Artificial Life and Robotics
https://doi.org/10.1007/s10015-021-00718-7

ORIGINAL ARTICLE

A new method to estimate the pose of an arbitrary 3D object without prerequisite knowledge: projection-based 3D perception

Yejun Kou[1] · Yuichiro Toda[1] · Mamoru Minami[1]

Received: 8 June 2021 / Accepted: 8 November 2021

Abstract
Recognizing a target object and measuring its pose are important functions of robot vision. Most recognition methods require prerequisite information about the target object to conduct the pose estimation, which limits the usability of the robot vision. To overcome this issue, the authors proposed a new approach to estimate an arbitrary target's pose using stereo-vision, which was inspired by the parallactic character in human perception. The authors continued the previous research presented in AROB 2020 and expanded the ability of projection-based 3D perception (Pb3DP). Through tracking the trajectory of the target's motion with a hand–eye robot, it has been confirmed that the Pb3DP method can provide a feasible result in the visual servoing of an unknown target object. In this paper, the authors introduce the methodology of the Pb3DP approach in detail and show the effectiveness of the method through the experimental results of visual servoing in 6 DoF using a stereo-vision hand–eye robot.

Keywords Pose estimation · Arbitrary target object · Projection-based

1 Introduction

Detecting the target and estimating the pose are vital functions in robot vision. Researches and applications in this field have been well studied, including visual servoing, SLAM, and "bin-picking" tasks to pick up industrial parts. All these approaches require a kind of information about the target objects' appearance as a prerequisite. Unlike the methods that need a priori information, this paper considers a situation that the target is an unknown 3D object existing in 3D space. It means that the visual information about the target object, such as the color, shape, and size, are considered unknown before the estimation process starts. Our motivation aims to present a system that can estimate the pose of an arbitrary target object in a real-world setting. Meanwhile, the whole estimation process should be conducted in real time to ensure that the system is functional, while the target is moving. Thereby the pose estimation process of this system should be fast and reliable to make it adaptable to the visual servoing that uses dynamic images provided by stereo-vision.

Some pioneering studies in robot vision, for example, visual servoing, are a technique that uses feedback information extracted from the visual sensor to control the robot's motion. The class of visual servoing methods can be divided into three varieties: position-based visual servoing (PBVS) [1], image-based visual servoing (IBVS) [2], and 2-1/2-D method [3]. Those methods need some pre-defined information about the target object to generate the feature points for conducting visual servoing, which limits the adaptability and usability. In a real-world environment, the object's shapes are irregular and not uniform. Therefore, natural things could not be represented by prepared models. Thereby, a method that could estimate the pose of an arbitrary target without prerequisite information is indispensable to expand the applicability of robot vision into natural outdoor fields.

Recently, some RGB-D camera employed devices, such as Kinect and Realsense, have received much attention. Those devices can estimate the distance to the target through projecting infra-red light, and they do not require prerequisite information about the target object. Therefore, Kinect

This work was presented in part at the 26th International Symposium on Artificial Life and Robotics (Online, January 21–23, 2021).

✉ Yejun Kou
ptlg9dvi@s.okayama-u.ac.jp

[1] Okayama University Graduate School of Nature Science and Technology, 1-1 Tsushima-naka, 3-Chome, Kita-ku, Okayama 700-8530, Japan

Published online: 27 January 2022

is thought to have become the major 3D measuring device in the indoor environment for its low cost and reliable performance. However, the infra-red has a very high absorption rate in water, so the Kinect cannot be used in the underwater environment. Furthermore, the ambient light such as sunlight will lead to a quick saturation of the infra-red sensor, and then, the actual reflection light from the target is hard to be detected. Therefore, the applicable environment of Kinect is limited, and also, it cannot output the target's pose in the scene, but the projection-based 3D perception (Pb3DP) method can do it.

