ROBOMECH Journal

RESEARCH ARTICLE Open Access

Occlusion handling for a target-tracking robot with a stereo camera

Yuzuka Isobe[1*], Gakuto Masuyama[2] and Kazunori Umeda[2]

Abstract

This paper presents an occlusion-handling method for a target-tracking robot with a stereo camera. One of the main challenges with the robot is to continue tracking when the illumination changes and occlusion occurs. In order to cope with the challenge, we use both color and disparity images acquired from a stereo camera. The tracking system is composed of three phases: candidate extraction, target identification, and occlusion handling. First, by using only three-dimensional (3D) information, target candidates are extracted. Second, the target is identified from the candidates based on a combination of both color and location features of the target and candidates. The combination depends on illumination changes that are supposed by changes in the white balance. Finally, the state of occlusion is estimated by results of both the analysis of the positional relationship between the candidates and the identification of a target. The proper procedure for the state is implemented. In the off-line experiments, the proposed method is compared with previous methods. Then, the proposed method is applied to a mobile robot, and an on-line experiment is carried out. Through the experiments, the effectiveness of the proposed method is verified.

Keywords: Occlusion handling, Target tracking, Mobile robot, Illumination changes, Stereo camera

Introduction

Autonomous mobile robots must have various abilities to assist humans. Tracking a specific person is one of these abilities. This skill can be applied to the carrying of luggage for a person, surveillance, and communication. With a wide range of applications, this ability is expected to be used both indoors and outdoors, from industry to daily life. Currently, target-tracking robots have been deployed in shopping centers [1], military areas [2], golf courses [3], and other places [4]. In order to utilize robots in wide fields and dynamic environments, it is essential for robots to have high perceptual capabilities. Color cameras are most commonly used to give mobile robots such capabilities. Color information acquired from the camera provides features of a target for effective target tracking. However, the effectiveness is influenced by illumination changes and occlusion.

Changing illumination is the leading cause of changes in the color information. One strategy for coping with the problem is to use both the color and location features of a target as humans do. Our previous method [5] was also based on the combination of both features. In that method, a parameter was adopted to show how illumination conditions change. Based on the parameters, how each feature is relied on changes. This compensates for weaknesses of using each feature singly. However, the method did not introduce any method of handling occlusion.

Occlusion occurs when a target is invisible in the frame and cannot be detected. It leads to the loss of a target because a similar candidate might be identified as the target. Additionally, due to occlusion, robots cannot determine whether tracking should be recovered due to loss or continued. One solution is based on estimating the state of occlusion occurring. Because it can be recognized that a target is invisible in the frame during occlusion, loss is prevented when a target is not supposed to be detected. Furthermore, estimation indicates which situation is occurring, loss or occlusion.

*Correspondence: isobe@sensor.mech.chuo-u.ac.jp
[1] School 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

Springer Open

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In this paper, we propose an occlusion-handling system for a target-tracking robot. The system exploits stereo vision, which can be produced stably during illumination changes. In a further development of our previous target-tracking method, the state of occlusion is estimated, and the procedure is selected in accordance with the state. The process of estimating the occlusion state is implemented with only three-dimensional (3D) information. Three occlusion states are defined: no occlusion, partial occlusion with the exact target detection, and partial/total occlusion with no one being detected. Based on the state, the color or location models of a target are updated, and locations of the other obstacles/people are predicted by trackers.

The paper is organized as follows. We review state-of-the-art target-tracking robots in "Efforts to overcome the occlusion problem" section . In "Proposed system" section , we detail the proposed system. "Experiments" section describes two types of target-tracking experiments, on-line and off-line. In the off-line experiments, the proposed system is compared with the previous one. On-line, in real-world outdoor environments, the proposed system is applied to a mobile robot, and its effectiveness is verified. Finally, "Conclusion" section concludes the paper and discuss future works.

Efforts to overcome the occlusion problem

Related works

Several techniques can be used in the attempt to carry out target tracking. Many of them introduce time-series filters to improve the accuracy of tracking. While someone overlaps a target, the estimation of the target's position reinforces robustness to occlusion. Some methods [6–8] use the filter without detecting the occlusion state. However, it is ambiguous as to whether to implement the procedure to recover or continue tracking.

The problem of occlusion is also a challenge in the field of human detection with a fixed camera. Researchers have proposed more occlusion-handling methods with a fixed camera than with a moving camera. Algorithmically, some of these fixed camera-based methods can be applied to the occlusion detection of target-tracking robots. Common methods using a fixed camera are based on classifiers that are built using machine learning algorithms. With benchmark datasets or preparing samples, whether occlusion occurs or not is classified as learning. The Support Vector Machine (SVM) is one of the most common algorithms. Wang et al. [9] use a linear SVM classifier for human detection to detect occluded regions. For human features, Histograms of Oriented Gradients (HOG) are combined with Local Binary Pattern (LBP). The idea is that densely extracted blocks of HOG features are prone to responding to the linear SVM score with negative inner products. HOG and LBP features have the advantage of being feasible for use during illumination changes. To develop the method, Shu et al. [10] introduced part-based detection of humans. The human model is created using features of the parts in each human region. It provides the advantages of excluding the effect from the background and obtaining the regional features. Although these methods perform with high accuracy, scanning the windows where the features are extracted causes the computational cost to be too high for applying to tracking robots. Basso et al. [11] and Cielniak et al. [12] reduced the computational cost by using another learning algorithm, Adaptive Boosting (AdaBoost). Additionally, these methods have been successfully embedded into robot systems. In both methods, the color feature is used to train the classifier. Because color is an unambiguous feature, the classifier performs better and is composed of a larger number of weak classifiers than classifiers that are trained using features other than color. However, these methods are affected more easily by illumination changes.

Without any learning algorithm, some approaches to detecting occlusion are presented according to their analysis of the appearance of a target or human. Pan et al. [13] proposed a content-adaptive progressive occlusion analysis. Occlusion detection is based on scanning the regions of interest (ROI). The occlusion situation is determined by analyzing the pixels in the ROI. Iterative scanning for target detection leads to not only high performance in the experiments but also high computational cost. By evaluating both the distance between objects and the changes of object size in an image, Yilmaz et al. [14] proposed a contour-based tracking to cope with the occlusion problem. Once the evaluation has detected an occlusion, modeling the contour changes alleviates the effect of shape variation from frame to frame. In [15], target-tracking and occlusion-handling methods were shown and applied to a mobile robot. The colors of a target's parts are used as features. The number of pixels in each region identifies the current situation from three cases of occlusion. Based on the case, the tracking procedure is implemented appropriately. In these methods, when the distance between objects is close and the target's size is changed, it causes the modeling to fail. Changes in size occur for two reasons, one is occlusion, the other is the changing distance of the target–getting close to a camera or farther away from it. Because these methods use only a color image, it is unclear which reason explains the modeling failure in the situation. Contrastively, disparity-based occlusion detection is carried out in [16]. This study uses the changes in both the distance between humans and a stereo camera, and the size of human regions. However, the human regions are given by the result of background subtraction method. By using the method, it is easy to extract the human regions. However, it cannot be applied to a moving camera and dynamic environments. Also, the method with the change in the feature between frames is proposed by Tran et al. [17]. The method uses the change in the number of people as a feature which indicates when occlusion occurs or finishes. However, in dynamic environments, the number would be changed by not only occlusion but also the movement of people.

