The Twenty-Fifth International Symposium on Artificial Life and Robotics 2020 (AROB 25th 2020),
The Fifth International Symposium on BioComplexity 2020 (ISBC 5th 2020),
B-Con Plaza, Beppu, Japan, January 22-24, 2020

Robustness Improvement of Cognitive Performance for Human-following Autonomous Mobile Robot

Yu Zhou[1,2], Junxiang Wang[1], Yejun Kou[1], Hongzhi Tian[1], Mamoru Minami[1]

[1]Graduate School of Natural Science and Technology, Okayama University, Japan
(Tel: 81-86-251-8233, Fax: 81-86-251-8233)
[2]Graduate School of Mechanical and Electrical Engineering, Henan University of Science and Technology, China
(Tel: 86-379-64231879, Fax: 86-379-64231879)
[1]pl1x30sd@s.okayama-u.ac.jp

Abstract: Nowadays, the production field is increasingly starting to use the automated guided vehicle (AGV) to assist employees in daily work. The difficulty lies in enabling the robot to recognize the position and size of the moving object in real-time. To meet the demand, we propose a real-time human-following and recognition system for AGV based on visual servoing. Using the dual-eye camera, it can estimate the relative position and size of the target and control AGV to achieve human tracking in real-time. Besides, a Real-time Multi-step Genetic Algorithm (RM-GA) and newly designed projection-based 3D perception (Pb3DP) method are used to improve the robustness of the recognition system against changes of light. The experimental results confirmed that the proposed system could recognize the relative position, detect the size of the target accurately without build complex models in advance, then drive the mobile robot to follow it. Besides, it provides high robustness against disturbances that the influence of the captured camera images under different lighting conditions.

Keywords: Visual servoing, Projection-based, 3D pose estimation, Arbitrary targets, Illumination tolerance, Mobile robot

1 INTRODUCTION

A widespread phenomenon in nature is that most creatures rely on two main eyes to obtain visual information [1]. Since the dual-eye can stereoscopically visualise the information through visual differences even in a changing environment, thereby, animals can locate the target in space, then estimate the size, distance, and the 3D pose from it. However, it is difficult for a robot without stereo vision to achieve that demand, especially if the target object is unique without an artificial marker, and the shape is arbitrary. Further, the object is moving under changing light environment.

To solve these problems, the model-based method is one way using the model of a target object [2]. Although it can detect the distance of the target object from a monocular vision, its accuracy is lower than that of stereo vision [3]. Besides, stereo vision is more sensitive to the object's pose variation than monocular vision. Researches are using RGB-Depth(RGB-D) camera, one RGB camera and depth sensor with infrared light, to improve the distance detection of monocular vision and conduct picking and placing or visual servoing tasks [4]. RGB-D sensors such as the Microsoft Kinect, Intel RealSense, and the Asus Xtion. A depth image is computed by calculating the distortion of a known infrared light pattern which is projected into the scene [5]. These studies still rely on target detection or segmentation from a single image and cannot directly use the depth point cloud for target detection. However, RGB-D camera generates depth point cloud corresponding to the individual image. Therefore, many studies utilise the deep learning methods for target detection [6]-[8]. But, it requires many pictures and pre-training time. Some other studies use model-based method to simplify preliminary preparations [9]. But both of them cannot avoid the disadvantage of RGB-D camera, i.e., missing depth data caused by the depth sensor. Some pixels do not have corresponding depth data [10]. And bright ambient illumination can affect the contrast of infrared images in the active light sensors, resulting in outliers or holes in the depth map [11]. Unlike optical infrared and electric-field sensing, stereo vision is more robust to varying target material properties and light conditions [12]. It is not dependent on capacitance, reflectivity, or other material properties, as long as the target surface has some visible features. For the above reasons, the research proposed in this paper is based on stereo vision, i.e., dual RGB cameras. With a dual-eye camera, we proposed a human-following autonomous mobile robot system based on the new projection-based 3D perception (Pb3DP) method. The non-contact size, distance detection and human-tracking experiments were conducted under different light conditions. The results show that the system is robust to light changes under different lighting conditions during tracking.

