OCCLUSION HANDLING FOR TARGET TRACKING WITH A MOBILE ROBOT

Yuzuka Isobe[1], Gakuto Masuyama[2], Kazunori Umeda[2]
[1]Graduate School of Science and Engineering, Chuo University, Japan
isobe@sensor.mech.chuo-u.ac.jp
[2]Faculty of Science and Engineering, Chuo University, Japan
masuyama@mech.chuo-u.ac.jp, umeda@mech.chuo-u.ac.jp

Keywords: Occlusion Detection, Target Tracking, Stereo Camera, Mobile Robot, Outdoor Environment

Abstract

This paper addresses a target-tracking method for a mobile robot with the purpose of resolving the occlusion problem. The approach is based on both color and disparity images acquired from a stereo camera. To improve the robustness of the method against occlusion, two new techniques are adopted. First, human candidates are detected based on measured 3D points. This allows the system to detect people who are partly occluded. Second, the occlusion-detection technique is used to carry out tracking procedures appropriate to occlusion states. According to the positional relationship between a target and persons/objects, three occlusion states are defined. The effectiveness of the proposed method is tested through target-tracking experiments in real-world environments. The experimental result is compared with our previous target-tracking method, which is not equipped with any occlusion handling technique. The comparative robustness of the proposed method is demonstrated.

1 Introduction

The ability to track a specific person is fundamental to autonomous mobile robots. The ability can be applied to a wide range of fields from industry to daily life. Carrying baggage for people is the main application of target tracking. Benefits include a reduction in manpower costs, as well as a resolution of the shortage of manpower in, for instance, depopulated areas. Coexisting with humans requires the ability to work in dynamic environments, such as with varying illumination and obstacles or people other than a target. Therefore, how to distinguish a target from other people or how to handle occlusion caused by others and under various illuminations must be considered.
Our previous target-tracking system [1] deals with the problem of illumination changes. High performance of the system has been verified. Therefore, in next subsection, we analyze the challenges in handling occlusion.

1.1 Related work

In the literature, several target-tracking methods have been proposed that consider the presence of multiple people and occlusion. Many approaches for tackling the occlusion problem are based on a time-series filter, such as a Kalman Filter [2,3,4,5]. Basso *et al.* [6] accomplished tracking based on the data association of the models obtained from an Unscented Kalman Filter and AdaBoost. Resolving the problem in accordance with a learning-based algorithm and a particle filter was achieved by Cielniak *et al.* [7]. However, these methods depend on the duration of predictions of time-series filters. Therefore, when the duration exceeds a certain time, tracking becomes weaker. Conversely, in [8], a method that changes the tracking procedure based on the occlusion detected in a situation has been proposed. Although the method does not change with the duration of occlusion, because the method uses only color information as a target's feature, it is weak in an environment where the clothing color of other people is the same as that of the target.

In this paper, we propose an occlusion-handling method for tracking a target independently of the duration of occlusion.
In the next section, we will detail the proposed method. Section 3 reports the results of our target-tracking experiments. As compared with our previous method, the effectiveness of the proposed method is verified. In last section, our conclusions and future works are presented.

2 Details of the proposed system

The proposed system is composed of a stereo camera mounted on a mobile robot. By exploiting both color and disparity images, robust target tracking is achieved, even in cases of varying illumination and occlusion. The target-tracking system follows three procedures: candidate extraction, target identification, and occlusion detection.
In the first phase, candidate extraction, the candidate regions of people in a 3D space are extracted by using a disparity image. To identify the region corresponding to a target from the candidate regions, a target model is created. The model consists of both the color and positional information of a target. After identification of a target using the model, occlusion is detected by the positional relationship between a target person and other objects. If this identification is not done, the predicted position of a target is used to determine whether occlusion is occurring. Analysis determines whether the predicted position of a target overlaps other people/objects. Occlusion detection results are classified into three types of occlusion states. Depending on each occlusion state, appropriate procedures for updating the model and extending the duration of Kalman Filter are executed.

