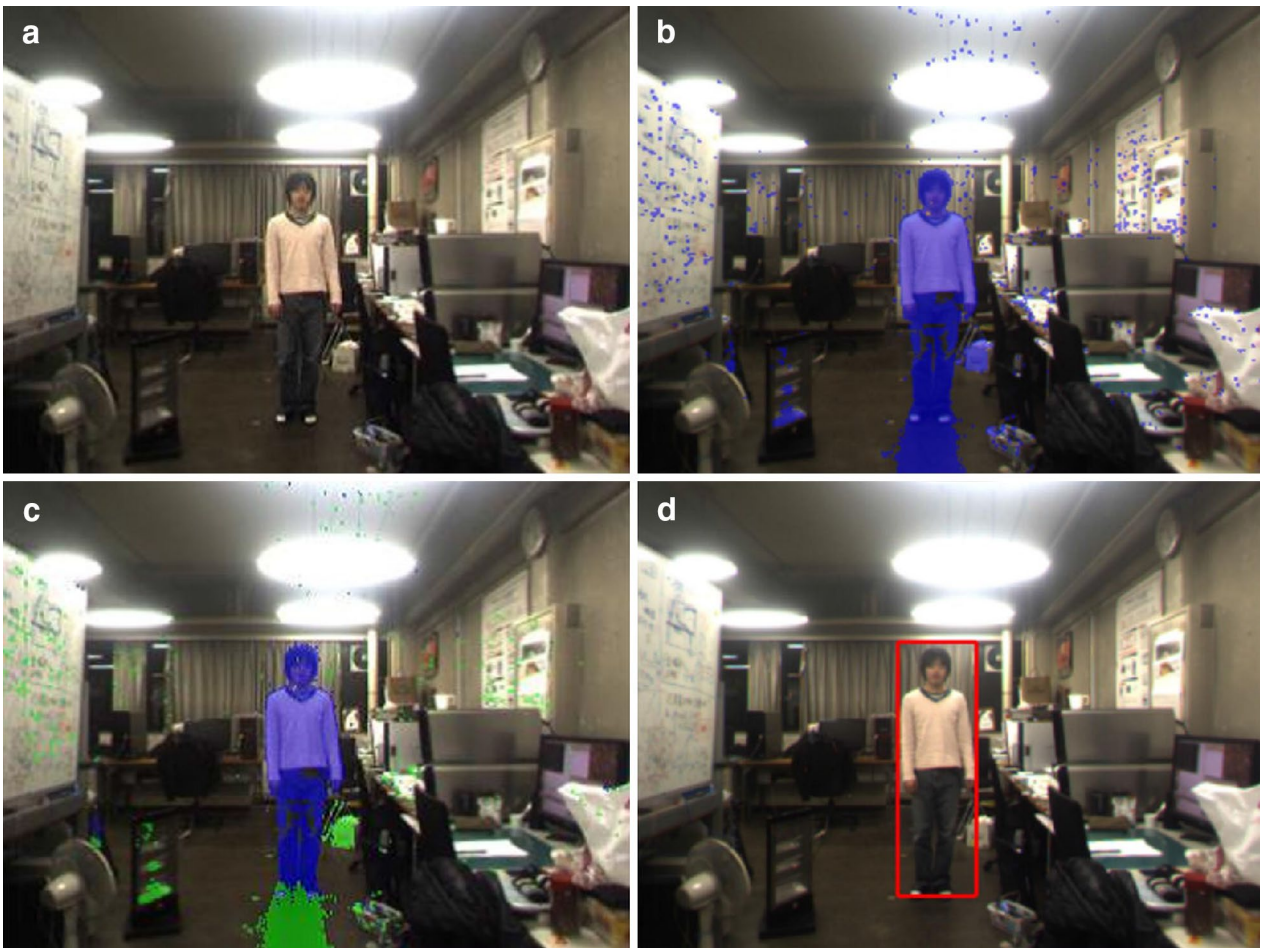

Fig. 2 Procedure of subtraction stereo. **a** Input image, **b** subtraction stereo, **c** Shadow removal, **d** foreground extraction

disparity regions. The foreground and foreground disparity regions are obtained from background subtraction using an image and a disparity image, respectively. If a human actually exists inside the window, it is expected that the corresponding foreground and foreground disparity regions tend to be observed in the center of the window. Therefore, a weighting coefficient for the score is introduced.

$$W_u = k - \left(u - \frac{w}{2}\right)^2, \tag{1}$$

where k is a parameter, and w is the width of the window. A center of uv coordinate is in the upper left of the window. Figure 4 illustrates the calculation of W_u.

The score S is then given by

$$S = \frac{\sum_{u=0}^{w-1}\sum_{v=0}^{h-1} W_u M_{uv}}{\sum_{u=0}^{w-1}\sum_{v=0}^{h-1} W_u (G_{uv} + D_{uv})}, \tag{2}$$

where G_{uv} takes 1 if the foreground region exists at (u, v) and takes 0 otherwise. D_{uv} is the same as G_{uv} but takes 1 for the foreground disparity region, and $M_{uv} = G_{uv}D_{uv}$. w and h are the width and height of the window, respectively. Figure 5 depicts an example of the foreground and foreground disparity, where red and blue rectangles represent illustrative target pixels. In the scene, M_{uv} takes 1 and 0 in red and blue region, respectively. Using the M_{uv} and other values, the score S is calculated. If the score S is under a threshold thr_S, the corresponding window is eliminated. After the elimination of misdetections, we obtain the detection results. The center position of each window is then accepted as input of the target-tracking step.

