2015 IEEE/SICE International Symposium on System Integration (SII)
December 11-13, 2015. Meijo University, Nagoya, Japan

Target Tracking for a Mobile Robot with a Stereo Camera Considering Illumination Changes

Yuzuka Isobe, Gakuto Masuyama, and Kazunori Umeda

Abstract— **Tracking a specific person in dynamic environments is a fundamental task of mobile service robots. Image information is essential to identify a target person, however, the information is not reliable under varying illumination. In this paper, we propose a target-tracking system using a stereo camera. Color and location information is used for the target's feature, which is useful to distinguish a target from the other people. An evaluation value to identify a target is defined as weighted sum of the color and location features. The weight to the features is derived from a parameter of illumination changes. The parameter of illumination changes provides the system with capability of robust tracking even under varying illumination. We confirmed robustness of the proposed system through target-tracking experiments in outdoor environment where the lighting condition changes extremely.**

I. INTRODUCTION

Tracking a specific person is an essential ability for mobile service robots. Robots that have this ability are expected to be applied in offices, shopping centers [1], military areas [2], golf courses [3] and so on. For the realization of tracking a specific person in such dynamic environments, robots are required to be aware of other people's presence [4].

In order to perform the tasks in real-world unstructured and dynamic environments, robots must have sufficient perceptual capabilities. There are various sensors and sensor modalities that give robots such capabilities [5], [6], [7].

One of the most common sensors used for human tracking are cameras. Numerous methods for target tracking with a mobile robot are based on color information. Some of them use only color information in order to extract a target region [8], [9], [10], [11]. Because color information is easily affected by illumination changes, these methods might be prone to causing mis-tracking. Hu *et al.* [12] illustrate a target-tracking system in which, using both color and edge information, the extraction of target region is achieved. The approach proposed by Chakravarty *et al.* [13] is based on extraction of the candidate regions of humans with a laser range finder and comparison of the colors of the candidates and the color of a target obtained from a panoramic camera. These methods abate tracking errors but still are not robust under changing illumination.

In [14], Takemura *et al.* accomplish the task in both indoor and outdoor conditions by a combination of location information from a laser range finder and color information from a stereo camera. Based on a combination of four types of percepts (faces, torsos, sound sources, and legs), Fritsch *et al.* [15] carry out tracking. The system has not been applied in an outdoor environment because the environment where each percept can be acquired is severely restricted. The solution based on fusion of thermal and color cameras is demonstrated by Cielniak *et al.* [16]. A thermal camera is used for human detection and gives a contour model of humans. Identification of each person is achieved by the appearance model of the color distribution, which is robust under changing lighting conditions. Furthermore, an occlusion detection method is investigated. This system is verified through a tracking experiment in cluttered indoor environments. Thermal information is proper for recognizing humans but may not allow robots to be used in outdoor applications.

Contrary to the methods using color information that are affected by illumination changes, Satake *et al.* [17] adopt stereo-based human detection. A target is identified by using the scale-invariant feature transform (SIFT) features of the target's clothing texture, which is resilient to changes in lighting conditions. However, there is the problem of computational costs. Petrovic *et al.* [18] also describe a stereo vision-based method using only 3D information for human detection and tracking. This method can be used in indoors and outdoors, but in the experimental environments, there are no people other than the target. It might be difficult to apply the method in crowded environments, because it does not use any feature specific to the target person.

In this paper, a target-tracking system for a mobile robot is proposed, focusing on a problem of lighting conditions: in order to be applied in dynamic real environments, the system is required to be robust in spite of illumination changes. We address these problems using a stereo camera, which can offer robust measurement under changing illumination. A stereo camera has the advantage of capturing both disparity and color images simultaneously. The system is based on both color and 3D information obtained from a stereo camera. For adjusting to illumination changes, color and location features are weighted and combined according to changes in the lighting conditions.