The concept of Pb3DP originated from a photo-model-based method (PMB) [4], which was developed in our laboratory that can estimate a 3D target object's pose with stereo-vision using a 2D photo-model. The performance of tracking a 3D target's pose using the PMB method has been reported in [5] and [6], and the pose tracking experiments towards underwater target have been successfully conducted and presented in [7]. Different from the current technologies [8–10] that are thought to be derived from [1–3], the PMB method provides poses of 3D target objects using a 2D photo-model of the target object. Considering the fact that the pose of a 3D object can be estimated through a 2D model, an extended idea that the 2D photo-model could be substituted by one camera image in one of the stereo-vision cameras happened to appear as a new idea. Since there is no need for a pre-defined photo-model, the application of Pb3DP has expanded the ability that can estimate an arbitrary target's pose. Meanwhile, different from the RGB-D camera [11, 12], the Pb3DP only uses visible light information,"not using the infra-red light," which means that the Pb3DP can be useful in underwater [13]. And the RGB-D methods cannot work outdoors in sunshine [14], but the Pb3DP can be workable.

The previous research presented in AROB 2020 [15] has been limited in the confirmation of pose estimation. The authors focus on expanding the ability of the Pb3DP method into visual servoing by controlling a hand–eye robot to track a target's trajectory through Pb3DP's estimation results. The following contents are organized as follows: Sect. 2 introduces the mechanism of projections and the real-time establishment of a searching model. Section 3 explains the optimization method used in the Pb3DP method, and the visual servoing experiments are explained in Sect. 4.

2 Projection-based 3D perception

2.1 A concept of Pb3DP

The authors proposed a new method named Pb3DP to recognize an unknown target and estimate its pose without any pre-defined information. The searching model of the Pb3DP method can be made from one of the stereo-vision camera images. This approach of establishing a model in one camera image requires no prerequisite information about the target object, it ensured that the Pb3DP method could be applicable to any unknown target object.

This section focuses on introducing the methodology of the Pb3DP method. The concept of Pb3DP is introduced in Fig. 1 in the order from (1) to (5). (1): A real target is depicted with dark color and be projected onto left and right cameras images—those are natural light projections—are indicated by Ⓐ. The natural projection is thought to be completed based on the true pose, $\boldsymbol{\phi}$, but it is unknown. (2): Target object is selected in the left camera scene as a model depicted by Ⓑ. (3): The selected 2D target model is inversely projected into the 3D space with an assumed pose $\hat{\boldsymbol{\phi}}$, shown by Ⓒ and the inverse projection is indicated by Ⓓ. The inversely projected 2D model in the 3D space is depicted with light color, compared with the actual target which is also depicted in the 3D space. (4): The 2D flat model in 3D space is re-projected back to the right camera image with the same assumed pose $\hat{\boldsymbol{\phi}}$, depicted by Ⓔ. (5): If the re-projected model with an assumed pose $\hat{\boldsymbol{\phi}}$ coincides with the real target in the right camera image, then the assumed pose is deemed to represent the real target's pose in 3D space, that is, $\hat{\boldsymbol{\phi}} = \boldsymbol{\phi}$.

The following explanation about the Pb3DP method can be divided into two parts: (1) how the target's pose is estimated in this section and (2) how to apply the Pb3DP method in real time is described in Sect. 3. The first part includes the model generation process and the inverse projection from 2D to 3D of a 2D image model with the assumed pose $\hat{\boldsymbol{\phi}}$ that has been explained in Fig. 1. The second part is to detect

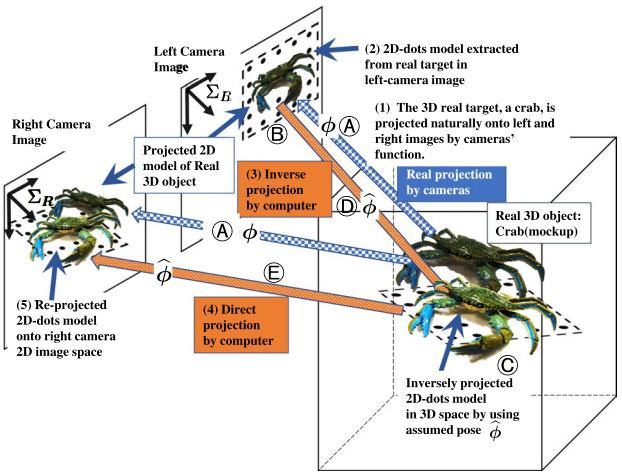


Fig. 1 The schematic of Pb3DP. However, in this figure, since this figure is depicted on the assumption of $\hat{\boldsymbol{\phi}} \neq \boldsymbol{\phi}$, the two images, naturally projected crab by right camera and the inversely projected crab with light color, do not coincide with each other in right camera image. If $\hat{\boldsymbol{\phi}} = \boldsymbol{\phi}$, they would coincide

the pose $\hat{\boldsymbol{\phi}}_i$ that represents true target's pose $\boldsymbol{\phi}$. It is a time-consuming task to find $\boldsymbol{\phi}$, but could be optimized by a real-time multi-step GA approach (RM-GA) [16]. The following contents will introduce the methodology in detail.