2.1 Candidate extraction

In our previous system, candidates of people were detected by the segmentation method [9]. Because the density of the projected 3D points corresponding to human regions is high on a plane, the method uses an overlooked plane, onto which a point cloud in a 3D space acquired from a stereo camera is projected. However, human regions can be successfully extracted in a restricted condition when no occlusion occurs and the entire region of the person (from head to foot) is in a disparity image. If such is not the condition, the density lowers.

This problem might lead to frames in which a target is not considered to be a human candidate, even when parts of the target region are still visible in the images. In this paper, we use the term *no-detection frames* to refer to frames where a target is not detected due to partial or total occlusion. Especially in populated environments, objects and persons can possibly occlude one another. Under long-term occlusion, *no-detection frames* are expected to increase, leading to an unsteady estimation of a target's position.

Therefore, the candidate-extraction method is proposed, so that *no-detection frames* are decreased in the case of the partial occlusion of a target. The density, not on an overlooked 2D plane but in a 3D space, is used to extract human candidates. By using a point cloud in a 3D space, robustness is developed in situations where the entire target's region is not visible due to partial occlusion. A 3D space is defined by a body coordinate system, as shown in Figure 1. The steps of candidate extraction are explained as follows: first, the point cloud in the space is obtained. Figure 2 depicts an example of a captured image. The point cloud corresponding to three people in the figure is shown in Figure 3. Then, to simplify the segmentation process, the space is divided into boxes. In each box, the number of points is counted. The human region is mostly composed of the boxes with the largest numbers. Therefore, the boxes in which the number exceeds a certain threshold are extracted (see Figure 4) and labeled.

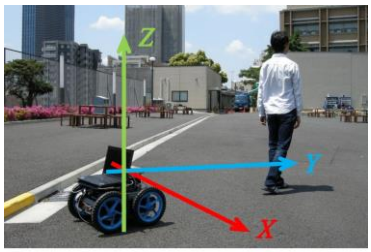


Figure 1. X-Y-Z coordinates of the proposed system

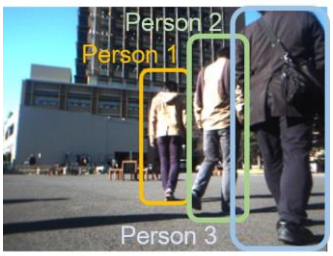


Figure 2. Example of color image: There are three people in the image.

Note that the labeling of four-connected components is implemented to reduce the computational costs. The groups consisting of labeled boxes are merged or split by mean-shift clustering. After clustering, each group of boxes is acquired. In the last step of candidate extraction, candidate groups are determined based on height, width, and depth. Figure 5 shows the result of the extraction; three groups of boxes are considered to be the human candidate.

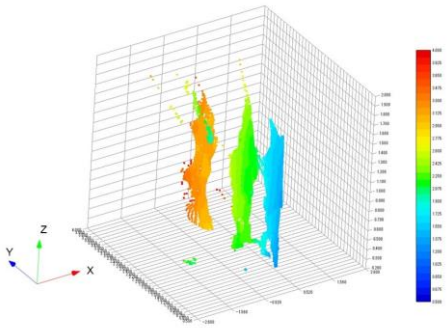


Figure 3. Point cloud; The blue points mean close to the camera, and the red ones mean far from the camera.