The rest of the paper is organized as follows. Section II explains the algorithm of our method. Section III presents human-following experiments outdoors, to compare the proposed system with other six methods and test the proposed system by controlling a mobile robot on-line. Finally, conclusions and future works are shown in Section IV.

Yuzuka Isobe is with the School of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, Japan isobe@sensor.mech.chuo-u.ac.jp

Gakuto Masuyama and Kazunori Umeda are with the Faculty of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, Japan {masuyama, umeda}@mech.chuo-u.ac.jp

978-1-4673-7241-1/15/$31.00 ©2015 IEEE

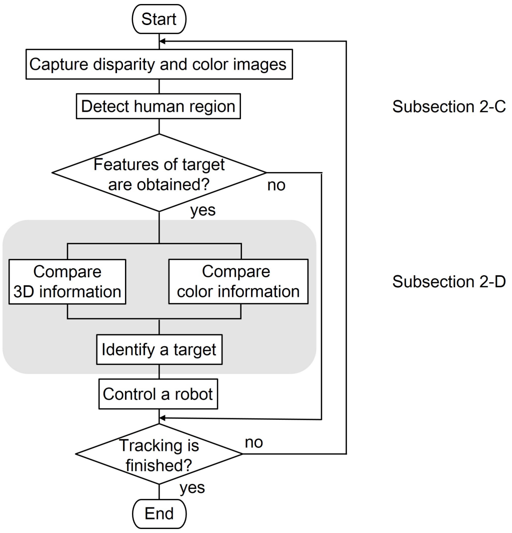


Fig. 1. Flow chart of the proposed system

II. TARGET TRACKING ALGORITHM

A. System Overview

The proposed system utilizes a stereo camera attached to a mobile robot. The algorithm for tracking a target is shown in Fig. 1, which is an improved system based on our previous one [19]. First, disparity and color images are acquired by a stereo camera. A disparity image, which can be captured with little influence from sunlight and illumination changes, is used for human detection. When objects are identified as humans, the color and location information of each person is compared with that of a target person which is predetermined. In order to make this system resilient to illumination changes, the dissimilarity between detected humans and a target human is given by fusion of both types of information based on lighting conditions. A robot is controlled according to the angle and the distance between the robot and the target. Repeating these processes, continuous tracking is achieved.

B. Problems of the Previous Method

Our previous target-detection method uses only color information, hue and saturation. Though hue and saturation are relatively consistent under illumination changes, they are affected by extreme and sudden changes in lighting conditions and white balance, so it is difficult to distinguish a target from the others. Fig. 2 shows one example of extreme white-balance changes. After Fig. 2(a) had been captured, the next frame was captured as shown in Fig. 2(b) with the interval of about 0.05 s. The influence of white-balance changes on color information can also be observed in Fig. 3. In this graph, the blue line shows degrees of dissimilarity between the color information of a calculated target on each frame and that of a predetermined target, and the changes in white balance are given as a red line. White balance has two parameters, i.e., red and blue gains, which have integer values from 0 to 1023, in a color image captured by a stereo camera, Point Grey Research Bumblebee2. The change in white balance is calculated by adding the amount of the changes of the red and blue gains between the first and the current frames. As the white-balance changes severely, the dissimilarity tends to change as well.