Our previous method

In our previous method [18], the state of occlusion is estimated, and, then, the proper procedure is followed. The estimation is achieved using only two factors: the result of target identification and the analysis of the positional relationships between the target and others.

In the preprocessing of target identification, candidate targets are extracted. To help cope with illumination problems during candidate extraction, only 3D information is used. The 3D information is steadily acquired from a stereo camera, even under varying illuminations. Based on the 3D information, a point cloud in a 3D space is produced. By using the point cloud, candidate regions can be extracted, even when the entire region of the target (from head to foot) cannot be visible due to partial occlusion. Furthermore, it reduces estimation errors regarding a target's position during occlusion. Also in the method of [5], a target is identified from the candidates using a combination of both color and location models based on illumination parameters.

The positional relationship is analyzed based on the 3D information, and the occlusion state is estimated from the results of the analysis. The occlusion states are classified into three types: no occlusion occurs (STATE 1), so little occlusion occurs that a target is identified (STATE 2), and so much occlusion occurs that no one is identified as a target (STATE 3). In accordance with the state, it is determined whether each color or location is registered as a target feature.

The results of the experiments showed the method's effectiveness. However, there may be challenging situations with the method [18], as shown in Fig. 1. In each figure, the region of identified target is depicted as a red rectangle. A target is drastically occluded by person A when the illumination changes extremely. In the method, the reliability of both color and location features is determined by the degree of changes in the illumination. In this situation, target tracking relies heavily on the location feature. Then, during total occlusion of approximately 21 frames (3.0 s), estimation errors of the target's position accumulate. Finally, the estimation is close to the position of A, and mis-identification occurs.

In order to cope with the problem, an occlusion-handling method is developed. In the method, candidates other than the target are also tracked from frame to frame. To prevent the escalation of computational costs, tracking is implemented only during STATE 2 and 3. In the next section, we will detail the proposed method.

Proposed system

The proposed system is composed of a stereo camera mounted on a mobile robot. The target-tracking system consists of three procedures: candidate extraction, target identification, and occlusion detection.

The first phase of the system is candidate extraction. By projecting 3D information acquired from a disparity image into a 3D space, the candidate regions of a target are extracted. Second, a target region is distinguished from the others. In order to achieve the distinction, a target model is produced. The model is composed of a target's color and location features. Finally, in the process of occlusion detection, the positional relationship between a target and the others is analyzed. Using the results of both the analysis and target identification, the occlusion state is estimated. Depending on the state, the appropriate procedure is followed.

Candidate extraction

In our previous system [5], the segmentation method [19] was applied for candidate extraction. The method utilized an overlooked plane, and 3D information in each pixel of a disparity image was projected onto the plane. The density of the projected points tended to be high in regions corresponding to humans. The human regions were extracted based on density. The method required the entire region of the person (from head to foot) to be visible because the height information was squeezed on the plane. When partial occlusion occurs and a target region is partly visible but not entirely, the target may not be detected even as human. As mentioned above, according to how long occlusion continues, estimation errors regarding the target's position accumulated. To avoid this accumulation, extracting a partially occluded region of the candidate is also desirable in this phase.

Therefore, in this paper, the candidate-extraction procedure utilizes a 3D space against an overlooked 2D plane. By using 3D information acquired from a disparity image, a point cloud is obtained in the space. The space is defined by the X–Y–Z coordinate, as shown in Fig. 2. The procedures for candidate extraction are explained as follows: first, a point cloud is produced. For instance, Fig. 3 depicts an illustrative captured image. The point cloud acquired from the corresponding disparity image is shown in Fig. 4. Note that the points on the ground have been eliminated. The cloud includes groups of points that

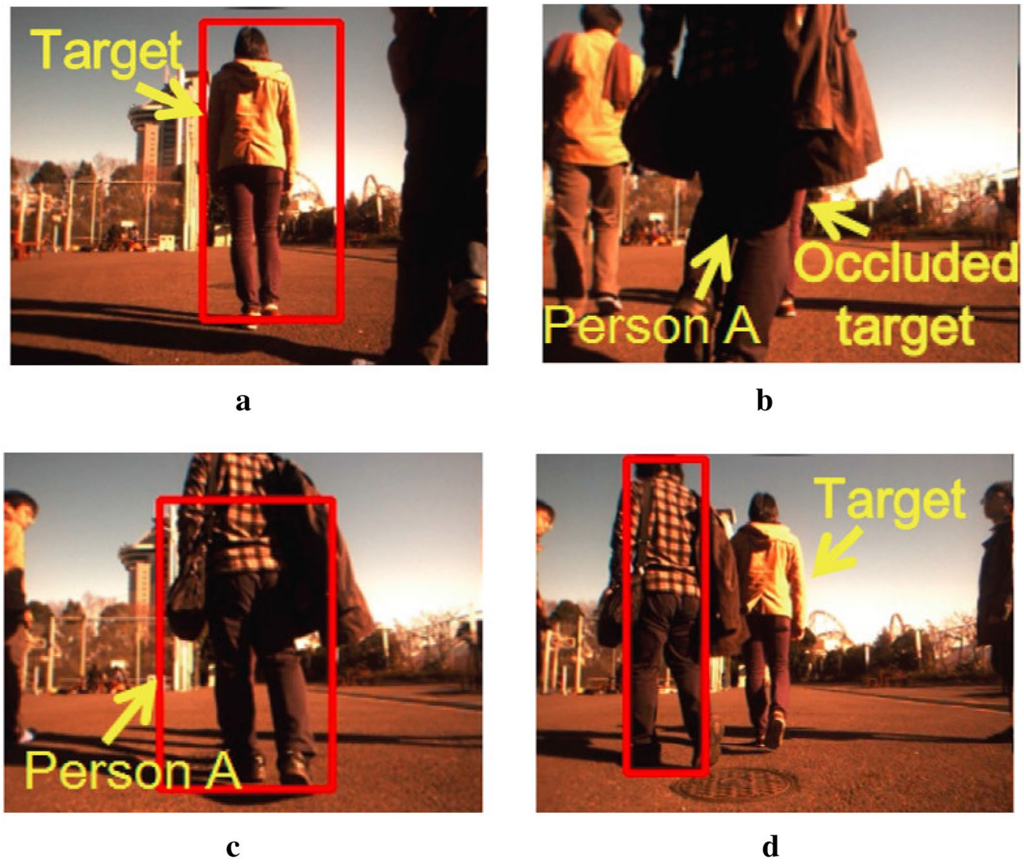


Fig. 1 Mis-identification of a target due to long-term occlusion. Each red rectangle indicates the result of target identification. During occlusion, estimation errors of the target's position accumulate. The errors cause mis-identification. **a** # 1774, **b** # 1793, **c** # 1802, **d** # 1810