2 METHODOLOGY COMPARISON

As mentioned above, RGB-depth(RGB-D) camera method is being used by many researchers. Therefore, this chapter will compare the methodology of the RGB-D camera

©ISAROB

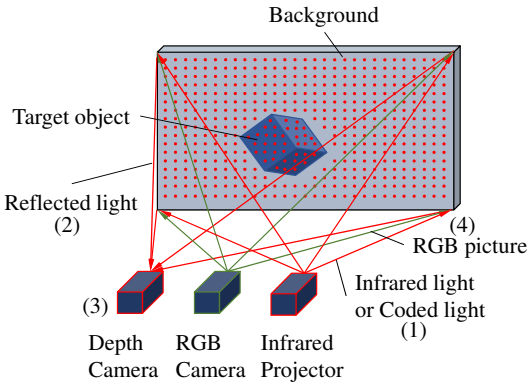

Fig. 1. The schematic of RGB-DC method

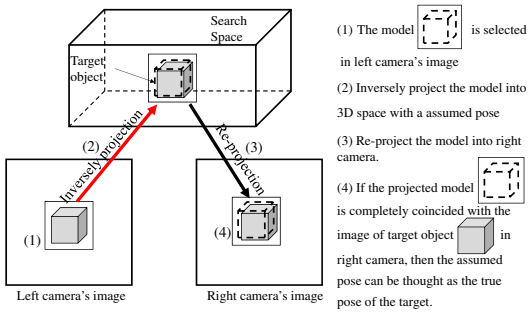

Fig. 2. The schematic of Pb3DP method

method and Pb3DP method.

2.1 RGB-Depth Camera Method Using Stereo Vision

In terms of depth camera technology, Microsoft and Intel are the primary researchers in this field. Its representative products are Microsoft Kinect and Intel RealSence. In principle, the depth camera technology used by the two companies mainly uses the Time-of-Flight (ToF) law, which is shown as Fig.1. Its workflow is as follows: (1) infrared light or coded light is emitted from the infrared projector to the scene(including background and object) to be measured. Then, the infrared receiving camera(depth camera) Fig.1.(3) will capture the reflected light Fig.1.(2). Based on the time it takes the light to travel from the camera to the scene and back, the distance from the camera to the scene can be estimated. This method can obtain the depth map or depth point cloud image of the target scene, and it can read the colour of the target object with its matching RGB camera, Fig.1.(4).

The main advantage of this method is that the depth information in the scene can be discretised, and distance measurement can be performed in a dark environment, which is conducive to extracting objects from the background or removing the background. However, due to the discretisation process, the critical information of depth cloud points in the image may be lost, especially when the object is moving or partially covered, which makes the measurement inaccurate. On the other hand, since the depth information and position information of the images are obtained using different cameras, which leads to poor performance in real-time.

2.2 Projection-based 3D Pose Estimation Method Using Stereo Vision

In the proposed projection-based 3D perception method, the main purpose is to use the image of the arbitrary target's image to estimate it's pose. The schematic is shown as Fig.2, (1) target object is selected in the scene in one of the stereo cameras, (2) the selected 2D target is inversely projected in 3D space with assumed pose, (3) the target in 3D space is projected again into the other camera scene, (4) if the target projected through assumed pose happens to be matched with the real target in the camera scene, then the assumed pose represents real target's pose in 3D space. In addition, Real-time Multi-step GA (RM-GA) [13] is exploited as the optimization method to process the dynamic image.

The main advantage of this method is that as long as an arbitrary target object is selected in the image, the pose of the object relative to the camera coordinate system and the robot coordinate system in space can be known immediately. And the target does not need a specific marker. Then, according to the pose of the recognized object in space, the robot can track the object in real-time, even when it is moving. In other words, it only needs to select any object in the picture to complete the model construction, without prior knowledge or long-term target training process. The object parameters, including the distance and position to the camera or mobile robot and the size, can be obtained immediately. Then these parameters can be more directly used for real-time object dynamic tracking. The disadvantage is that when the ambient light is too dark, the recognition effect will be reduced.