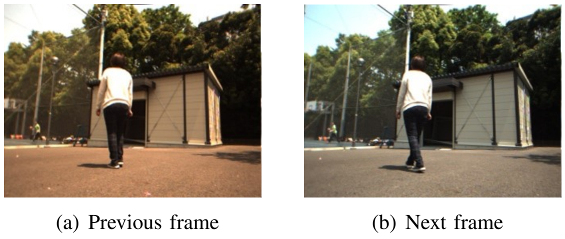

(a) Previous frame
(b) Next frame

Fig. 2. Example of extreme white-balance changes

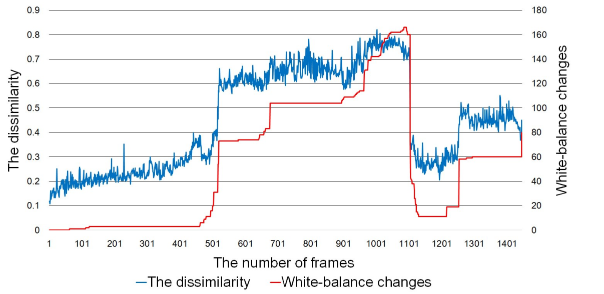


Fig. 3. The influence of white-balance changes on color information

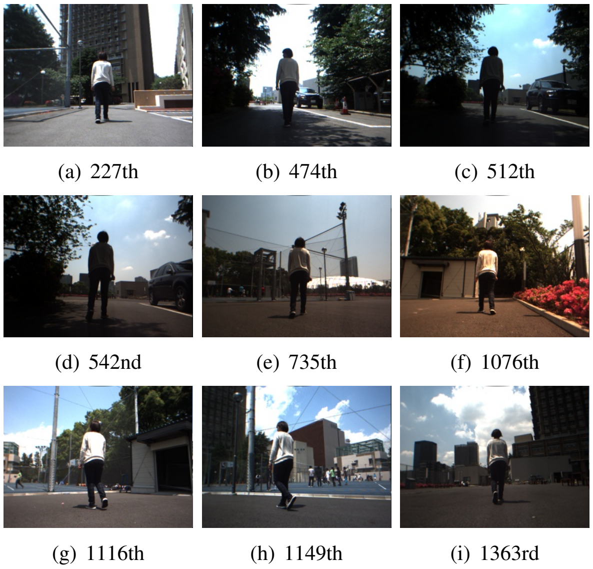
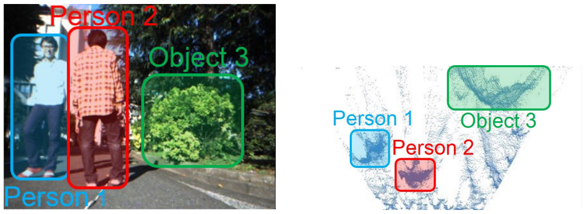

(a) 227th
(b) 474th
(c) 512th
(d) 542nd
(e) 735th
(f) 1076th
(g) 1116th
(h) 1149th
(i) 1363rd

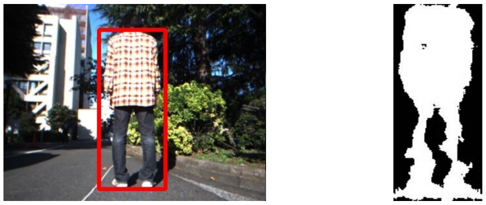
Fig. 4. Scenes with varying illumination

Additionally, Fig. 4 shows the color images at each number of the frames in Fig. 3. Based on these figures, we assume that white-balance changes reflect the changes in illumination, and by using the changes in white balance, the resilience of the target-tracking system to lighting conditions is improved. Therefore, in the proposed system, the changes in white balance are adopted as the parameters that show illumination changes.



(a) Input image (b) Projected points

Fig. 5. Extraction of the candidates of human regions



(a) Example of extraction of the candidates (b) Binary image

Fig. 6. Process of human detection

C. Human Detection

In order to extract the object regions, the segmentation method [20] is adopted. 3D information of each pixel of a disparity image is projected onto an overlooked plane (see Fig. 5). Fig. 5(b) is the plane of an input image as shown in Fig. 5(a). The candidates of human regions are extracted according to the density of the projected points. Then, using a disparity image, the contours are depicted as shown in Fig. 6(b), given by the disparity image of the candidate region indicated by a red rectangle in Fig. 6(a). Thus, human regions are detected based on the contours of the candidate objects.