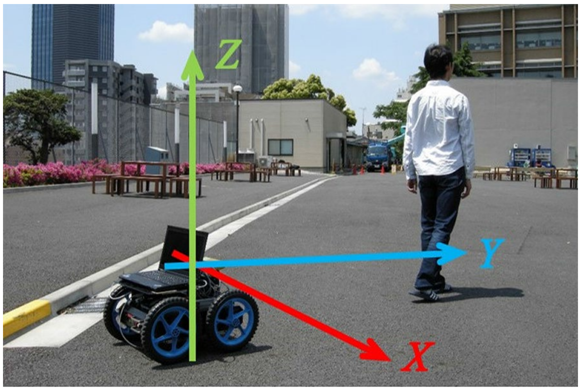


Fig. 2 X–Y–Z coordinate of the proposed system

correspond to three people in Fig. 3. Second, in order to obtain the density of the points, the space is divided into boxes, and the number of the points in each box is counted. The density in each box would represent the existence of the candidates. Therefore, if the density in a box exceeds a certain threshold, the box is assumed to be a part of the candidate. Third, boxes with high densities are extracted (Fig. 5) and labeled. In the labeling procedure, four-connected components are defined as belonging to the same label. Fourth, mean-shift clustering is implemented to merge or split labeled regions. Finally, candidates' regions are extracted by thresholding with respect to the height, width, and depth of the regions. The result of candidate extraction for the illustrative condition is shown in Fig. 6. Three regions are extracted as the candidates of a target. The candidate-extraction method gives robustness to both partial occlusion and illumination changes. In contrast to methods that use the contour features of people, our method does not require the entire contour to be visible. Furthermore, because the method is composed only of 3D information, it is not affected by varying illumination.

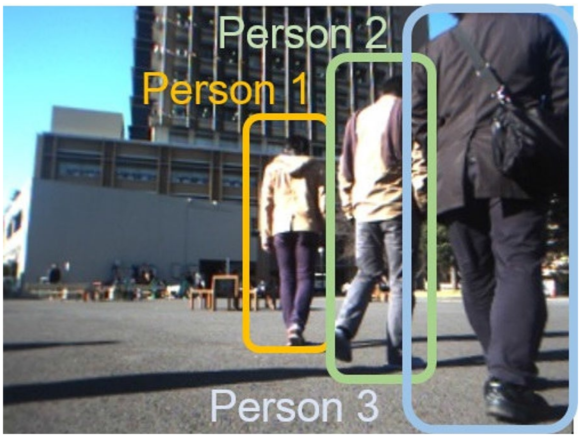


Fig. 3 Example of a color image

Target identification

To compare extracted candidates with a target, the target's model is used. Both color and location features are components of the model.

The dissimilarity of color features is based on comparisons of the hue and saturation histogram. The color histograms of each candidate H_c and the color model of a target H_t are compared by the following equation:

$$R_{color} = \sqrt{1 - \sum_h \sum_s \sqrt{H_c(h,s) H_t(h,s)}.} \tag{1}$$

Note that $H_c(h,s)$ and $H_t(h,s)$ indicate the normalized frequencies of hue (h) and saturation (s). Additionally, the color model is compared with pre-registered one, and if the dissimilarity is under threshold, the color model is updated.

The location model of a target is given by a Kalman filter. The filter is defined based on the assumption between frame k and $(k+1)$:

$$\boldsymbol{X_{k+1}} = F_k \boldsymbol{X_k} + \boldsymbol{u_k} + \boldsymbol{v_k} \tag{2}$$

$$\boldsymbol{z_k} = H_k \boldsymbol{X_k} + \boldsymbol{w_k} \tag{3}$$

where $\boldsymbol{X_{k+1}}$ and $\boldsymbol{X_k}$ are the state at respective frame, F_k and H_k are the transition and observation model, $\boldsymbol{u_k}$ is the control vector, $\boldsymbol{z_k}$ is the observation at frame k ,and