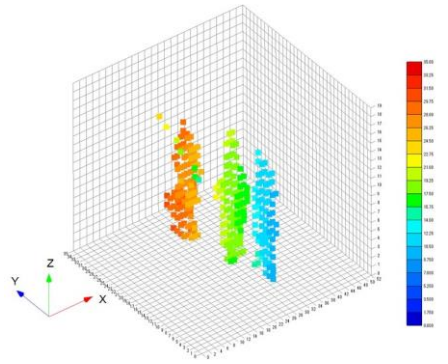
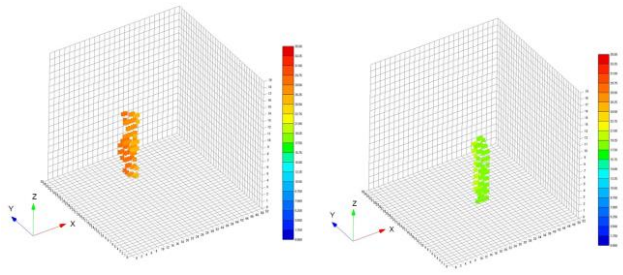
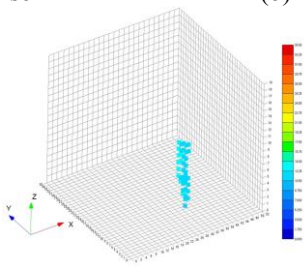


Figure 4. Divided boxes; The colors of the boxes are corresponding to Figure 3.



(a) Person 1 (b) Person 2



(c) Person 3

Figure 5. Result of labeling and mean-shift clustering of the point cloud

2.2 Target identification

For target identification, both the color and positional information of the candidate is used. The previous target-identification method is then applied. The method evaluates the dissimilarities between a target and detected persons, and the information is integrated according to a parameter of illumination changes. The parameter is defined as changes in the white balance obtained by a stereo camera. The method allows target tracking to be executed under varying illumination environments.

The values of a color image, corresponding to the points in the boxes of the candidates, are used to build hue-saturation histograms. The dissimilarity of the color distribution between each candidate's and a stored target's histograms are calculated as follows:

$$R_{color} = \sqrt{1 - \sum_{h}\sum_{s}\sqrt{H_{input}(h,s)H_{template}(h,s)}} \quad (1)$$

where $H_{input}(h,s)$ and $H_{template}(h,s)$ are normalized histograms of the hue (h) and saturation (s) of each candidate and a target, respectively.

Positional information of each candidate is determined by the centroid of the region on the X-Y plane. Each candidate's position is compared with the estimated position of a target. For comparison, the Euclidian distance is used as the dissimilarity of positions, which is computed as follows:

$$E = k\sqrt{(X_s - X_e)^2 + (Y_s - Y_e)^2} \quad (2)$$

where (X_s, Y_s) is each position of a candidate, (X_e, Y_e) is the estimated position of a target, and k is a certain constant for transforming E to a dimensionless number.

These dissimilarities of color and position are weighted according to changes in illumination. The total dissimilarity is calculated by the sum of these weighted dissimilarities. Target identification is achieved by the total dissimilarity.

2.3 Occlusion detection

Based on the probability that occlusion is occurring or will occur in some frames, the occlusion state is classified into three types: STATES 1, 2, and 3. To deal with target tracking in every state, the occlusion state should be detected. In each state, the proper procedure must be adopted. Classification of occlusion states is achieved by using the relationship between a target and others on an overlooked plane, as shown in Figure 6. In this figure, every blue rectangle indicates a stereo camera, and the areas inside the dotted lines are fields of view of the lenses attached to a stereo camera.

(1) STATE 1

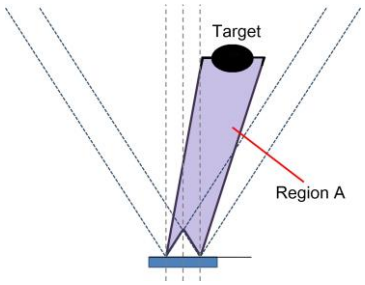
In Figure 6(a), Region A is defined as the inner region built by lines that connect the lens to the left and right edges of the target's region with margins. STATE 1 is defined as when no object/person is present in Region A. In this state, since a target is not occluded, the possibility of obtaining a target's true color model is high. Therefore, a procedure for updating the target's model is applied whenever the total dissimilarity is below a certain threshold.