D. Target Detection

Once people have been detected, these regions are shown as rectangles in the color image. The color information of each region is extracted based on the binary image which is given by the disparity image of the region. Hue and saturation that are resilient to illumination changes are used as the color information for the regions. The dissimilarity of color information between detected humans and a target human is calculated as follows:

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

where $H_{input}(h,s)$ and $H_{template}(h,s)$ are a histogram of the hue(h) and saturation(s) of the input and template information, respectively. Additionally, at intervals of a few frames, the color information of a target is compared with preregistered one, and if the dissimilarity is under threshold, the color information is updated.

In addition to color information, another component of the proposed target-detection method is the location information of detected humans and a target human. The location information is given by X, Y values in the robot coordinate (see Fig. 7). A Kalman filter is adopted to estimate the latest X, Y values of a target's position from the previous values. The dissimilarity of the information between detected humans and a target is computed as follows:

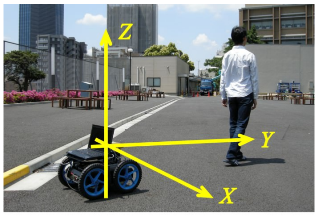


Fig. 7. Robot coordinate system

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

where (X_s, Y_s) is the latest position of the humans and (X_e, Y_e) is the estimated position of a target, and k is adopted to transform E to a dimensionless number (usually $k = 1.0 \text{ m}^{-1}$).

R_{color} and E are the evaluation values of color and location information, respectively. If these values are under the respective color and location thresholds, the total dissimilarity between detected humans and a target human is calculated. The total dissimilarity is defined as follows:

$$D = \begin{cases} (1-\alpha)R_{color} + \alpha E & (\alpha < \alpha_{th}) \\ E & (otherwise) \end{cases}, \tag{3}$$

where α is the parameter that represents illumination changes and has the relation $\alpha = p|W|$, where W is the amount of the white-balance change, and p is a constant. Value of p is determined so as to hold the relation $0 \leq \alpha < 1$. In (3), α_{th} denotes the threshold of illumination change. When the illumination changes so much that the color information changes significantly, α is equal to or more than α_{th}. The human region with the smallest D value is considered to be the target region, if the D value is under a certain threshold. The effect of the relationship between α and α_{th} against D is interpreted below.

(1) In the case of $\alpha < \alpha_{th}$

The evaluation values of the human regions are weighted according to the amount of white-balance change. This is because the reliability of the color information is changed by illumination variation. If the illumination varies little, the color information is reliable. However, under varying illumination, since the color information (particularly hue value) changes easily, the reliability may significantly decrease. On the other hand, the location information is robust to illumination changes. However, using it alone causes target tracking to be difficult if the behavior of a target does not accord with the prediction model. Consequently, by using both color and location information, the resilience in illumination changes can be improved compared with only

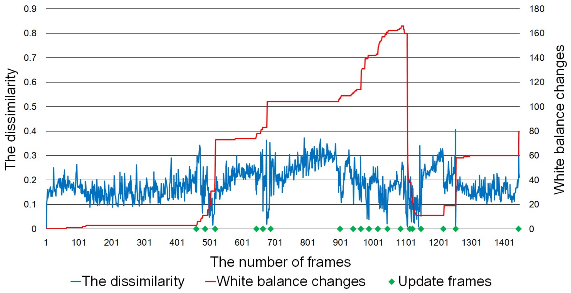


Fig. 8. Influence of white-balance changes on the proposed system. The blue line gives the D value, and the red line does the white-balance changes. The green marks show the frames in which the color template is updated.

one information.

(2) In the case of $\alpha \geq \alpha_{th}$

The color information is supposed to be unreliable in this case (see Fig. 3). Therefore, the dissimilarity is given by only location information. In order to prevent the tracking from depending solely on location information, the template color information is updated to adjust to the lighting condition if D is under the threshold. When the template color information is updated, white balance is also registered. Then, $|W|$ is calculated as an absolute value of the difference between white balance of a current frame and the registered one.