Tracking

A particle filter [10] is used to track multiple persons in the scene. To improve the robustness to occlusion, we introduce three features in the data association process:

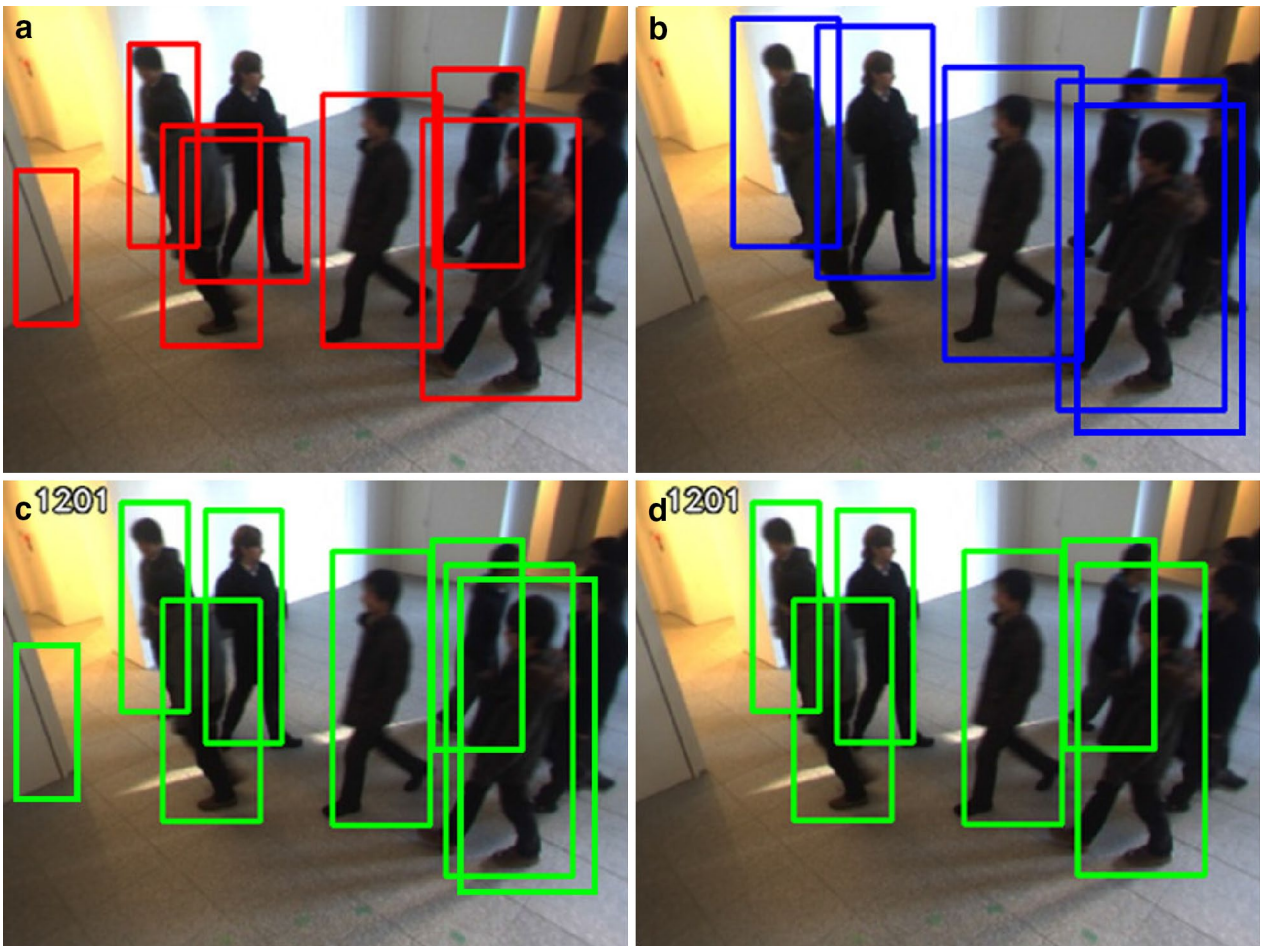

Fig. 3 Examples of detection. **a** Subtraction stereo, **b** joint HOG, **c** combined detection windows, **d** results after eliminating misdetections

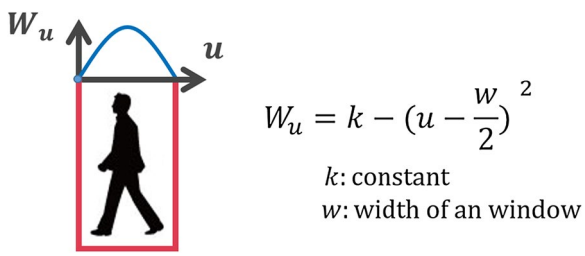

Fig. 4 Weight for detection score

distance, traveling direction, and color features. Especially, the color feature contributes to the improvement by utilizing time series of spatially segmented tracking data.

Data association in particle filtering

Particle filtering is one of the most common methods for solving the filtering problem. It suits requirements such as robustness to the movement of a target or non-Gaussian noise. The flow of a particle filtering is as follows.

In the tracking phase, we distribute particles for persons initially detected. Each particle transits its state based on the predicted position of the target in the current frame. The likelihood of each particle is then calculated using the distance between the targets and the particles in world coordinates. The current state of tracker is centroid of the weighted particles. In each frame, the association between the particle and the target is specified. In this paper, three features are used for this data association process. Particles are resampled based on their likelihood, and the above procedure is repeated. In the following, we introduce evaluation values into the data association process.