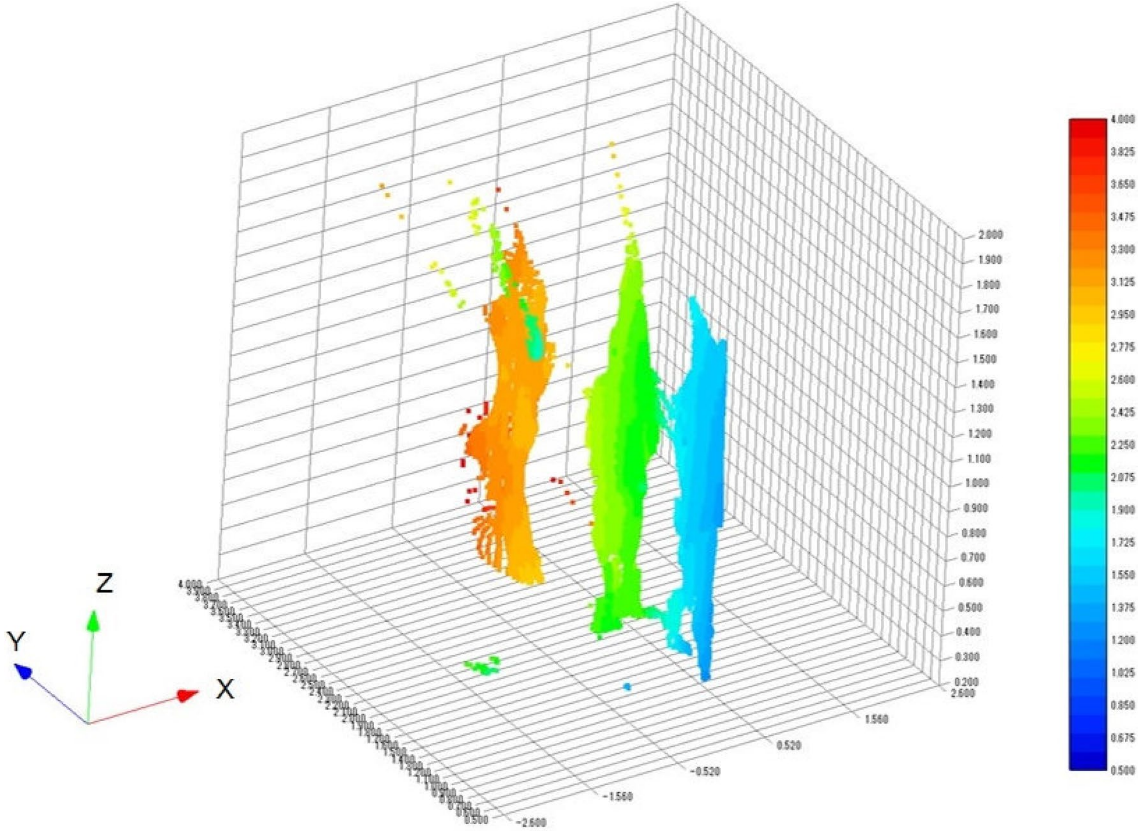


Fig. 4 Point cloud

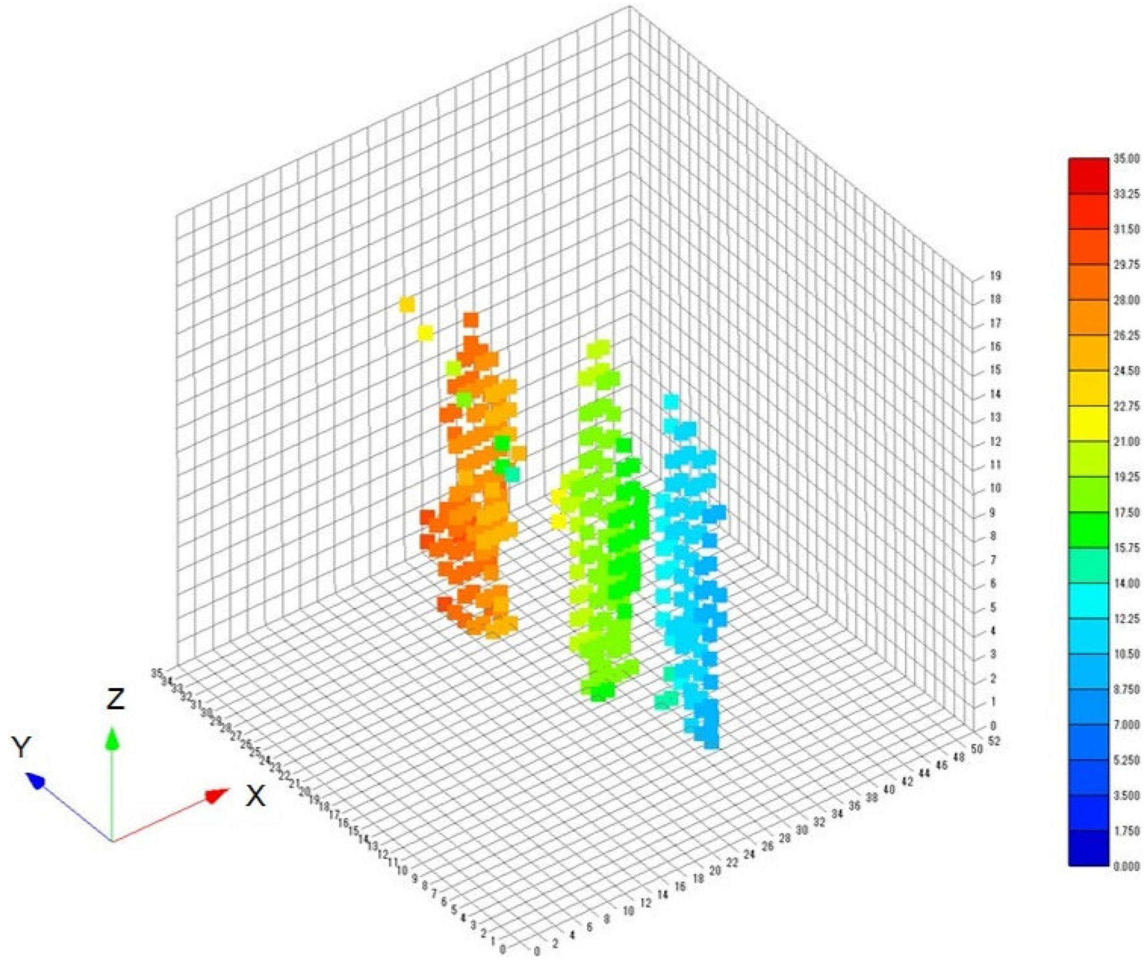


Fig. 5 Divided boxes

$\boldsymbol{v_k}, \boldsymbol{w_k}$ are the noises. The state and observation of a target at frame k, respective $\boldsymbol{X_k}$ and $\boldsymbol{z_k}$ is defined

$$\boldsymbol{X_k} = \begin{pmatrix} X_k \\ \dot{X}_k \\ Y_k \\ \dot{Y}_k \end{pmatrix}, \quad \boldsymbol{z_k} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} \tag{4}$$

where X_k and x_k are the location of X, and Y_k and y_k are Y direction. The transition and observation model F_k and H_k is calculated by the equations:

$$F_k = \begin{pmatrix} 1 & \Delta k & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta k \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$H_k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

The location feature between each candidate and the model is compared

$$R_{location} = k\sqrt{(X_c - X_t)^2 + (Y_c - Y_t)^2}, \tag{6}$$

where (X_c, Y_c) is the location of a candidate, (X_t, Y_t) is the location model of a target, and k is a fixed value for normalizing $R_{location}$.

As with our previous method [5], the total dissimilarity between candidates' features and the model is calculated by combining these dissimilarities in accordance with the illumination changes. The total dissimilarity is defined as follows:

$$D = \begin{cases} (1-\alpha)R_{color} + \alpha R_{location} & (\alpha < \alpha_{th}) \\ R_{location} & (otherwise) \end{cases}, \tag{7}$$

where α is the parameter that represents illumination changes and has the relation $\alpha = p|W|$, W is the amount of the white-balance change, and p is a constant. The amount of white balance W means the difference of the