Profile of the D value of a target is depicted in Fig. 8. It is obtained off-line in the same environments as shown in Fig. 4. The green marks show the frames in which the color template is updated. As shown in Fig. 8, the template was updated 19 times. It can be seen that the value is given independently of the changes in white balance.

III. EXPERIMENTAL RESULTS

A. *Off-line Experiments under Illumination Environments*

Firstly, we tested the performance of the system by comparing with six settings given in Table I, in outdoor environments where illumination varies. For all settings, each threshold is given as same value. And, for the proposed equation (I), the dissimilarity is calculated with $k = 1.0$ m^{-1}, $\alpha = 0.25 \times |W|$, and $\alpha_{th} = 0.8$. The environments are classified into six illumination scenes, as shown in Table II and Fig. 9. In this table, *Condition* means the lighting condition (e.g. back means back lighting); *Number of people* shows how many the other people were present at a frame on average; *Shadow* indicates how often the shadow appeared.

The paths of a mobile robot (Segway Japan, Blackship) with a stereo camera (Point Grey Research, Bumblebee2) were controlled manually as following a target. Average frame rate was 6.9 fps [1]. The effectiveness of each method is verified by two evaluation values, *Precision* and *Recall*, which represent accuracy and completeness, respectively.

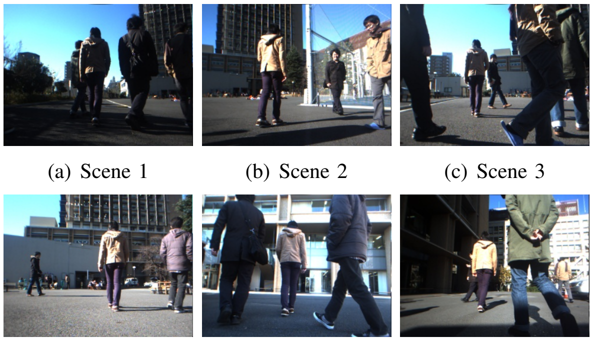
$$Precision = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}} \qquad Recall = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{C}} \tag{4}$$

TABLE I
COMPARED SETTINGS OF FEATURES AND DISSIMILARITY EQUATIONS

	Feature	Dissimilarity equation
I	Color and Location Information	(3)
II	Color Information (frequently updated)	R_{color}
III	Color Information (no updated)	R_{color}
IV	Location Information	E
V	Color and Location Information	$0.9R_{color} + 0.1E$
VI	Color and Location Information	$0.5R_{color} + 0.5E$
VII	Color and Location Information	$0.1R_{color} + 0.9E$

TABLE II
THE DETAILS OF OFF-LINE EXPERIMENTAL SCENES

Scene	Condition	Number of people	Occlusion the number of occlusion/ the average frames/ the maximum frames	Shadow
1	back	1.8	5.0	no appearance
			5	
			11	
2	direct	1.9	4	no appearance
			5	
			7	
3	side	1.1	8	no appearance
			4	
			7	
4	direct	1.1	15	no appearance
			7	
			14	
5	side	1.2	7	continuous appearance by buildings
			7	
			12	
6	direct	1.3	14	no appearance
			7	
			17	



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Scene 6

Fig. 9. Off-line experimental scenes. The scenes are classified according to illumination environments

A: The number of frames in which the target is correctly detected.
B: The number of frames in which a non-target is detected.
C: The number of frames in which no objects are detected as a target.

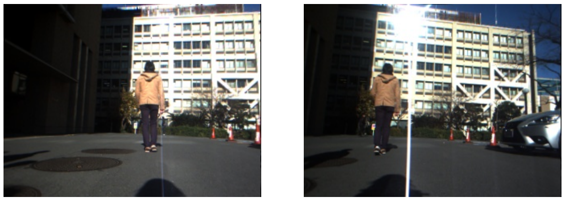
Table III is the result of the calculation of *Precision* [%] and *Recall* [%]. If a denominator of the value is computed as zero, it is indicated by "*". The bold letters show the results, which well indicate the usefulness of the proposed method.