Distance

The first evaluation value is simply the distance between a human's center point and a tracker's weighted centroid in world coordinates:

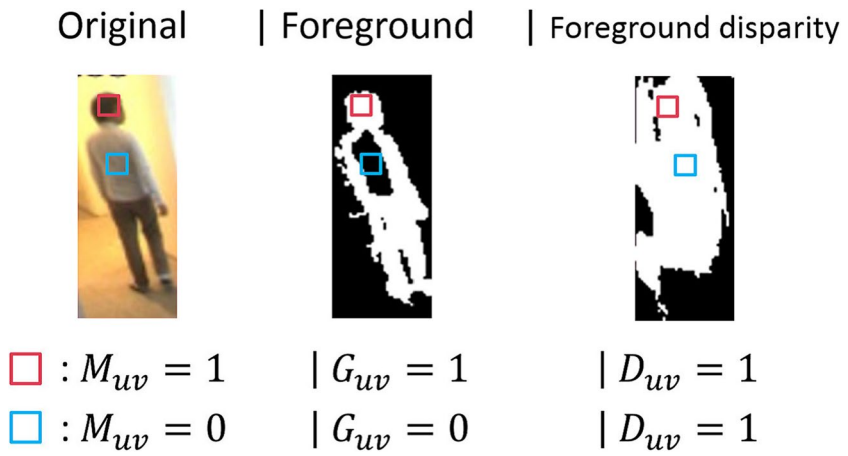

Fig. 5 Scoring based on foreground and foreground disparity images. M_{UV} represents logical product of G_{UV} and D_{UV}

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

The distance is obtained in overlooking plane, where X_p and Y_p represent the weighted centroid of the tracker, and X_h and Y_h represent the center point of the human.

Traveling direction

Another evaluation value is obtained from the difference of the traveling directions of (a) the particle calculated based on the history of movement, and (b) each human in each frame. Details about this evaluation value is presented in [11].

The direction of (a) is calculated from the history of movement of each particle. Note that the direction of (b) must be calculated within the data association process. Therefore, we cannot obtain each person's position at this point. Instead, we use the neighboring center point of the human and a weighted centroid of the tracker.

Figure 6 illustrates an example of related directions of humans and a tracker. The red arrow in the figure represents the direction of (a). The green arrow represents a candidate for direction (b), which is calculated based on the association of the tracker with the person colored green. The blue arrow is the same as the green one with a correspondingly colored human. If the norm of the traveling speed is less than a threshold, we remove the candidates.

A candidate's direction that has the smallest angle with direction a) is used to determine the evaluation value. The evaluation value of traveling direction D_a is

$$D_a = \begin{cases} A^{\alpha} - 1 & (\alpha < \alpha_{thr}) \\ A & (otherwise) \end{cases}, \quad (4)$$

where A and α_{thr} are the parameters. The parameters A and α_{thr} are set at 1.2 and 80°, respectively.

Color

The basic idea for the color feature is to segment detection windows into multiple blocks as shown in Fig. 7. We utilize eligibility and co-occurrence in each block. The segmentation is expected to contribute to improving reliability as compared to color features obtained from all information from one detection window.

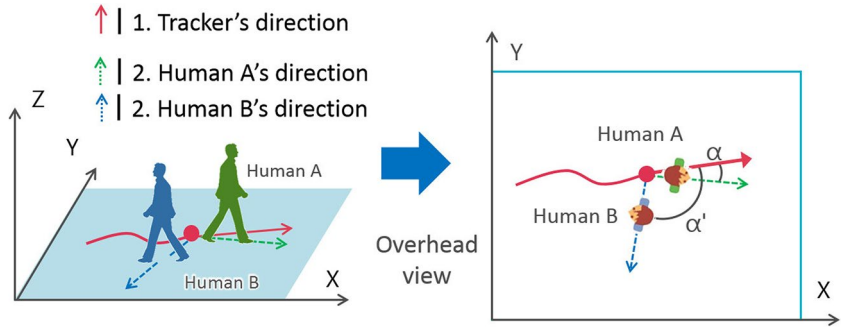

Fig. 6 Angular difference in the traveling directions of humans and particles

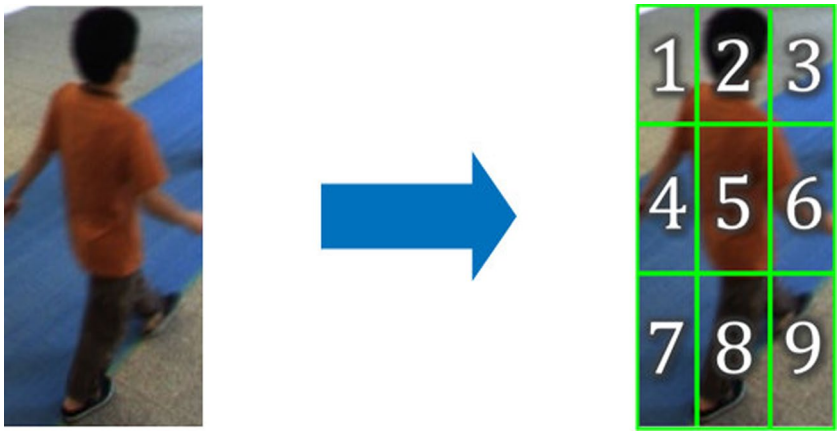

Fig. 7 Block segmentation

Hue and saturation, which are relatively robust to illumination changes, are used to measure the similarity of color information:

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

where L represents the Bhattacharyya coefficient, p and q are normalized histograms of hue and m represents the number of the hue value. Only pixels that have saturation values of more than a threshold are used to make the histograms in (5).