(2) STATE 2

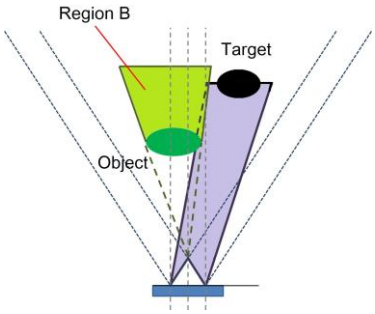
If objects violate Region A, a target is or will be partly/totally occluded. When there are objects in Region A, Region B is given, as shown in Figure 6(b). Region B indicates the region in which objects between a target and a robot can occlude. When someone/something is in Region A but a target is correctly identified, this is regarded as STATE 2.

(3) STATE 3

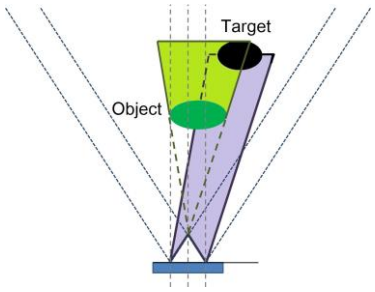
If a target is not identified and the estimated position of a target overlaps with Region B, it is considered that, because of occlusion, a target model cannot be acquired. This is defined as STATE 3 (see Figure 6(c)). In the case of STATE 3, it is assumed that the occlusion causes *no-detection frames*. The duration of target-position estimation by Kalman filter is extended until the estimation position is moving out of Region B. This allows for re-identification after long-term occlusion.



(a) STATE 1: No occlusion occurs.



(b) STATE 2: Partial/Total occlusion occurs or will occur, but a target is still correctly identified



(c) STATE 3: A target cannot be identified because of occlusion.

Figure 6. Details of occlusion states

3 Experiments

Experiments have been conducted to compare the performance of the proposed method with that of the previous method. The new method was tested by the use of images that had been prospectively captured in environments where the illumination varied and multiple people were present. During capturing process, a human operator controlled a robot (Segway Japan, Blackship) to follow a target. A stereo camera, Bumblebee2 (Point Grey Research), was used to capture both color and disparity images.

An entire experiment was composed of 2184 frames that were captured and saved at 6.9 fps. We classified the environments into seven scenes according to the illumination. The details of the scenes are explained in Table 1 and Figure 7. In this table, *condition* means the lighting condition (e.g., *back* means "back lighting"); *number of people* indicates how many other people were present in one frame, on average; *occlusion* indicates the number of occlusions, the average number of occluded frames, and the maximum number of occluded frames; *shadow* indicates how often a shadow appeared. In both methods, the allowable duration of the situation when no one is identified as a target, is fixed to five frames. In the previous method, this means that occlusion is allowed to continue until five frames. In the new method, when the occlusion state is not in STATE 3 and a target is not identified, the five-frames restriction is adopted. During five frames, in both methods, a target's position has been estimated and updated. After the frames, a positional model of a target is initialized.

The previous and proposed methods are evaluated by results of target-tracking experiments using acquired images. For the evaluation, the following values, *precision* and *recall*, represent accuracy and completeness, respectively.

$$Precision = \frac{A}{A+B}, Recall = \frac{A}{A+C}.$$

A: The number of frames in which the target is correctly identified.

B: The number of frames in which a non-target is identified.