[1] All captured color and disparity images were saved for off-line experiments in this procedure. Note that the frame rate was lowered by limitation of data transfer speed to HDD.

TABLE III

RESULTS OF PRECISION AND RECALL IN OFF-LINE EXPERIMENTS.

	I	II	III	IV	V	VI	VII
Scene 1 (91 frames)							
Precision [%]	100	80.0	81.6	100	100	100	100
Recall [%]	93.8	74.1	67.8	93.8	93.8	93.8	93.8
Scene 2 (78 frames)							
Precision [%]	100	60.0	66.7	83.3	100	100	100
Recall [%]	**88.3**	16.1	6.8	53.6	**73.0**	**73.3**	**73.3**
Scene 3 (196 frames)							
Precision [%]	100	16.7	75.8	89.6	100	100	100
Recall [%]	93.8	6.6	63.9	78.4	93.8	93.8	93.8
Scene 4 (660 frames)							
Precision [%]	99.6	0.0	0.0	99.8	99.6	99.6	99.6
Recall [%]	87.9	0.0	0.0	77.8	87.9	87.9	82.0
Scene 5 (347 frames)							
Precision [%]	100	0.0	*	100	100	100	100
Recall [%]	73.0	0.0	0.0	72.3	73.0	72.7	73.0
Scene 6 (356 frames)							
Precision [%]	98.9	*	*	93.1	100	100	100
Recall [%]	**67.0**	0.0	0.0	58.5	**59.8**	**59.4**	**59.4**
Mean and Standard Deviation (SD)							
Mean of *Precision*	**99.7**	31.3	56.0	94.3	**99.8**	**99.8**	**99.8**
SD of *Precision*	**0.4**	32.8	32.8	6.3	**0.2**	**0.2**	**0.2**
Mean of *Recall*	**84.0**	16.1	23.1	72.4	**80.3**	**80.1**	**79.2**
SD of *Recall*	**10.0**	26.6	30.4	13.3	**12.5**	**12.7**	**12.2**


(a) High saturation value of target's clothes

(b) Low saturation value of target's clothes

Fig. 10. Example of saturation changes not according to white balance

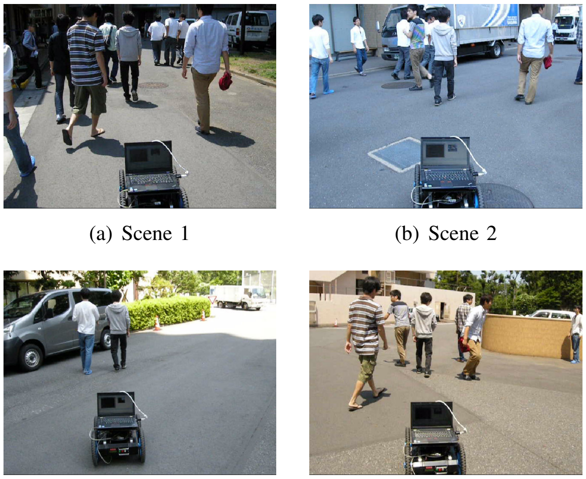
Proposed method (I) demonstrated the highest *Preision* values except in Scene 5. Additionally, in each scene, *Recall* values of the proposed method are higher than that of any other setting.

However, sometimes no object was detected in spite of target's presence, especially in Scene 6. The reason why this problem was occurred is that the changes in white balance do not accord with the changes in the color histogram of a target, as shown in Fig. 10. In the scene, the smear appeared due to reflected sunlight on windows of a building. The smear cause the changes in brightness of color images. The changes in the brightness affected the saturation of target's color. It shows that the changes in white balance do not completely conform to illumination changes.