Eligibility of each block

The eligibility of each block is determined based on tracking data obtained online. An example of the calculated weight W_{s_i} for the ith block is depicted in Fig. 8.

First, we assume that a block showing less variation is more eligible (blocks 2, 3, 5, and 7 in Fig. 8). The stationarity of the block should indicate whether the corresponding color information is specific for the target person. Enhancing the specific feature and ignoring the background or unstable color information are expected to improve the reliability of the color information.

L_i' in Fig. 8 represents the averaged difference of the Bhattacharyya coefficient between the ith block and the corresponding mean block image over the samples. Thus, large L_i indicates that the block provides a constantly observable feature. W_{s_i} represents the weight for L_i, which is the Bhattacharyya coefficient of the ith block in the current frame:

$$W_{s_i} = \frac{L_i'}{\sum_i L_i'}. \tag{6}$$

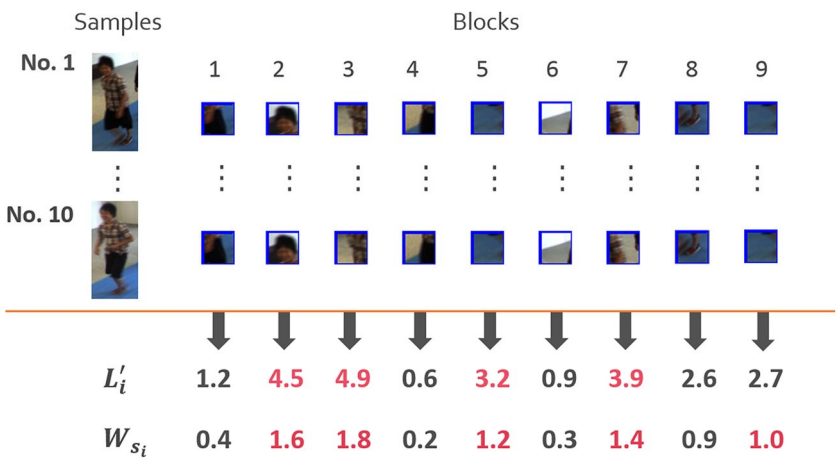

Fig. 8 Weight for each block. Images on the left are samples captured from tracking data. In this figure, detection windows are segmented into nine blocks

At the data association step, the evaluation value with respect to the eligibility of each block is then defined as $W_{s_i} L_i$.

Importance of color information

The evaluation value of each target candidate is further enhanced by the co-occurrence of eligible blocks. After the above procedure, we can find which block is relatively reliable for target tracking. A number of similar blocks in a candidate window with the eligible blocks could indicate the reliability of identification. We then weight the evaluation value of the color information D_c using the number of matched blocks c. The weight is defined as $W_c = c^2$ in this paper.

Finally, the evaluation value of the color feature is determined as

$$D_c = \sqrt{1 - W_c \sum_{i=1}^{n} W_{s_i} L_i}. \tag{7}$$

Occlusion handling

The validity of the evaluation value of the color feature in (7) is related to the occlusion problem. If the sample images in "Eligibility of each block" section include occluded frames, the eligibility might not work as intended. We then detect occluded blocks to make the process robust. The detection is simply executed by counting overlapping blocks as shown in Fig. 9.

If the number of occluded blocks is more than 30% of the total number of blocks in the window, the sample is rejected. In the calculation of (7), occluded blocks are ignored if the number of occluded blocks is more than 50% of the total number. The evaluation value of color information is then normalized by the number of accepted blocks.

Search for tracking targets

The whole evaluation value for data association D is then determined as

$$D = \lambda D_d D_a + (1 - \lambda) D_c. \tag{8}$$

The default value of λ is set at 0.5 and decreases with the number of tracking failures. If the data association fails (due to occlusion, for example), the positional information of the tracker is not updated. Then the reliability of the distance and direction information deteriorates. Therefore, it would be reasonable to enhance D_c when the tracker does not work properly. Figure 10 illustrates the data association process based on evaluation value D. For each particle, D is calculated with respect to each

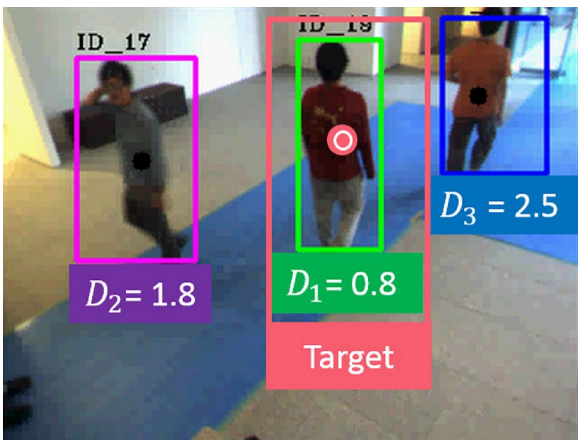

Fig. 10 Evaluation value for data association. The evaluation value, D, is shown below the detection window. The tracker is associated with the middle person in this figure

Fig. 9 Occlusion detection. The left image represents an original image. The right image represents occluded blocks, colored blue

detection window. The particle is associated with one of the windows, which shows minimal D. After the data association, the position and color information of each particle is updated.