3 HUMAN-FOLLOWING AUTONOMOUS MOBILE ROBOT SYSTEM

3.1 System overview

The human-following autonomous mobile robot system consists of three parts. The sensor part that takes in the image uses two cameras. The camera uses Sony's "FCB-IX11A" camera with a video rate of 30 fps. The traveling part uses a two-wheel-drive cart type mobile robot. The PC for recognition and control uses the PC of Interface, which has 4 PCI slots. An overview of the system is shown in Fig.3.

3.2 Pose and size estimation method

3.2.1 The Establishment of a Model

In the conventional visual servoing method, the model that created beforehand limits the visual servoing system because they can only recognize the assigned target objects. In order to realize the recognition of the arbitrary objects, the models in Pb3DP are designed to be created at any time. In this

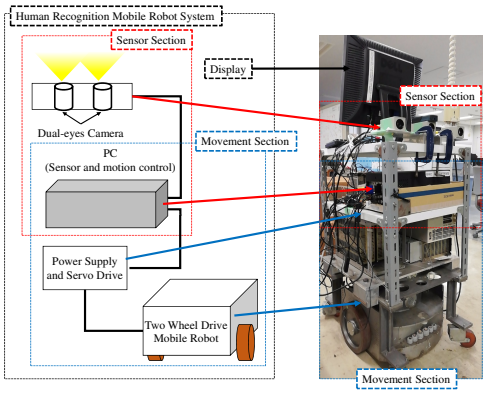

Fig. 3. System overview.

section, the establishment of the model will be described.

Figure 4 shows the process of model building. In the figure, a male is set as a target object. The model used in this method consists of a 2D point cloud, and each sampling point contains the colour information of the image at that point location. The colour information is used to evaluate the recognition results. In Fig. 4 (a), the original image from the left camera is read as a basement to generate a model, and the coordinates of the left camera image are defined as Σ_{IL}. Σ_{IL} 's origin is located in the centre of the left camera image. Select the size and location of the generated model manually. Sample points are then generated in the model area at regular intervals. The arbitrary position of the point in the model created in the left camera coordinate system is specified as ${}^{IL}\boldsymbol{r}_{Mi}^{\ j}$. As shown in Fig.4(b), the human body model is completely contained in the model area. However, since the shape of the model is set to a rectangle and the shape of the target is usually irregular, it is inevitable to include some background in the selected area. Therefore, it is necessary to distinguish the background from the model. Therefore, the model consists of an inner region (S_{in}) and an outer region (S_{out}), where S_{in} represents the target object and S_{out} represents the background. As shown in Fig.4(c), the outer area surrounds the inner area as a subtraction to obtain accurate and accurate recognition results. The outer area is generated at the same regular interval around the inner area.

Unlike the models used in the position-based and image-based method, the model in the Pb3DP method consists of 2D points instead of features, which means that the model can always be adjusted no matter what the target looks like. Besides, the Pb3DP method uses the raw image without any imaging processing to avoid processing time and image distortion that may occur in image processing technology.

3.2.2 The kinematics of stereo-vision

The coordinates of this system are shown as Fig. 5. It is utilised eye-in-hand configuration and two cameras to com-

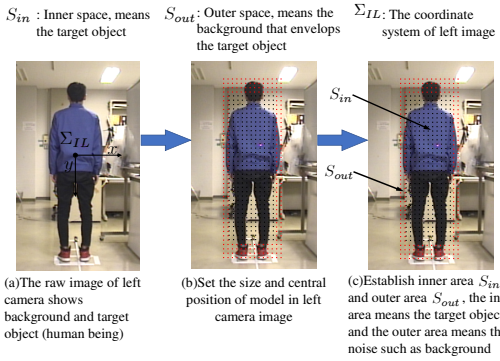

Fig. 4. Model generation process are described as (a)∼(c): (a) shows the raw image in left camera, (b) represents the model area set by assigned central position and size, (c) represents a inner area S_{in} and outer area S_{out} envelops S_{in}