TABLE IV

THE DETAILS OF ON-LINE EXPERIMENTAL SCENES

Scene	Condition	Number of people	Occlusion the number of occlusion/ the average frames/ the maximum frames	Shadow
1	back	2.7	3 9 12	no appearance
2	back	3.7	2 5 7	continuous appearance by buildings
3	direct	1.2	0 0 0	continuous appearance by buildings
4	direct	2.1	2 8 9	no appearance


(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

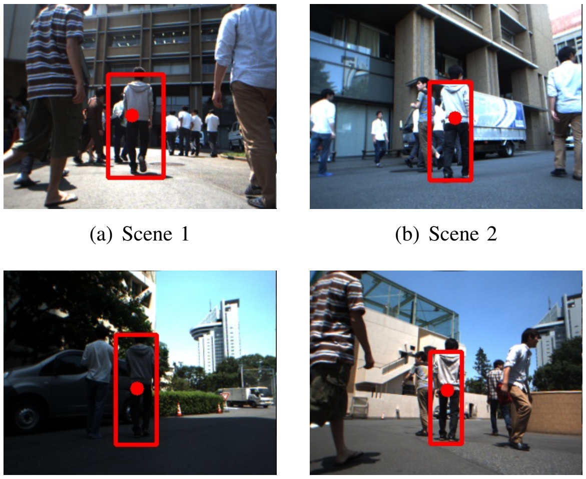
Fig. 11. On-line experimental scenes.These are classified according to illumination environments.

B. On-line Experiment in Outdoor Environments

The proposed system has been tested in outdoor environments where lighting conditions change extremely and multiple people are present. The system has been tested on the mobile robot equipped with the stereo camera. The robot is controlled by a PID controller to adjust the distance from robot to target and the angle computed by their 3D positions. In addition, each parameter in Eq. (3) and the evaluation method are described in section III-A. The effectiveness of the proposed system is verified by (4).

The time length of the trial is 151 s. The experimental environments can be classified into four scenes based on lighting conditions, as shown in Table IV. These scenes are depicted in Fig. 11. Fig. 12 shows the examples of the results of target detection. In these figures, the rectangles represent the target regions, and the dots show the centroids of the white pixels of the binary images, with reference to Fig. 6(b). The evaluation of the experiment is given in Table V. Under all illumination environments, *Precision* and *Recall* values of higher than 91% and 88% are derived, respectively.

Both of these evaluation values in Scene 3 are lower than those of other scenes, because of high frequency of the changes in white balance for a few frames. The changes



(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

Fig. 12. Illustrative frames of target detection in on-line experiments

TABLE V

RESULTS OF PRECISION AND RECALL IN ON-LINE EXPERIMENTS

Scene	Number of frames	*Precision* [%]	*Recall* [%]
1	559	99.4	97.0
2	472	99.8	95.4
3	462	91.6	89.0
4	716	100	98.4

in white balance produced the frames in which no object was detected, despite a target's presence, because the target detection depended solely on location information. With increase of such frames, the prediction of the target's position becomes harder, leading to incorrect detection. Thus, the B and C frames increase, and both evaluation values decrease.

IV. CONCLUSION

In this paper, a target-tracking system for a mobile robot equipped with a stereo camera has been described. By weighting the evaluation values of color and location information of humans and a target based on lighting conditions, robust pursuit of a target is accomplished under varying illumination conditions. Our method demonstrated more than 91% and 88% of the *Precision* and *Recall* values in real outdoor environments. These values are high enough, and it can be said that the system is capable of tracking a specific person under various illumination conditions.

The current system could be further improved by adopting and verifying other parameters that reflect illumination changes or other features of a target in order to implement the system under more severe environments, such as the presence of more people, frequent changes in illumination, and cluttered environments.

REFERENCES

[1] Budgee. Five Elements Robotics. [Online]. Available: http://www.5elementsrobotics.com/

[2] M. Raibert, K. Blankespoor, G. Nelson, R. Playter, and the BigDog Team, "BigDog, the Rough-Terrain Quadruped Robot," *Proc. of the 17th World Congress of the Int. Federation of Automatic Control*, pp. 10822-10825, 2008.