Experiments

The validity of our system is verified through experiments using captured images in real environments. First, the human detection component is tested; second, the

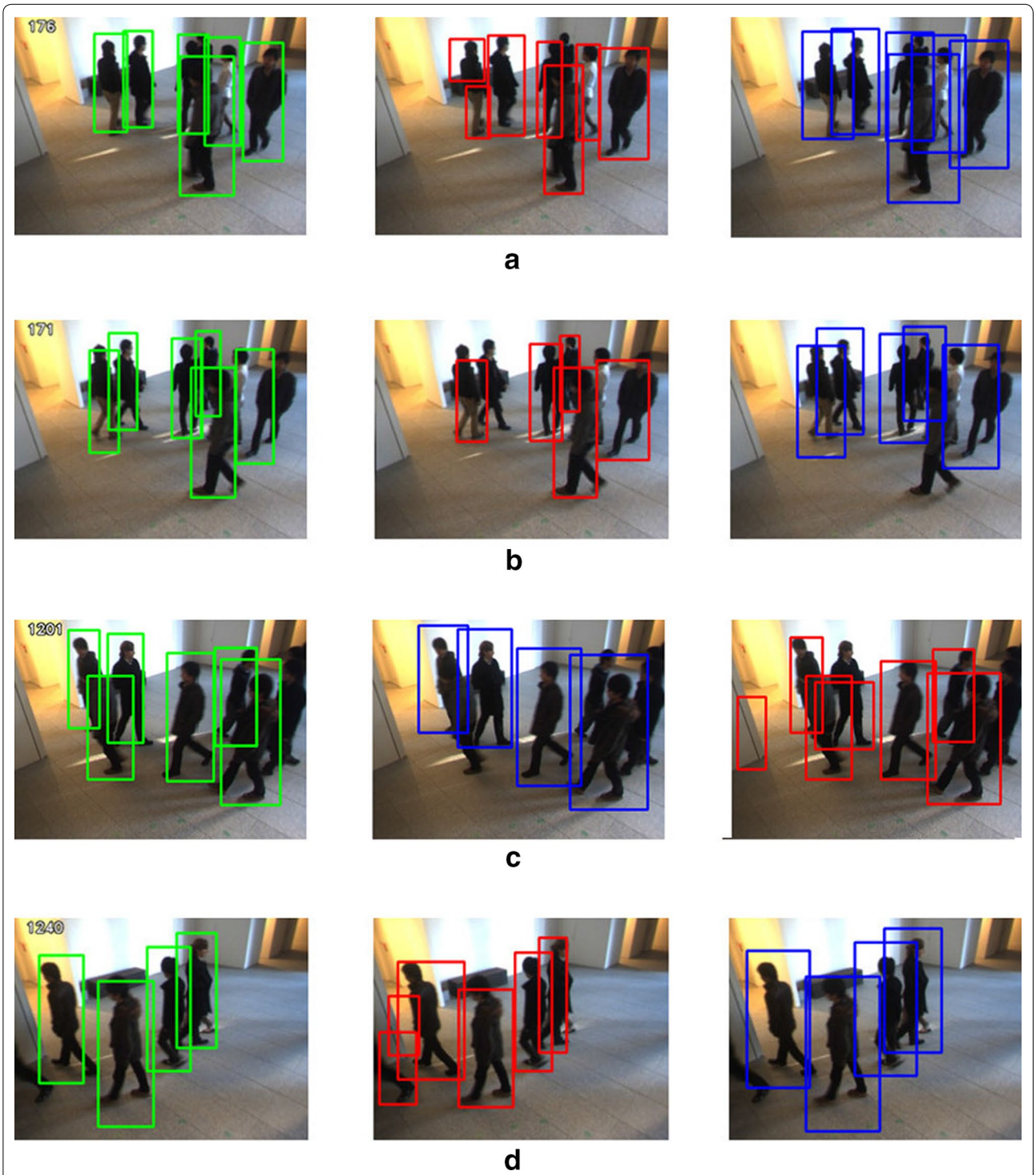

Fig. 11 Results of detection. **a** 171st frame, **b** 176th frame, **c** 1201st frame, **d** 1240th frame

Table 1 Results of detection performance

	Subtraction stereo	Joint HOG	Proposed method
TP	78.0	76.4	*83.2*
FN	22.0	23.6	*16.8*
FP	9.8	1.9	*0.9*
P	88.8	97.6	*98.9*
R	78.0	76.4	*83.2*
F	83.1	85.6	*91.0*

Italic values indicate better results than other detection methods

performance of the entire tracking system is discussed. The purpose of the experiment is to evaluate the validity of the proposed detection and tracking methods under typical environmental conditions in the indoor environment. Thus, the experimental conditions does not change significantly during the experiment.

Experimental settings

A Point Grey Research Bumblebee2 camera was used with a resolution of 320 × 240 pixels. The camera was mounted on a tripod stand. The installation angle and height of the camera were 30° and 2.3 m, respectively. The video was captured at 20 frames/s. Videos were taken on the 1st floor of Building 2 on the Chuo University Korakuen campus. In the video, captured scenes included partially or completely occluded persons as well as persons stopping suddenly, changing directions, and running.