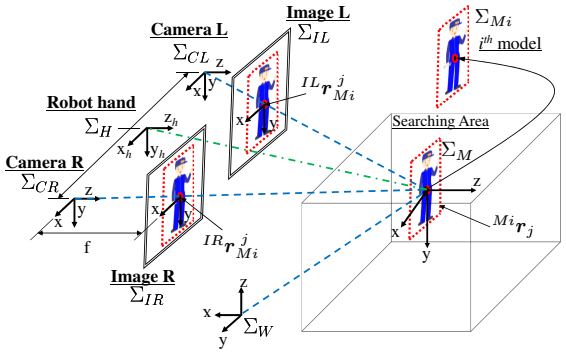

Fig. 5. The coordinate systems of the projection-based method

plete stereo vision. The coordinate systems of two cameras and target object consist of world coordinate system Σ_W, i^{th} model coordinate system Σ_{M_i}, hand position coordinate system Σ_H, left and right camera coordinate system Σ_{CL} and Σ_{CR}, left and right image coordinate system Σ_{IL} and Σ_{IR}, coordinate system of target object Σ_M, and they are shown in Fig. 5. The position vectors of an arbitrary j^{th} point in the i^{th} 3D model coordinate Σ_{Mi} based on each coordinate system are as following:

- ${}^{W}\boldsymbol{r}_{Mi}^{\ j}$: position of an arbitrary j^{th} point on i^{th} 3D model based on Σ_W.
- ${}^{Mi}\boldsymbol{r}_j$: position of an arbitrary j^{th} point on i^{th} 3D model in Σ_{Mi}, where ${}^{Mi}\boldsymbol{r}_j$ is a constant vector.
- ${}^{CL}\boldsymbol{r}_{Mi}^{\ j}$ and ${}^{CR}\boldsymbol{r}_{Mi}^{\ j}$: position of an arbitrary j^{th} point on i^{th} 3D model based on Σ_{CL} and Σ_{CR}.
- ${}^{IL}\boldsymbol{r}_{Mi}^{\ j}$: the position of j^{th} point of i^{th} model in left image coordinate system Σ_{IL}
- ${}^{IR}\boldsymbol{r}_{Mi}^{\ j}$: the position of j^{th} point of i^{th} model in left image coordinate system Σ_{IR}

The homogeneous transformation matrix from the left camera coordinate system Σ_{CL} to the 3D model coordinate system Σ_{Mi} is defined as ${}^{CL}\boldsymbol{T}_M({}^H\boldsymbol{\phi}_{Mi}, \boldsymbol{q})$, where ${}^H\boldsymbol{\phi}_{Mi}$ is i^{th} model's pose based on the robot hand coordinate system Σ_H and $\boldsymbol{q}$ means robot's joint angle vector. The pose of i^{th} 3D model, including three position variables and three orientation variables in quaternion based on Σ_H, are represented as

$$ {}^H\boldsymbol{\phi}_{Mi} = [{}^Hx_{Mi}, {}^Hy_{Mi}, {}^Hz_{Mi}, {}^H\varepsilon_{1Mi}, {}^H\varepsilon_{2Mi}, {}^H\varepsilon_{3Mi}]^T. \tag{1} $$

Meanwhile, the projective transformation matrix is given as following

$$ \boldsymbol{P}({}^Cz_j) = \frac{1}{{}^Cz_j}\begin{bmatrix} f/\eta_x & 0 & {}^Ix_0 & 0 \\ 0 & f/\eta_y & {}^Iy_0 & 0 \end{bmatrix}. \tag{2} $$