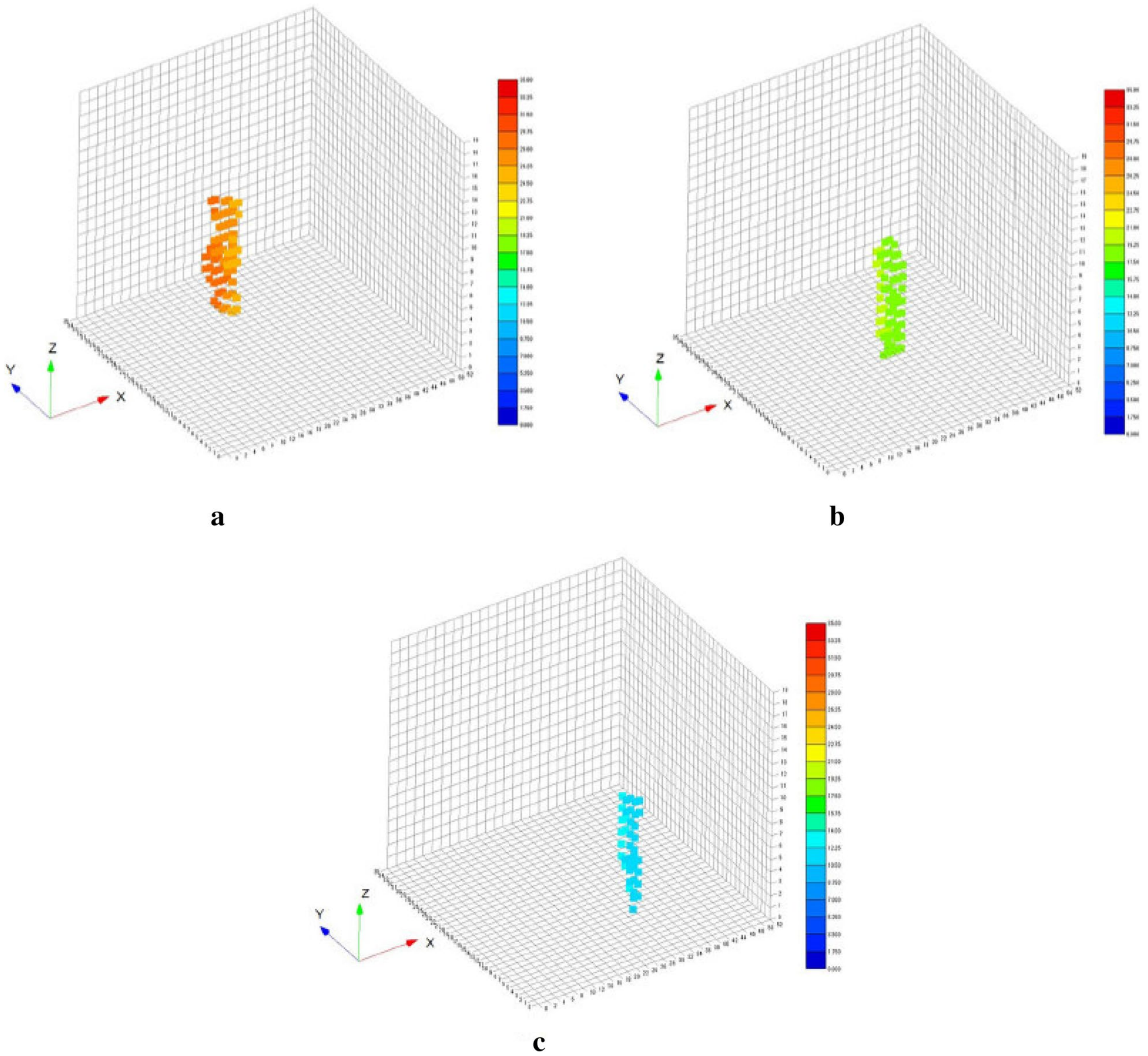


Fig. 6 Result of labeling and mean-shift clustering of point cloud. **a** Person 1, **b** Person 2, **c** Person 3

value between the present frame and the last frame in which the color model is updated. The value of p is determined so as to hold the relation of $0 \leq \alpha \leq 1$. In (7), α_{th} denotes the threshold of illumination change. When illumination changes remarkably ($\alpha \geq \alpha_{th}$), we use only the location feature.

Occlusion handling

Previously, we proposed an occlusion-detection method [18] that used a disparity image. First, by using the method, the state of occlusion in the latest frame is determined. The occlusion-handling procedure is performed in accordance with the state.

Occlusion detection

The state of the positional relationship between a target and other objects/people is analyzed on an overlooked plane (Fig. 7). In the figure, the blue rectangle shows a stereo camera, and the dotted lines indicate the cameras' fields of view. D is the closest distance that can be measured by the stereo camera. Region A is defined as the inner region, which is structured by lines that connect the left camera with the left edge and the right camera

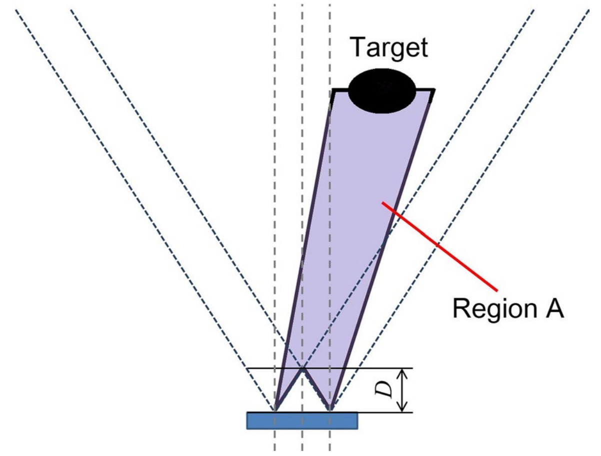


Fig. 7 Definition of Region A. The blue rectangle shows a stereo camera, and the dotted lines indicate the cameras' fields of view. *D* is the closest distance that can be measured by the stereo camera. Region A is defined as the inner region, which is structured by lines that connect the left camera with the left edge and the right camera with the right edge of a target's region with margins

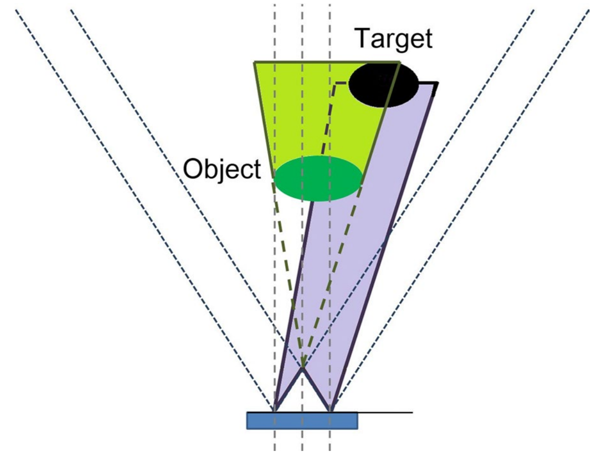


Fig. 8 Definition of Region B. The green region indicates where the object hides in the image

with the right edge of a target's region with margins. When there are objects/people in Region A, the partial/total region of a target is (or is going to be) hidden in an image, i.e. occlusion occurs. While occlusion continues, Region B is structured as the region hidden by the object in Region A (Fig. 8).

By using Regions A and B, and also based on the result of target identification, the state of occlusion is classified into three types: STATE 1, 2, and 3.

I. STATE 1: When no object/person violates Region A, the occlusion state of the frame is STATE 1. No occlusion occurs in this state.

II. STATE 2: When objects/people are present in Region A, occlusion is regarded to occur. If an identified target is partially occluded, the occlusion state is STATE 2. In this situation, the edges of the target's region are incorrectly determined because the correct edges might be hidden. Due to the margins on the edges, it is allowed to define this state even when the target is partially in Region B. The width of the margins depends on uncertainly of candidate extraction because the candidate region is given by boxes. Therefore, as shown in Fig. 9, when the target's region is not hidden but is close to Region B, it is also classified as this state. In other words, STATE 2 shows the situation when a target is going to be occluded in a few frames.

III. STATE 3: When the estimated region of a target is in Region B and a target is not identified, the occlusion state is determined to be STATE 3. The region of a target is defined as the model of the target's position acquired by a Kalman filter with margins. The width of the margins is determined by the width of the target region that is identified with STATE 1 just prior to STATE 3. In this state, only the estimated location of the target can be assessed, due to occlusion.

The occlusion-handling procedure in each state

As defined above, situations with/without occlusions are shown for all STATE. Then, the strategy for continued tracking in each STATE is detailed as follows.