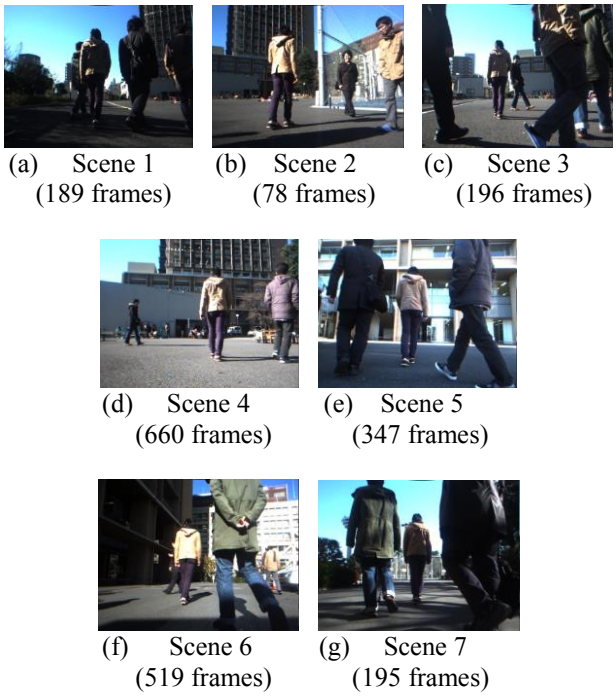
C: The number of frames in which no objects are identified as the target despite the target's presence (*no-detection frames*).

The calculation results of *precision* and *recall* are shown in Table 2. By using the proposed method, both evaluation values improved, as compared with those of the previous method. This is due to the fact that even when partial occlusion occurs, a target can be detected as a candidate by the proposed candidate-extraction method. This caused the estimation of a target's position to be steady. One such instance is shown in Figure 8. In the figure, the red rectangle indicates the identified target region. In this situation, a person who is closer to a robot than a target occludes half of the target region. Most of the upper body is occluded. Therefore, the previous method failed to detect the target even as a candidate because it uses the density of the points on the overlooked plane. However, in the proposed method, using the density in a 3D space is effective. The region can be determined to correspond to a candidate, resulting in correct target identification. Furthermore, occlusion detection causes target-tracking to be irrespective of the duration of occlusion. In all scenes, when using the proposed method, the *precision* values are computed at 100%.

The accuracy of occlusion detection is calculated as shown in Table 3. There are four reasons for incorrect detection.

Table 1. The details of experimental scenes

Scene	Condition	Number of people	Occlusion number of occlusion/ average frames/ maximum frames	Shadow
1	back	1.8	5 5 11	no appearance
2	direct	1.9	4 5 7	no appearance
3	side	1.1	8 4 7	no appearance
4	direct	1.1	15 7 14	no appearance
5	side	1.2	7 7 12	continuous appearance by buildings
6	direct	1.3	14 7 17	no appearance
7	side	1.6	3 10 21	continuous appearance by trees



(a) Scene 1 (189 frames) (b) Scene 2 (78 frames) (c) Scene 3 (196 frames)

(d) Scene 4 (660 frames) (e) Scene 5 (347 frames)

(f) Scene 6 (519 frames) (g) Scene 7 (195 frames)

Figure 7. Off-line experimental scenes; The scenes are classified according to illumination environments

1. Incorrect disparity image: disparity information of the persons/objects could not be correctly acquired. This occurs when occluding persons/objects were too close or an occurring smear covered the target in images.
2. Failure of tracking: a target is lost because the target's position was incorrectly estimated. A target is not identified despite the target's presence in images.
3. Failure of perceived objects: Even though nothing was present, something was perceived. This might be caused by noise from a stereo camera. The occlusion state was classified as STATE 2.
4. Incorrect clustering of a target's region: Region B was built by some parts of a target itself. It is considered that some parts of a target occlude other parts.

In the experiment, the rate of the reasons of incorrect detection was calculated (see Figure 9). This shows that the main reason for misdetection is incorrect sensing, which is caused by the measurement principle of a stereo camera. In the image obtained from one camera, there is an occluding person, but in the image from the other one, there is not. In such a case, disparity is not acquired. This situation is different from a situation in which occlusion is not occurring. By fusing other sensors, measurements might be carried out more completely, even where a stereo camera cannot measure.

4 Conclusion

In this paper, a target-tracking system with a mobile robot is proposed that is robust despite occlusion. Occlusion detection and the implementation of the proper procedure based on the detected occlusion state lead to target tracking, irrespective of the duration of occlusion. In the experiments, higher evaluation values could be obtained with the proposed method than with the previous method.