Therefore, the arbitrary point of target object naturally projected result in Σ_{IL} and Σ_{IR} can be given as,

$$ \begin{aligned} {}^{IL}\boldsymbol{r}_M &= \boldsymbol{P}({}^{CL}z_j){}^{CL}\boldsymbol{r}_M \\ &= \boldsymbol{P}({}^{CL}z_j){}^{CL}\boldsymbol{T}_H{}^H\boldsymbol{T}_M(\boldsymbol{\phi}_M, \boldsymbol{q}){}^M\boldsymbol{r} \end{aligned} \tag{3} $$

$$ \begin{aligned} {}^{IR}\boldsymbol{r}_M &= \boldsymbol{P}({}^{CR}z_j){}^{CR}\boldsymbol{r}_M \\ &= \boldsymbol{P}({}^{CR}z_j){}^{CR}\boldsymbol{T}_H{}^H\boldsymbol{T}_M(\boldsymbol{\phi}_M, \boldsymbol{q}){}^M\boldsymbol{r} \end{aligned} \tag{4} $$

On the other hand, the inverse projection transformation matrix $\boldsymbol{P}^+$ can be achieve based on Eq.(2) as

$$ \boldsymbol{P}^+({}^Cz_j) = {}^Cz_j\begin{bmatrix} \frac{\eta_x}{f} & 0 & 0 & 0 \\ 0 & \frac{\eta_y}{f} & 0 & 0 \end{bmatrix}^T \tag{5} $$

where, the ${}^Cz_j$ is the distance from the coordinate of Σ_{Mi} to Σ_{CL}, which is assumed by RM-GA.

$$ \begin{aligned} {}^{Mi}\boldsymbol{r}_j &= {}^{Mi}\boldsymbol{T}_{CL}{}^{CL}\boldsymbol{r}^j_{Mi} \\ &= {}^{Mi}\boldsymbol{T}_{CL}\left[\boldsymbol{P}^+({}^{CL}z^j_{Mi}){}^{IL}\boldsymbol{r}^j_{Mi} + (\boldsymbol{I}_4 - \boldsymbol{P}^+\boldsymbol{P})\boldsymbol{l})\right] \end{aligned} \tag{6} $$

Then the position from the perspective of Σ_H to model can be calculated by the following equation:

$$ {}^H\boldsymbol{r}^j_{Mi} = {}^H\boldsymbol{T}^j_{Mi}{}^{Mi}\boldsymbol{r}_j \tag{7} $$

Following the previous step, the upper-left corner coordinates, lower-left corner coordinates, and model size in the model are set as follows:

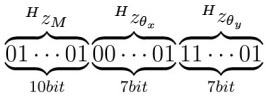

Fig. 6. Gene information

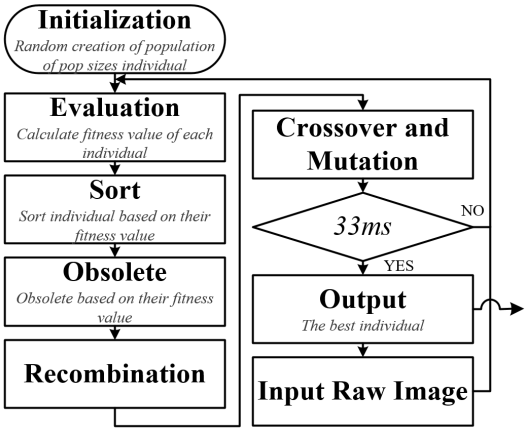

Fig. 7. Flowchart of RM-GA

- (M_{TLx}, M_{TLy}): x and y coordinates of the upper-left corner of the model of the selected model based on Σ_H.
- (M_{BRx}, M_{BRy}): x and y coordinates of the lower-right corner of the model of the selected model based on Σ_H.
- ${}^H\boldsymbol{h}_M$ and ${}^H\boldsymbol{w}_M$: height and width of the target in Σ_H

Then, based on the position of any point in space in the Σ_H, the size of the object can be calculated with Eq.(8) in the selected model .

$$ \begin{cases} {}^H\boldsymbol{h}_M = M_{BRy} - M_{TLy} \\ {}^H\boldsymbol{w}_M = M_{BRx} - M_{TLx} \end{cases} \tag{8} $$

3.3 Real-time Multi-step GA (RM-GA)

In Pb3DP method, searching all possible pose of target object through calculating the fitness value is time-consuming for real-time pose estimation. Therefore, the problem of recognizing the target object's pose can be transformed into a optimization to find the maximum value of fitness. In Pb3DP, we employed Real-time Multi-step GA (RM-GA) to satisfy the real-time recognition in 30 FPS. The reason why we choose RM-GA has been discussed in [13].