2.2 The establishment of a model

The model used in the Pb3DP is a rectangle area in the left camera image that encloses the target object. It consists of a 2-D point cloud; each of the sampling points contains the color information of the image at the location of the point, which is used to evaluate the recognition result. The model has two portions: the inner area that mainly presents the target object, and the outer area, which presents the background around the target object. The model can be made instantaneously using the target's image shown on the camera. It means as long as the target object can be seen in the camera's image; the Pb3DP method can make the model of the target without any prerequisite about the target object. This way of model's establishment ensured that the Pb3DP method could recognize an arbitrary target object.

The components of the model are the same as the previous research of the authors. The detail of the inner area and outer area can be referred to [17].

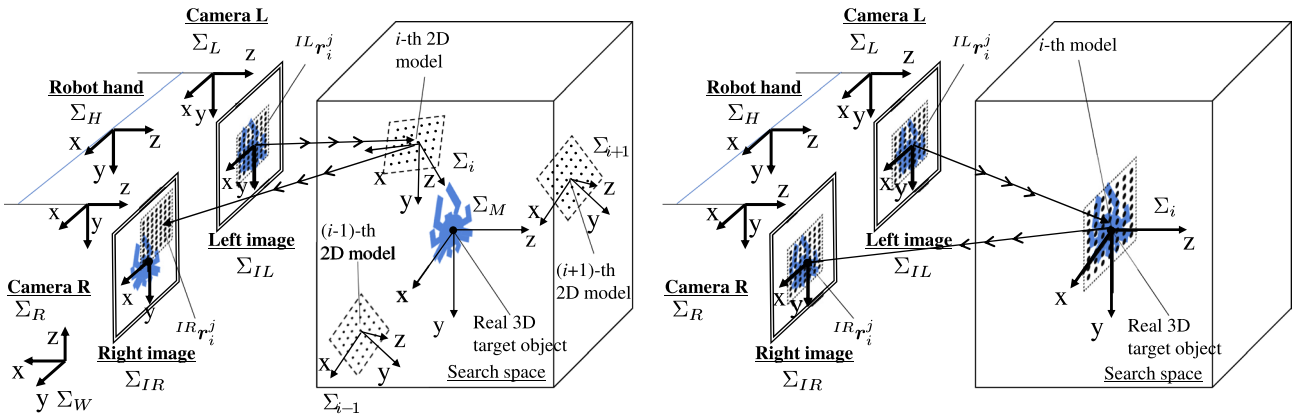
2.3 Projections and model selection

Figure 2a shows a similar situation as Fig. 1, but it represents how the true pose $\boldsymbol{\phi}$ could be found in genetic algorithm (GA) process: a target's image in left camera is selected as 2D image model, and the 2D image model is inversely projected into 3D space with assumed pose $\hat{\boldsymbol{\phi}}_{i-1}$, $\hat{\boldsymbol{\phi}}_i$, and $\hat{\boldsymbol{\phi}}_{i+1}$; the coordinate systems of those inversely projected model are defined as Σ_{i-1}, Σ_i, Σ_{i+1}. The Fig. 2b shows the situation that $\hat{\boldsymbol{\phi}}_i = \boldsymbol{\phi}$ through the convergence conducted by RM-GA. Let us define the coordinate systems and symbols that used in Pb3DP method. The coordinate systems can be referred to Fig. 2:

- $\Sigma_{L,R}$: the coordinate systems of left and right cameras.
- $\Sigma_{IL,IR}$: the coordinate systems of left and right camera images.
- Σ_H: the coordinate system of robot's hand.
- Σ_M: the coordinate system of model whose pose represents the target object's pose.
- $\Sigma_{i-1}, \Sigma_i, \Sigma_{i+1}$: the coordinate systems of the "i-1"th, "i"th, and "i+1"th model that inversely projected into 3D space with assumed pose $\hat{\boldsymbol{\phi}}_{i-1}$, $\hat{\boldsymbol{\phi}}_i$, $\hat{\boldsymbol{\phi}}_{i+1}$.
- ${}^{IL}\boldsymbol{r}_i^j, {}^{IR}\boldsymbol{r}_i^j$: the position vector of j-th point on i-th model in the left and right camera image coordinate.
- Σ_W: the world coordinate system.