We evaluated the performance of our method by precision, recall, and F-measure:

$$P = \frac{TP}{TP + FP}, \tag{9}$$

$$R = \frac{TP}{TP + FN}, \tag{10}$$

$$F = \frac{2PR}{P + R}, \tag{11}$$

where *TP*, *FP*, and *FN* are the number of true positive, false positive, and false negative, respectively. *P*, *R*, and *F* indicate precision, recall, and F-measure, respectively. Here, the tracking was assumed to have succeeded if the assigned ID did not change for more than three frames while the person was in the measurement range. The ID

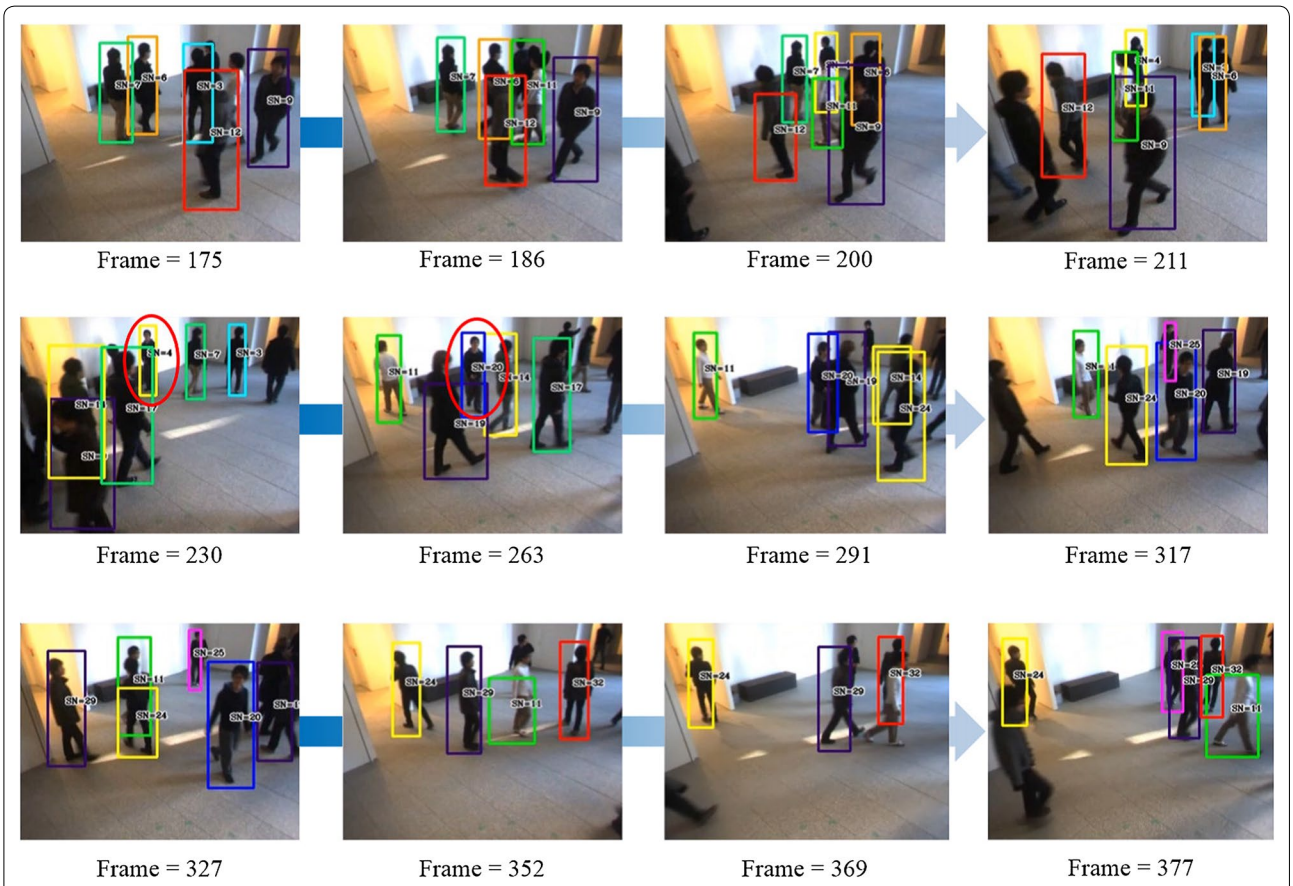

Fig. 12 Failure in integrating detection windows

Table 2 Results of tracking

Distance	Direction	Color	SN (number)	SR (%)
D_d	–	–	43	59.7
D_d	D_a	–	48	66.7
D_d	–	D_{cf}	56	77.8
D_d	D_a	D_{cf}	58	80.6
D_d	–	D_c	62	86.1
D_d	D_a	D_c	*66*	*91.7*

SN represents the number of successfully tracked persons. SR represents success rate of tracking

Italic values indicate better results than other combinations

was assigned when the tracking continued for ten frames. Interruption due to nondetection was not counted as a failure unless the ID changed after redetection.

Thresholds in "Integration" section are $thr_d = 40$ for the integration of detection windows; $thr_{SS} = 12.5$ and $thr_{JH} = 11$ for duplicate detections, and $thr_S = 220$ for score S.

The number of histogram bins for color information is 64, and a detection window is divided into 5×10 blocks. These parameters are determined based on preliminary experiments.

In the tracking phase, 10 frames captured after ID assignment are used for the model data in "Eligibility of each block" section. The number of particles per one person was 500. Parameter α in (8) was updated as

$$\alpha \leftarrow \alpha - 0.04m, \tag{12}$$

where m is the tracking failure count. m is initialized when the tracking was successful.

Detection

To confirm the improvement of the detection performance using the integrated detection method in "Detection", we compared the proposed method with the subtraction stereo and Joint HOG methods.

Representative output images are illustrated in Fig. 11. The results are shown in Table 1. The windows in the figure represent the detection results, where the colors of the windows represent the visibility.