In proposed RM-GA, each chromosome includes 24 bits for searching three parameters: ten for position and fourteen for orientation as shown in Fig.6. Figure 7 shows the flowchart of the Real-time Multi-step GA. The RM-GA operation is conducted in the sequence as evaluation, sorting, obsolete, crossover and mutation. These operations are repeated several times in 33[ms] to generate the best individual.

Fig. 8. Initial position of the human target

4 RECOGNITION AND TRACKING EXPERIMENT

To confirm the effectiveness of the recognition and tracking performance of the system, this chapter is mainly divided into two parts. Firstly, the accuracy experiments in the static state were carried out, primarily to determine the measurement accuracy of the system for distance and size. The second part is the tracking experiments under normal light and backlight conditions.

4.1 Static cognitive experiment

The role of static cognitive experiments is mainly to confirm the effectiveness and accuracy of cognitive systems. Especially in the estimation of the target distance and size, the RM-GA [13] is used to perform 5,000 consecutive cognitions(33ms/times) to find the optimal match between the model and the object and then calculate the distance and size. The steps of experiment are as follows: (1) Regard a male with hight=176.5mm and width=557.2mm as subject and create a model at 4000mm(initial position), shown in Fig.8; (2) Keeping the above-selected model size unchanged, use the Real-time Multi-step GA (RM-GA) [13] to recognise the 5000 generations at 4000mm under normal indoor light (182lux-188lux); (3) Collect and record experimental data in real-time.

The results are shown in Fig.9 and Tab.1. It can be seen that during the evolution of the 5000-generation RM-GA, the relative error (RE) of the distance and size of the recognition system to the target was less than 0.4 %, and the relative standard deviation (RSD) was less than 0.7 %.

Table 1. Data analysis for static experiment

Class	**Fitness**	**Distance**	${}^{H}\boldsymbol{h}_{M}$	${}^{H}\boldsymbol{w}_{M}$
Unit	-	(mm)	(mm)	(mm)
AVERAGE	0.777	4012.152	1760.622	558.246
STDEVPA	0.013	28.096	11.772	3.733
RSD	1.713%	0.700%	0.669%	0.669%
RE	-	0.304%	0.27%	0.188%