As shown in Fig. 3, there is a 2D model in 3D space with the coordinate Σ_M attached, consisted of sampling points that lie on the plane. Because the model is a 2D flat, the z-axis coordinate of an arbitrary jth sampling point on the model is fixed to be 0, i.e., ${}^M z_j = 0$. The position vectors in the world coordinate system Σ_W of an arbitrary jth point on a 2D model placed in space are set and defined as follows:

- ${}^W\boldsymbol{r}_{Mj}$: 3D position vector in Σ_W of a jth point on a 2D model defined by Σ_M.
- ${}^M\boldsymbol{r}_j$: 2D position vector on x-y plane in Σ_M of a jth point on a 2D model whose x-y plane coincides with the 2D model plane, where ${}^M\boldsymbol{r}_j$ is a constant vector, since Σ_M is attached to the model.



(a) Distributed 2D models in 3D space before converging to real 3D target object

(b) Converged 2D models onto 3D real target by RM-GA

Fig. 2 The coordinate systems; symbols used in Pb3DP method

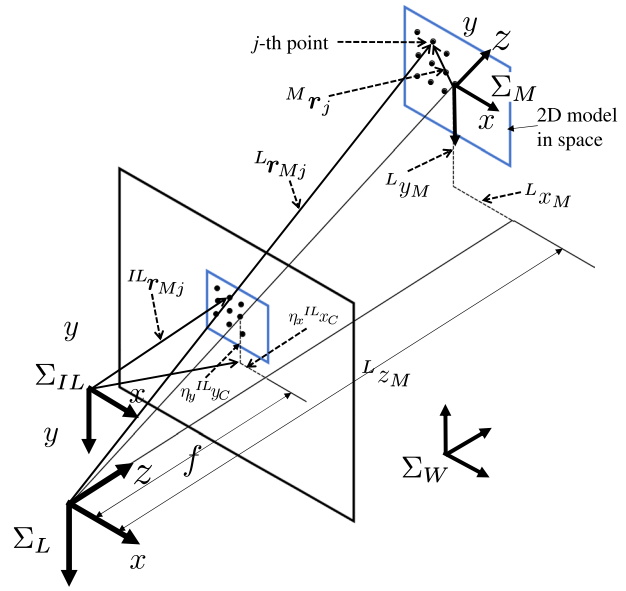


Fig. 3 Projection schematic diagram: The jth point ${}^{M}\boldsymbol{r}_j = [{}^{M}x_j, {}^{M}y_j, {}^{M}z_j]^{\mathbf{T}}$ on a model shown by Σ_M in space is converted into the point represented by Σ_L as ${}^{L}\boldsymbol{r}_j = [{}^{L}x_j, {}^{L}y_j, {}^{L}z_j]^{\mathbf{T}}$, and the point ${}^{L}\boldsymbol{r}_j$ is projected to ${}^{IL}\boldsymbol{r}_j = [{}^{IL}x_j, {}^{IL}y_j]^{\mathbf{T}}$ on the left camera image defined by Σ_{IL}

- ${}^{L}\boldsymbol{r}_{Mj}$: 3D position vector in Σ_L of a jth point on a 2D model in space in the left camera coordinates system Σ_L, as shown in Fig. 3.
- ${}^{IL}\boldsymbol{r}_{Mj}$: 2D position vector of an jth point on a model in left image coordinate system Σ_{IL}.

Given the homogeneous matrix connecting Σ_M to Σ_L as ${}^{L}\boldsymbol{T}_M$, the relation between ${}^{L}\boldsymbol{r}_{Mj} = [{}^{L}x_{Mj}, {}^{L}y_{Mj}, {}^{L}z_{Mj}, 1]^{\mathbf{T}}$ and ${}^{M}\boldsymbol{r}_j = \left[{}^{M}x_j, {}^{M}y_j, {}^{M}z_j, 1\right]^{\mathbf{T}}$ is represented by

$$ {}^{L}\boldsymbol{r}_{Mj} = {}^{L}\boldsymbol{T}_M \, {}^{M}\boldsymbol{r}_j. \tag{1} $$