Overall performance improved with the integration of subtraction stereo and Joint HOG methods. While decreasing nondetection (FN), misdetection (FP) also decreased as we intended. However, it is still difficult to reliably detect when the target's body is almost occluded. Figure 12 depicts examples in which the integration of detection windows failed. The persons were detected in quite close distance. Large part of the misdetection windows were then foreground and foreground disparity regions. Thus, the integration process did not work well. We are considering further integration with a body part detection method to improve robustness to occlusion.

Tracking

We tested the tracking performance of each combination of evaluation values. The results are shown in Table 2. Seventy-two people appeared in the video. Success in the table is represented by the number of people successfully tracked. Figure 13 depicts representative images from the tracking scene. Rectangles in the figure represent detection windows. The windows are colored in accordance with their assigned IDs for visibility. As can be seen, occlusion frequently occurred in the environment.

The tracking success rate improved remarkably with the introduction of the proposed data association

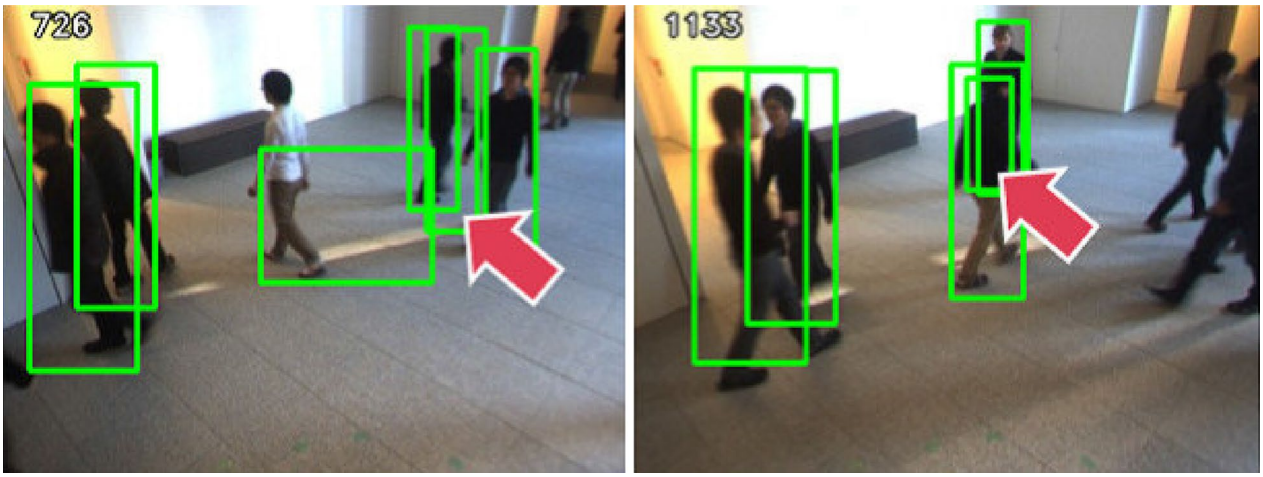

Fig. 13 Tracking scenes

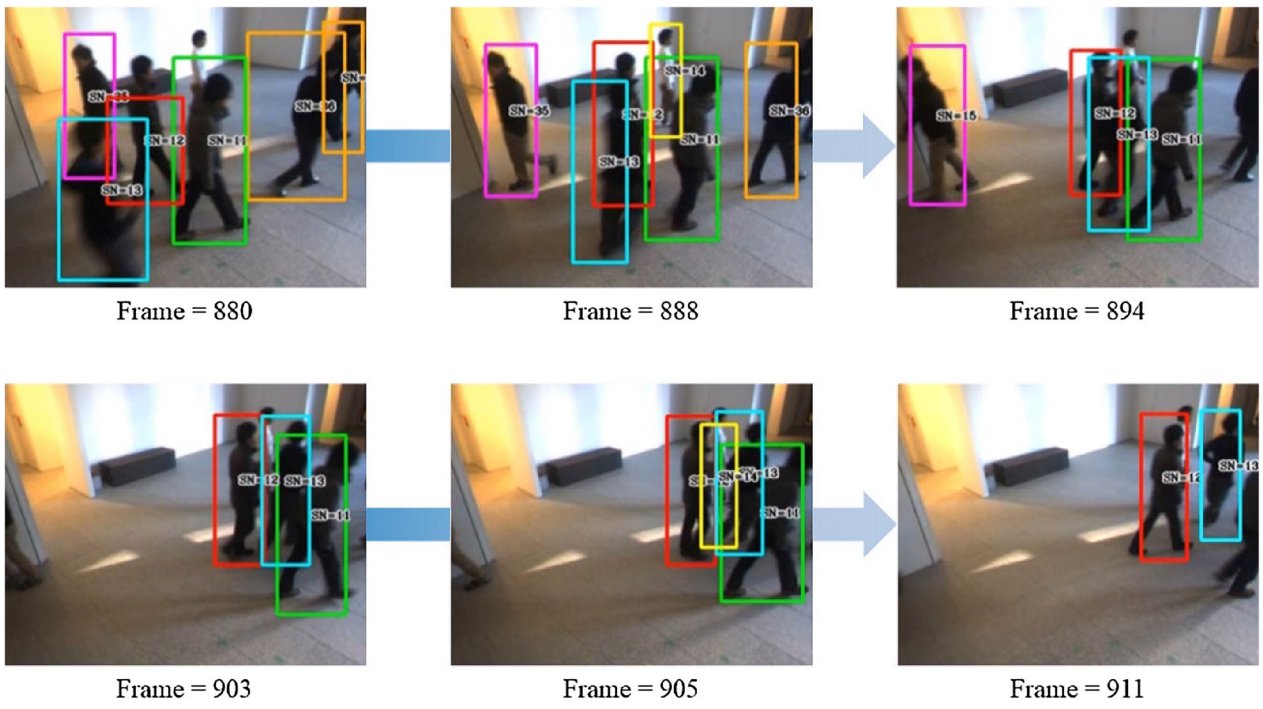

Fig. 14 Tracking scenes of a running person

method. When occlusion occurs, prediction errors of a tracker increase. In this situation, the unintended alternation of IDs or the assignment of a new ID was frequently observed in our experiments when using only distance and direction features. Block segmentation-based color features enhanced dissimilarity between a target and other people. Therefore, the proposed method could improve robustness to occlusion. However, tracking of a person circled at 230th and 263rd frames failed, as the ID changed. This is due to (a) consecutive nondetection in the frames and (b) rapid turning of the target after nondetection frames. The turn happened right after the first nondetection frame, and the prediction error increased remarkably. The tracker then could not find the person again. We believe that this problem would be fixed by the further integration of other detection methods. Figure 14 depicts scenes where one of the observed person (detected by an aqua rectangle; ID: 13) was running. Only the proposed method ($D_d + D_a + D_c$) succeeded to correctly track the running person. Exact data association contributed to the update process of position information.

Conclusion

A human tracking method using a single stereo camera is presented in this paper. The proposed method has advantages in that it is (1) robust to occlusion and (2) easy to install. By integrating two independently working detection methods, misdetection and nondetection frames can be reduced. A color feature for data association in particle filtering is introduced. The color feature is based on the eligibility and co-occurrence of segmented blocks of a detection window. The worsening of tracking performance due to occlusion is reduced by weighting the blocks. The validity of the proposed method was verified using measurement data at the entrance of a building. Although occluded scenes frequently appeared, the proposed method achieved tracking success at 91.7%.