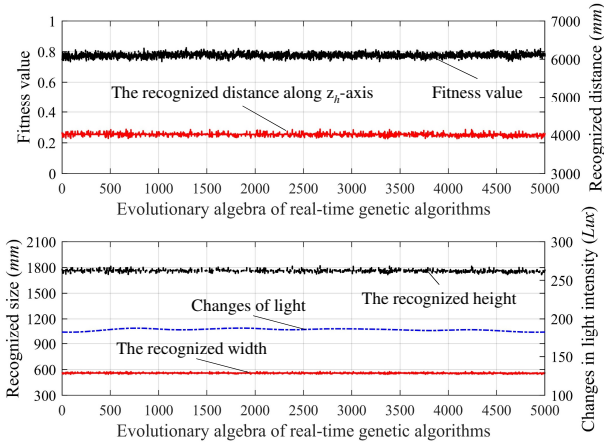

Fig. 9. Result of static experiment

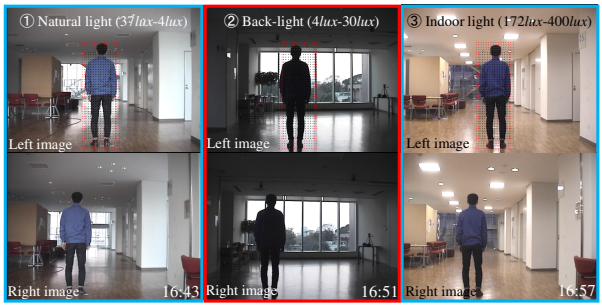

Fig. 10. Tracking experiments under different lighting conditions

4.2 Tracking experiment

Based on the static cognitive experiment, the purpose of the tracking test is to determine whether the cognitive system can still capture and tracking the target when it is moving, especially under different light environments. The experimental steps are as follows: (1) Regard a male with hight=176.5mm and width=512.5mm as subject and create a model at 4000mm(initial position), shown in Fig.8; (2) Keeping the above-selected model size unchanged, and track the movement of human object under normal indoor light (172lux-400lux), natural light (37lux-4lux) and back-light (4lux-30lux) conditions, shown in Fig.10; (3) Collect and record experimental data in real-time.

The tracking results are shown from Fig.11 to Fig.13. Among them, Fig.11 and Fig.12 are the results under natural light conditions. It can be seen from the figure that the size of the object at 4000mm is (613.14mm,1757.68mm). Under natural light, $Fitness_{min1}$= 0.39, and the recognition distance fluctuates between 3740.47 to 4283.2, average value is Dis_{Ave1}=4005.41mm. Under the backlight, $Fitness_{min2}$= 0.65, and the distance ranges between 3673.82 to 4525.39, with an average value of Dis_{Ave2}=4058.10. Under the indoor lighting, $Fitness_{min3}$= 0.76, and the distance fluctuates between 3619.14 to 4490.23, Dis_{Ave3}=4131.56mm.

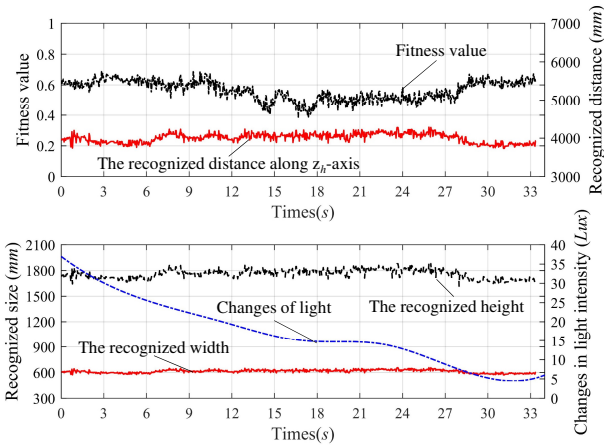

Fig. 11. Results of tracking under natural light condition

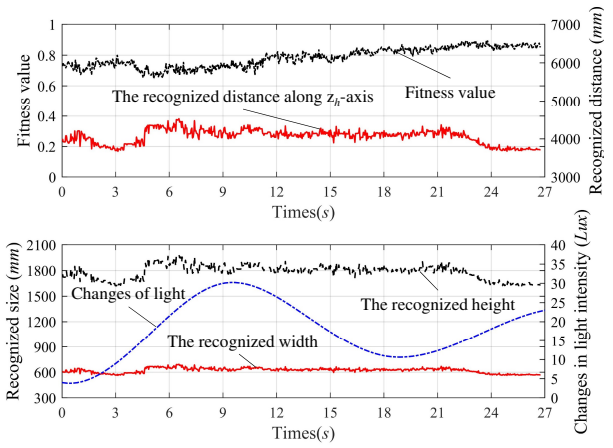

Fig. 12. Results of tracking under back-light light condition

5 CONCLUSION

To achieve better tracking and size recognition in a variety of light intensity environments, especially in changing light or backlight conditions. This paper proposes a method based on a Real-time Multi-step Genetic Algorithm (RM-GA) and newly designed projection-based 3D perception (Pb3DP) method. The experimental results show that the system can detect the distance and size with high accuracy in the static test. In the dynamic tracking experiment, it can still reach excellent tracking accuracy even under changing light(including low light and backlight) conditions. However, it should be mentioned that the system cognitive fitness is lower in a changing(from bright to dark) and low light intensity environment. Authors believe that the problem primarily lies in the initial fixed model parameters, which are difficult to match with the initial model parameters in a changing light environment. Therefore, the next research direction will try to use real-time models for cognitive experiments. In short, this study effectively validates the robustness of the proposed system in a variable light environment and offers a new solution for robotic visual servoing.