In our future work, we will address the problem of unmeasurable boundaries due to the principle of stereo. Furthermore, target tracking in environments with greater changes in illumination and more cluttered environments is also our goal.

Table 2. Evaluation of experimental results in outdoor scenes

Scene	Proposed method		Previous method	
	precision [%]	recall [%]	precision [%]	recall [%]
1	100	100	100	87.6
2	100	100	100	77.1
3	100	98.9	100	87.9
4	100	98.9	99.6	83.2
5	100	97.1	100	68.8
6	100	83.5	97.0	56.3
7	100	99.4	100	70.6

Figure 8. Result of target detection under occlusion

Table 3. Accuracy of the determination of occlusion state

Scene	Accuracy [%]
1	98.4
2	100
3	96.9
4	98.0
5	97.1
6	93.3
7	99.5
Total	96.9

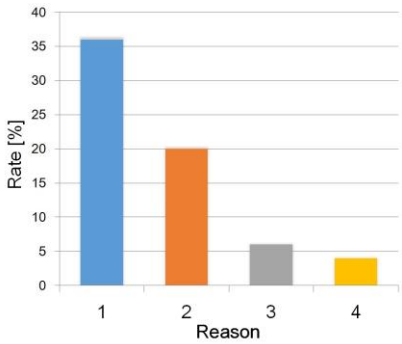


Figure 9. Frequency of the reasons for misdetermination. 1 is the reason of incorrect disparity images, 2 is that of failure of tracking, 3 is that of failure of perceived objects, and 4 is that of incorrect clustering of a target's region.

5 References

[1] Isobe, Y., Masuyama, G., and Umeda, K., "Target Tracking for a Mobile Robot with a Stereo Camera Considering Illumination Changes," Proc. of the 2015 IEEE/RSJ Int. Conf. on Intelligent Robotics and Systems, 2015 (in press).

[2] Satake, J., Chiba, M., and Miura, J., "Visual Person Identification Using a Distance-dependent Appearance Model for a Person Following Robot," Int. Journal of Automation and Computing, Vol. 10, Issue 5, pp. 438–446, 2013.

[3] Bellotto, N. and Hu, H., "Computationally Efficient Solutions for Tracking People with a Mobile Robot: an Experimental Evaluation of Bayesian Filters," Int. Journal of Autonomous Robots, Vol. 28, Issue 4, pp. 425–438, 2010.

[4] Chakravarty, P. and Jarvis, R., "Panoramic Vision and Laser Range Finder Fusion for Multiple Person Tracking," Proc. of the 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 2949–2954, 2006.

[5] Germa, T., Lerasle, F., Ouadah, N., and Cadenat, V., "Vision and RFID Data Fusion for Tracking People in Crowds by a Mobile Robot," Int. Journal of Computer Vision and Image Understanding, Vol. 114, Issue 6, pp. 641–651, 2010.

[6] Basso, F., Munaro, M., Michieletto, S., Pagello, E., and Menegatti, E., "Fast and Robust Multi-people Tracking from RGB-D Data for a Mobile Robot," Journal of Intelligent Autonomous Systems 12, Advances in Intelligent Systems and Computing, Vol. 193, pp. 265–276, 2013.

[7] Cielniak, G., Duckett, T., and Lilienthal, A.J., "Data Association and Occlusion Handling for Vision-based People Tracking by Mobile Robots," Int. Journal of Robotics and Autonomous Systems, Vol. 58, Issue 5, pp. 435–443, 2010.

[8] Bai, P., Qiao, H., Wan, A., and Liu, Y., "Person-Tracking with Occlusion Using Appearance Filters," Proc. of the 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1805–1810, 2006.

[9] Ubukata, T., Terabayashi, K., Moro, A., and Umeda, K., "Multi-object Segmentation in a Projection Plane Using Subtraction Stereo," Proc. of the 20th IEEE Int. Conf. on Pattern and Recognition, pp. 3296–3299, 2010.