I. STATE 1: In this state, the color model of a target is considered to be correctly obtained, because no object/person occludes a target. Therefore, the color model is updated to adjust to illumination changes. In addition, the location model of a target is also updated to reduce the estimation errors of the target's position.

II. STATE 2: Though a target is visible, the region may not be completely visible. The color model of a target is not updated. The location model is updated when the width of the target's region exceeds a threshold. Another problem during occlusion is the incorrect identification of a target. When the target region is occluded, the candidate would be erroneously identified and mistaken for the correct target. With tracking objects/people between frames, the problem could be avoided. To reduce the computational cost, objects/people (whether they are extracted as the candidates of the target or not) are

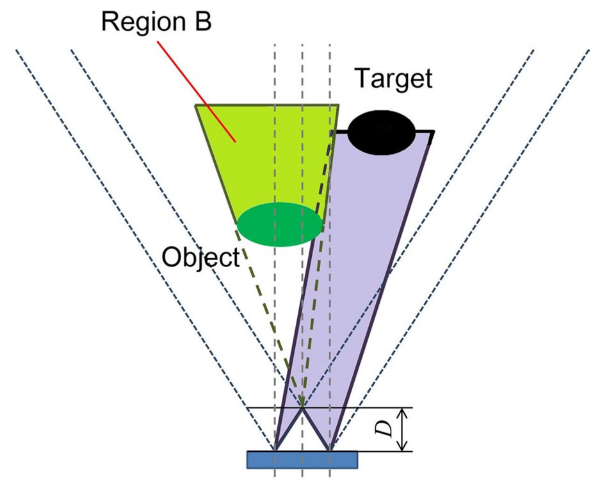


Fig. 9 The situation when the object does not violate Region A. This state is also determined to be STATE 2 because of the margins attached to the target's region

tracked using a Kalman filter only during STATE 2 and 3. The color features of objects/people are not used for tracking. Tracking is implemented based on Euclidean distance, as shown in Eq. 6. In STATE 2, the objects in Region A are tracked between frames. Once the objects invade Region A, the location features of the objects are registered and are tracked during STATE 2 or 3. Even when the registered feature is not similar to any object, the estimated position of the feature is updated using the measurement position. If the object position is not similar to that of any registered object, the position is newly registered.

III. STATE 3: Occlusion causes a target to be hidden and not identified. Therefore, the duration of the estimation of a target's position is extended until the estimated position is moving out of Region B. Non-target objects/people have also been tracked in this state. All of the objects/people within 0.5 m of a target's estimated position are tracked. This procedure allows for re-identification after long-term occlusion.

Additionally, if some objects are tracked and the value of the illumination parameter α exceeds a certain threshold, target identification might be incorrect. When illumination changes and the neighbor candidate partly/totally occludes a target, identification failure might occur. The $R_{location}$ of the identified target is compared with the distance between the position of the identified target and each registered position. Through the comparison, if the $R_{location}$ is the smallest value, the target is regarded to have been identified correctly. On the contrary, if another distance value is the smallest, the identified target is not considered to be a correct target but a corresponding candidate. It follows that a target cannot be identified and it leads to STATE 3.

Experiments

The proposed system has been tested in outdoor environments with both illumination changes and occlusion. We have conducted two types of experiments, off-line and on-line. In the off-line experiments, images that had been prospectively captured were used. The proposed method and our previous methods were applied to the images and compared. In the on-line experiment, a mobile robot tracked a target. In each experiment, we used the Bumblebee2, of Point Grey Research, as a stereo camera, and Blackship, of Segway Japan, as a mobile robot. Additionally, the parameters of the proposed method is shown in Table 1. The amount of white balance changes $|W|$ is calculated as the sum of the changes of red and blue gains of the camera. Each gain changes in 1024 steps. In addition, the minimum width when the location model of a target is updated in STATE 2, is half value of the target's width when a target has been identified last.

Off-line experiments

Before the experiments, images had been captured by the stereo camera attached to the mobile robot. The robot was controlled to follow a target by a human operator. During controlling, 3260 frames were captured at 12.8 Hz.

Details of the experimental environments are shown in Tables 2 and 3. In Table 2, the number of times when occlusion occurred and the average and maximum duration of occlusion are shown. In Table 3, the number of frames and the average duration are shown for each number of people. Additionally, Fig. 10 shows examples of the color images that were used in the experiments.

In the off-line experiments, three types of methods were applied. Method I is the proposed method,

Table 1 The parameters of the proposed method in the experiments

The size of a box	One side 0.1 m
The closest distance which can be measured: D	0.2 m
The fixed value k in Eq. (6)	1.8 m^{-1}
The constant value p of the illumination parameter α	0.1
The threshold α in Eq. (7)	0.5
The width of margins with each edge of Region A	0.2 m
The threshold of target identification which relates the total dissimilarity	0.4

Table 2 The details of occlusion condition in off-line experiments

Number of times	Average duration	Maximum duration
437 frames (29 times)	15 frames	47 frames

Table 3 The details of the number of people, except the target, in off-line experiments

Number of people	Number of frames	Average duration
0	818 frames (15 times)	55 frames
1	1719 frames (30 times)	57 frames
2	584 frames (20 times)	29 frames
3	139 frames (7 times)	20 frames

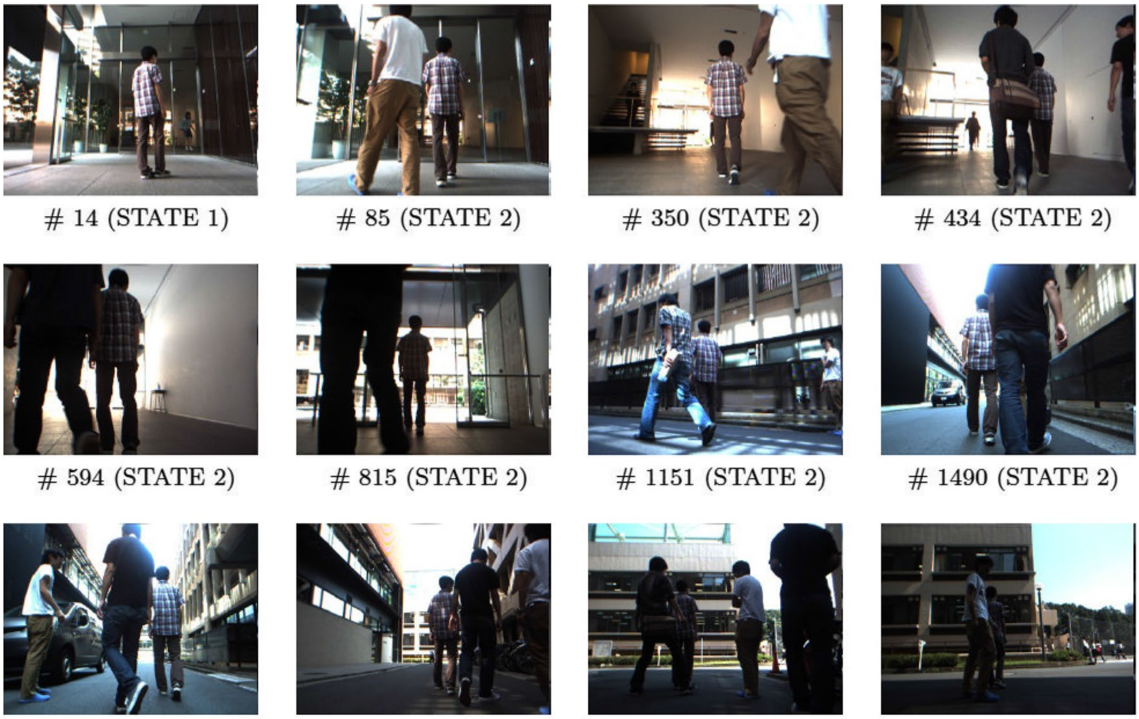
with procedures for detecting and handling occlusions. Method II is our previous method [5], with procedures for detecting occlusions and both updating the color model and continuing the duration of estimation of target's position. Method III is also our previous method [18] without any procedures for detecting or handling occlusions.