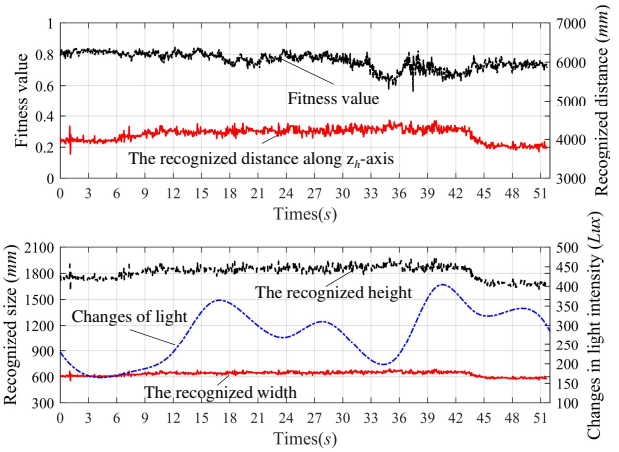

Fig. 13. Results of tracking under indoor light condition

REFERENCES

[1] Heesy, C.P. (2009), Seeing in stereo: The ecology and evolution of primate binocular vision and stereopsis. Evol. Anthropol., 18: 21-35

[2] Andrey V., & Philippe L. (2015), Analysis of CAD Model-based Visual Tracking for Microassembly using a New Block Set for MATLAB/Simulink, International Journal of Optomechatronics, 9:4, 295-309

[3] Jisung P., & Jinwhan K., (2019), Model-referenced pose estimation using monocular vision for autonomous intervention tasks, Autonomous Robots, ISSN 0929-5593

[4] K. Pauwels, & S. Vijayakumar, (2014), Real-time object pose recognition and tracking with an imprecisely calibrated moving RGB-D camera, IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, 2014, pp. 2733-2740

[5] Alhwarin F., & Scholl I., (2014), IR Stereo Kinect: Improving Depth Images by Combining Structured Light with IR Stereo. Pacific Rim International Conference on Artificial Intelligence, Springer, pp. 409-421

[6] D. Morrison et al., (2018), Cartman: The Low-Cost Cartesian Manipulator that Won the Amazon Robotics Challenge, IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, pp. 7757-7764

[7] Andy Z., & Shuran S., (2018), Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching, in 2018 IEEE International Conference on Robotics and Automation (ICRA), pp.3750-3757

[8] Max S., Christian L., (2018), Fast Object Learning and Dual-Arm Coordination for Cluttered Stowing, Picking, and Packing, in 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp. 3347-3354

[9] S. Trinh, & F. Chaumette, (2018), A modular framework for model-based visual tracking using edge, texture and depth features, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 89-96

[10] A. Dakkak, & A. Husain, (2012), Recovering missing depth information from Microsoft's Kinect, Proc. Embedded Vis. Alliance, pp. 1-9

[11] A. Kadambi, & A. Bhandari, (2014), 3d depth cameras in vision: Benefits and limitations of the hardware, Computer Vision and Machine Learning with RGB-D Sensors, Springer, pp. 3-26

[12] A. Leeper, & K. Hsiao, (2014), Using near-field stereo vision for robotic grasping in cluttered environments, Experimental Robotics, Springer, pp. 253-267

[13] Lwin, K. N., Myint, M., Mukada, N., Yamada, D., Matsuno, T., Saitou, K., & Minami, M. (2019), Sea Docking by Dual-eye Pose Estimation with Optimized Genetic Algorithm Parameters. Journal of Intelligent & Robotic Systems, 1-22