Our final purpose is to construct an easily installable system for measuring human flow [12] for applications such as marketing, surveillance, and safety control. For this purpose, we hope to improve the accuracy of detection by integrating another fast and accurate detection method [13]. The automatic estimation of camera parameters is also a future goal. We did not discuss robustness against illumination change in this paper. It can be expected that our method directly gets the benefits of a stereo camera: robustness to illumination change. Other conditions, e.g. other places, the number of subjects, observation time, and time period are also not investigated in this paper. Thus, experiments in other conditions and environments are also desirable.

Authors' contributions
All authors equally contributed to develop the method. GM wrote the manuscript. TK conducted the experiments and analyzed the data. KU supervised the research. All authors read and approved the final manuscript.

Author details
[1] Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. [2] Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 27 February 2017 Accepted: 15 September 2017
Published online: 29 September 2017

References

1. Chen X, Bhamu B (2014) Soft biometrics integrated multi-target tracking. In: Proc. 22nd int. conf. on pattern recognition, pp 4146–4151
2. Satake J, Miura J (2012) Stereo-based tracking of multiple overlapping persons. In: Proc. 21st int. conf. on pattern recognition, pp 2581–2585
3. Luber M, Spinello L, Arras KO (2011) People tracking in RGB-D data with on-line boosted target models. In: Proc. 2011 IEEE/RSJ int. conf. on intelligent robots and systems, pp 3844–3849
4. Tseng T, Liu A, Hsiao P, Huang C, Fu L (2014) Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In: Proc. 2014 IEEE/RSJ int. conf. on intelligent robots and systems, pp 4077–4082
5. Fukushi K, Itsuo K (2010) Real-time human tracking using multiple-view stereo images and continuity constraints of long duration. In: Proc. 2010 IIEEJ image electronics and visual computing workshop, 1C-2-1-7
6. Ubukata T, Shibata M, Terabayashi K, Moro A, Kawashita T, Masuyama G, Umeda K (2014) Fast human detection combining range image segmentation and local feature based detection. In: Proc. 22nd int. conf. on pattern recognition, pp 4281–4286
7. Umeda K, Hashimoto Y, Nakanishi T, Irie K, Terabayashi K (2009) Subtraction stereo: a stereo camera system that focuses on moving regions. In: Proc. 2009 SPIE-IS & T electronic imaging, 3D imaging metrology, vol 7239. pp 1–11
8. Moro A, Terabayashi K, Umeda K, Mumolo E (2009) Auto-adaptive threshold and shadow detection approaches for pedestrians detection. In: Proc. 2009 Asian workshop on sensing and visualization of city–human interaction, pp 9–12
9. Mitsui T, Fujiyoshi H (2009) Object detection by joint features based on two-stage boosting. In: Proc. 12th ICCV workshop, pp 1169–1176
10. Gordon N, Salmond D, Ewing C (1995) Bayesian state estimation for tracking and guidance using the bootstrap filter. J Guid Control Dyn 18(6):1434–1443
11. Kawashita T, Shibata M, Masuyama G, Umeda K (2014) Tracking of multiple humans using subtraction stereo and particle filter. In: Proc. 2014 IEEE int. workshop on advanced robotics and its social impacts, pp 63–68
12. Kawashita K, Shibata M, Masuyama G, Umeda K (2015) An easily applicable system for measuring flow of pedestrians using a stereo camera. Trans JSME 81(2):149–155 (in Japanese)
13. Benenson R, Mathias M, Timofte R, Gool LV (2012) Pedestrian detection at 100 frames per second. In: Proc. 2012 IEEE conf. on computer vision and pattern recognition, pp 2903–2910