The effectiveness of each method is verified by three evaluation values: precision, recall, and F-measure P, R, and F, respectively. These values represent accuracy, completeness, and the harmonic mean of precision and recall, respectively.

$$P = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}} \quad R = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{C}} \quad F = \frac{2PR}{P+R} \tag{8}$$

A: the number of frames in which the target is correctly detected, B: the number of frames in which a non-target is detected (mis-identification), C: the number of frames in which no objects are detected (dis-identification).

Table 4 is the result of the calculation of each evaluation value against each method. The result shows that the highest evaluation values of all methods are acquired using the proposed method. Comparing Method II with III, the precision value of Method II is lower than that of Method III. However, the recall and F-measure values of Method II are higher than those of Method III. This shows that, under occlusion, a target is readily lost without occlusion detection. The detection method helps complete identification but might cause mis-identification. Therefore, the proposed method, which aims to decrease mis-identification, is effective.



14 (STATE 1) # 85 (STATE 2) # 350 (STATE 2) # 434 (STATE 2)

594 (STATE 2) # 815 (STATE 2) # 1151 (STATE 2) # 1490 (STATE 2)

1819 (STATE 2) # 2432 (STATE 2) # 2966 (STATE 3) # 3107 (STATE 2)

Fig. 10 Captured color images used in off-line experiments. The frame number and occlusion state which is classified by using the proposed method, are shown under each image

Table 4 The result of off-line experiments

Method	*P* (%)	*R* (%)	*F* (%)
I (proposed)	99.8	91.7	95.6
II [5]	93.7	87.2	90.4
III [18]	95.1	79.2	86.4

Table 5 The number of frames when a target is lost

Method	I	II	III
Number of lost frames	39	221	757

Table 6 The details of occlusion conditions in the on-line experiment

Number of times	Average duration	Maximum duration
418 frames (25 times)	17 frames	41 frames

Situations in which a target is not identified are classified into two types. One is occlusion, the other is the loss of a target. With Method I or II, when a target is not identified and there is no one in Region A, the situation is defined as the frame when a target was lost. Using Method III, whenever target identification is not carried out, the situation is defined as the loss of a target. The number of frames when a target is lost with each method is shown in Table 5.

The number of Method I is the smallest of all methods, approximately 1% of all frames (3260 frames). The number of Method II is less than that of Method III. This also shows the necessity of occlusion estimation.

On-line experiment

The proposed system was applied to the mobile robot with the stereo camera in real-world environments. The robot's behavior was based on a PID controller so as to keep the distance between a target and the robot to 1 m and the angle of direction to 0 rad. An entire experiment was composed of 1498 frames that were captured at 12.1 Hz. The experimental environments are classified into five scenes based on the illumination. Details of the experimental environments are explained in Tables 6 and 7. Three of the values are also used for evaluation (see "Off-line experiments" section).

Figure 11 depicts the target-identification results. In the figure, red rectangles indicate the centroids of the target. The results of the evaluation are shown in Table 8.

Table 7 The details of the number of people, except the target, in the on-line experiment

Number of people	Number of frames	Average duration
0	260 frames (5 times)	52 frames
1	329 frames (18 times)	19 frames
2	396 frames (28 times)	14 frames
3	256 frames (19 times)	13 frames
4	66 frames (6 times)	11 frames
5	7 frames (1 times)	7 frames

The *precision* values are calculated as 100%. Both of the *recall* and *F-measure* values are higher than 94%. The results follow the purpose of this method to decrease target mis-identification. Even when there was a person near the target who caused occlusion, mis-identification did not occur. Additionally, the robustness to occlusion is shown by successful identification after 20 frames (about 1.9 s) of occlusion.

However, frames occasionally occurred in which no target was identified despite the target's presence, 92% (55 frames) of which were caused by occlusion. Due to estimation errors regarding the target's position, in 23 frames, dis-identification occurred after occlusion. In each frame, the target was regarded not as the target but as the other candidate, and associated across frames. It is impossible to estimate the target position correctly during long-term occlusion. Therefore, a recovery method is required that will help with re-identification whenever the target is lost.

Dis-identification by occlusion occurred in 21 frames due to illumination changes. Before or after the frames, the target was occluded, and illumination changed. However, the illumination parameter did not change. Figure 12 shows this situation. In the situation, only the brightness of the images changed; therefore, the white-balance values did not change. To deal with the problem, other factors that reflect the brightness changes should be adopted in the future.

Identification was prevented in the other 9 frames. In these frames, part of target's region was still visible. Because the color model of a target was produced using the entire region of the target, the color histogram of the small part did not correspond to the model.

Segmentation errors of the candidate's regions caused the errors in another 3 frames, as the target's region was not correctly extracted and identified as a target.