[3] STEWART GOLF X9. Stewart Golf Limited. [Online]. Available: http://www.stewartgolf.com/X9Follow/

[4] D. Katz, J. Kenney, and O. Brock, "How Can Robots Succeed in Unstructured Environments?," *RSS Workshop on Robot Manipulation: Intelligence in Human Environments*, 2008.

[5] M. Hebert, "Active and Passive Range Sensing for Robotics," *Proc. of the 2000 IEEE Int. Conf. on Robotics and Automation*, pp. 102-110, 2000.

[6] L. Mattos and E. Grant, "Passive Sonar Applications: Target Tracking and Navigation of an Autonomous Robot," *Proc. of the 2004 IEEE Int. Conf. on Robotics and Automation*, pp. 4265-4270, 2004.

[7] J. H. Lee, T. Tsubouchi, K. Yamamoto, and S. Egawa, "People Tracking Using a Robot in Motion with Laser Range Finder," *Proc. of the 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 2936-2942, 2006.

[8] M. Doi, M. Nakakita, Y. Aoki, and S. Hashimoto, "Real-Time Vision System for Autonomous Mobile Robot," *Proc. of the 10th IEEE Int. Workshop on Robot and Human Interactive Communication*, pp. 442-449, 2001.

[9] H. Kwon, Y. Yoon, J. B. Park, and A. C. Kak, "Person Tracking with a Mobile Robot using Two Uncalibrated Independently Moving Cameras," *Proc. of the 2005 IEEE Int. Conf. on Robotics and Automation*, pp. 2877-2883, 2005.

[10] D. Calisi, L. Iocchi, and R. Leone, "Person Following through Appearance Models and Stereo Vision using a Mobile Robot," *Proc. of VISAPP (Workshop on Robot Vision)*, pp. 46-56, 2007.

[11] A. Tsalatsanis, K. Valavanis, and A. Yalcin, "Vision Based Target Tracking and Collision Avoidance for Mobile Robots," *Journal of Intelligent and Robotics Systems*, Vol. 48, Issue 2, pp. 285-304, 2007.

[12] C. Hu, X. Ma, and X. Dai, "A Robust Person Tracking and Following Approach for Mobile Robot," *Proc. of the 2007 IEEE Int. Conf. on Mechatronics and Automation*, pp. 3571-3576, 2007.

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

[14] H. Takemura, Z. Nemoto, and H. Mizoguchi, "Development of Vision Based Person Following Module for Mobile Robots In/Out Door Environment," *Proc. of the 2009 IEEE Int. Conf. on Robotics and Biomimetics*, pp. 1675-1680, 2009.

[15] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual Person Tracking with a Mobile Robot," *Proc. of the 2004 Int. Conf. on Intelligent Autonomous Systems*, pp. 898-906, 2004.

[16] G. Cielniak, T. Duckett, and A. J. Lilienthal, "Improved Data Association and Occlusion Handling for Vision-Based People Tracking by Mobile Robots," *Proc. of the 2007 IEEE/RSJ Int. Conf. on Intelligent Robotics and Systems*, pp. 3436-3441, 2007.

[17] J. Satake, M. Chiba, and J. Miura, "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.

[18] E. Petrovic, A. Leu, D. Ristic-Durrant, and V. Nikolic, "Stereo Vision-Based Human Tracking for Robotic Follower," *Int. Journal of Advanced Robotic Systems*, Vol. 10, pp. 1-10, 2013.

[19] Y. Isobe, G. Masuyama, and K. Umeda, "Human Following with a Mobile Robot Based on Combination of Disparity and Color Images," *10th France-Japan and 8th Europe-Asia Congress on Mecatronics*, pp. 84-88, 2014.

[20] T. Ubukata, K. Terabayashi, A. Moro, and K. Umeda, "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.