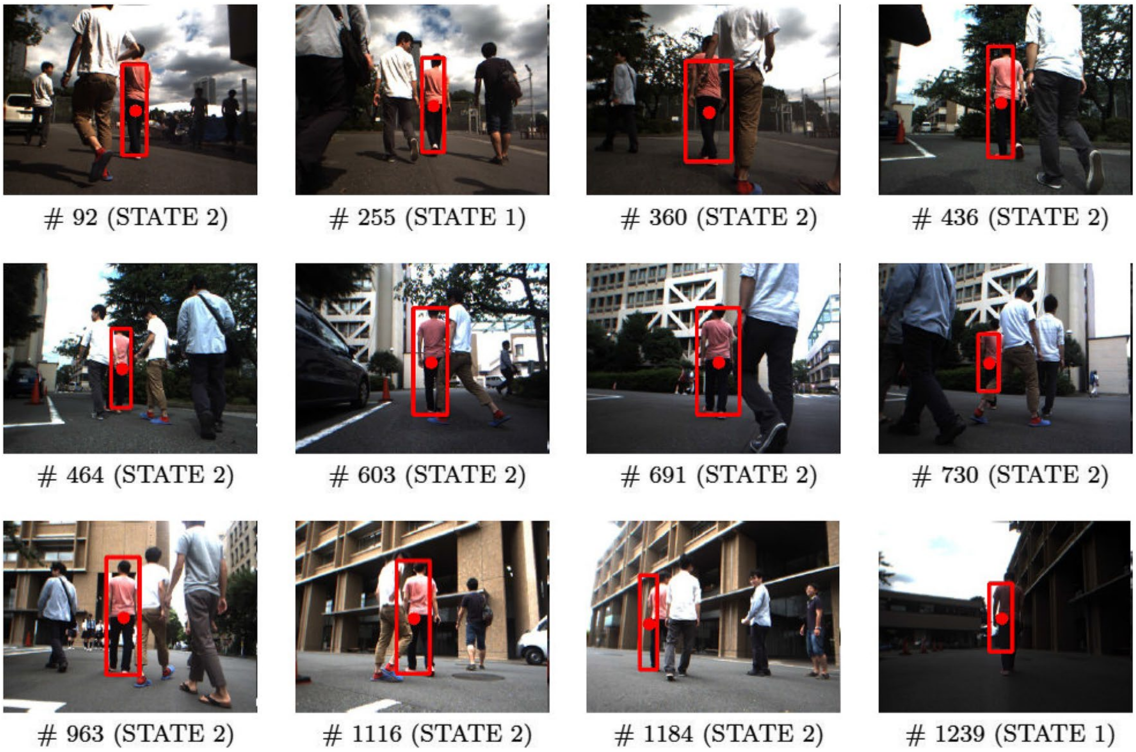

92 (STATE 2)
255 (STATE 1)
360 (STATE 2)
436 (STATE 2)
464 (STATE 2)
603 (STATE 2)
691 (STATE 2)
730 (STATE 2)
963 (STATE 2)
1116 (STATE 2)
1184 (STATE 2)
1239 (STATE 1)

Fig. 11 Result of target identification in the on-line experiments. The frame number and occlusion state which is classified by using the proposed method, are shown under each image

Table 8 Result of the on-line experiment

Precision *P* (%)	Recall *R* (%)	F-measure *F* (%)
100	94.5	97.2

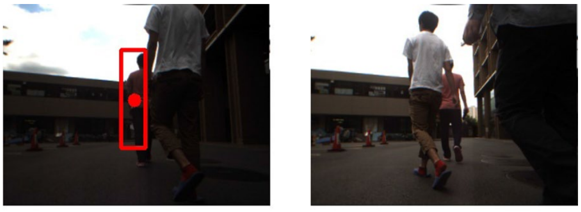

Before occlusion
After occlusion

Fig. 12 Brightness changes during occlusion

Conclusion

In this paper, an occlusion-handling method for target-tracking robots with a stereo camera has been proposed. We have focused on its weakness with occlusion and illumination changes. First, the occlusion state was classified into three types. Then, in accordance with the state, the proper procedure was implemented. Through off-line experiments, the effectiveness of the proposed method was verified by comparison with other methods. Also, the on-line experiment was carried out to demonstrate its robustness to occlusion and illumination changes.

In the future, a recovery method after the loss of a target should be adopted. Also, the parameters that indicate brightness changes in a color image must be used to enhance the robustness for illumination changes.

Authors' contributions

YI devised the basic concept of the overall system, technically developed the algorithm, carried out the experiments, analysed the data, and drafted the manuscript. GM and KU participated in the research design. All authors read and approved the final manuscript.

Author details

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

Acknowledgements

Not applicable.

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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 29 August 2017 Accepted: 18 February 2018
Published online: 01 March 2018

References

1. Budgee. Five elements robotics. http://www.5elementsrobotics.com/. Accessed 21 Feb 2016
2. Raibert M, Blankespoor K, Nelson G, Playter R, the BigDog Team (2008) BigDog, the rough-terrain quadruped robot. In: Proceedings of the 17th world congress of the international federation of automatic control, pp 10822–10825
3. STEWART GOLF X9. Stewart Golf Limited. http://www.stewartgolf.com/X9Follow. Accessed 21 Feb 2016
4. KSI. Doog Inc. http://www.doog-inc.com. Accessed 21 Feb 2016
5. Isobe Y, Masuyama G, Umeda K (2015) Target tracking for a mobile robot with a stereo camera considering illumination changes. In: Proceedings of 2015 IEEE/SICE international symposium on system integration, pp 702–707
6. Petrovic E, Leu A, Ristic-Durrant D, Nikolic V (2013) Stereo vision-based human tracking for robotic follower. Int J Adv Robot Syst 10:1–10
7. Takemura H, Nemoto Z, Mizoguchi H (2009) Development of vision based person following module for mobile robots in/out door environment. In: Proceedings of the 2009 IEEE international conference on robotics and biomimetics, pp 1675–1680
8. Ess A, Leibe B, Schindler K, Van Gool L (2008) A mobile vision system for robust multi-person tracking. In: Proceedings of international conference on computer vision and pattern recognition (CVPR), pp 1–8
9. Wang X, Han TX, Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: Proceedings of international conference on computer vision (ICCV), pp 32–39
10. Shu G, Dehghan A, Oreifej O, Hand E, Shah M (2012) Part-based multiple-person tracking with partial occlusion handling. In: Proceedings of international conference on computer vision and pattern recognition (CVPR), pp 1815–1821
11. Basso F, Munaro M, Michieletto S, Pagello E, Menegatti E (2013) Fast and robust multi-people tracking from RGB-D data for a mobile robot. In: Intelligent autonomous systems 12, https://doi.org/10.1007/978-3-642-33926-4
12. Cielniak G, Duckett T, Lilienthal A J (2007) Improved data association and occlusion handling for vision-based people tracking by mobile robots. In: Proceedings of the 2007 IEEE/RSJ international conference on intelligent robotics and systems, pp 3436–3441
13. Pan J, Hu B, Zhang JQ (2008) Robust and accurate object tracking under various types of occlusions. IEEE Trans Circuits Syst Video Technol 18(2):223–236
14. Yilmaz A, Li X, Shah M (2004) Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Trans Pattern Anal Mach Intell 26(11):1531–1536
15. Bai P, Qiao H, Wan A, Liu Y (2006) Person-tracking with occlusion using appearance filters. In: Proceedings of the 2006 IEEE/RSJ international conference on intelligent robots and systems, pp 1805–1810
16. Ma Y, Chen Q (2010) Depth assisted occlusion handling in video object tracking. In: International symposium on advances in visual computing, pp 449-460
17. Tran TA, Harada K (2013) Depth-aided tracking multiple objects under occlusion. J Signal Inf Process 4(3):299–307
18. Isobe Y, Masuyama G, Umeda K (2015) Occlusion handling for target tracking with a mobile robot. In: Proceedings of the fifth Asia international symposium on mechatronics (AISM2015), pp 176–181
19. Ubukata T, Terabayashi K, Moro A, Umeda K (2010) Multi-object segmentation in a projection plane using subtraction stereo. In: Proceedings of the 20th IEEE international conference on pattern and recognition, pp 3296–3299

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com