The detailed ${}^{L}\boldsymbol{T}_M$ is given by

$$ \begin{aligned} {}^{L}\boldsymbol{T}_M({}^{L}\boldsymbol{r}_M, {}^{L}\boldsymbol{\theta}_M) = &\begin{bmatrix} 1 & 0 & 0 & {}^{L}x_M \\ 0 & 1 & 0 & {}^{L}y_M \\ 0 & 0 & 1 & {}^{L}z_M \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos{}^{L}\theta_z & -sin{}^{L}\theta_z & 0 & 0 \\ sin{}^{L}\theta_z & cos{}^{L}\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &\begin{bmatrix} cos{}^{L}\theta_y & 0 & sin{}^{L}\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ -sin{}^{L}\theta_y & 0 & cos{}^{L}\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos{}^{L}\theta_x & -sin{}^{L}\theta_x & 0 \\ 0 & sin{}^{L}\theta_x & cos{}^{L}\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \end{aligned} \tag{2} $$

where ${}^{L}\boldsymbol{r}_M = [{}^{L}x_M, {}^{L}y_M, {}^{L}z_M]^{\mathrm{T}}$, ${}^{L}\boldsymbol{\theta}_M = [{}^{L}\theta_x, {}^{L}\theta_y, {}^{L}\theta_z]^{\mathrm{T}}$.

The jth point ${}^{L}\boldsymbol{r}_{Mj}$ on a model defined by Σ_L in space is projected to ${}^{IL}\boldsymbol{r}_{Mj} = [{}^{IL}x_{Mj}, {}^{IL}y_{Mj}]^{\mathrm{T}}$ on the left camera image defined by Σ_{IL} as follows, using $\boldsymbol{\phi} = [{}^{L}\boldsymbol{r}_M^{\mathrm{T}}, {}^{L}\boldsymbol{\theta}_M^{\mathrm{T}}]^{\mathrm{T}}$:

$$ \begin{aligned} {}^{IL}\boldsymbol{r}_{Mj} &= \boldsymbol{P}({}^{L}z_{Mj}) {}^{L}\boldsymbol{r}_{Mj} \\ &= \boldsymbol{P}(\boldsymbol{\phi}) {}^{L}\boldsymbol{T}_M(\boldsymbol{\phi}) {}^{M}\boldsymbol{r}_j. \end{aligned} \tag{3} $$

The projective transformation matrix $\boldsymbol{P}({}^{L}z_{Mj})$ is given as

$$ \boldsymbol{P}({}^{L}z_{Mj}) = \frac{1}{{}^{L}z_{Mj}} \begin{bmatrix} f/\eta_x & 0 & 0 & 0 \\ 0 & f/\eta_y & 0 & 0 \end{bmatrix}, \tag{4} $$

where

- ${}^{L}z_{Mj}$: z-axis position of the j-th point in Σ_L on the model Σ_M,
- f: focal length,
- η_x, η_y: coefficients [mm/pixel] in x-axis and y-axis of image frame.

The projection of right camera can be discussed in the same manner.

2.4 Inverse projection from left camera image to 3D space and re-projection to right camera image

For preparation of inverse projection of ${}^{IL}\boldsymbol{r}_{Mj}$ to 3D space, the pseudo-inverse projection matrix $\boldsymbol{P}^{+}({}^{L}z_{Mj})$ of $\boldsymbol{P}({}^{L}z_{Mj})$ defined by Eq. (4) is needed

$$ \boldsymbol{P}^{+}({}^{L}z_{Mj}) = {}^{L}z_{Mj} \begin{bmatrix} \eta_x/f & 0 & 0 & 0 \\ 0 & \eta_y/f & 0 & 0 \end{bmatrix}^{\mathrm{T}}. \tag{5} $$

Equation (3) can be modified into

$$ {}^{L}\boldsymbol{T}_M(\boldsymbol{\phi}) {}^{M}\boldsymbol{r}_j = \boldsymbol{P}^{+}(\boldsymbol{\phi}) {}^{IL}\boldsymbol{r}_{Mi} + (I_4 - \boldsymbol{P}^{+}\boldsymbol{P})\boldsymbol{l}. \tag{6} $$

If the position vector of j-th point in 2D model on Σ_M is selected to be at the origin of Σ_M, then ${}^{M}\boldsymbol{r}_j = [0, 0, 0, 1]^{\mathrm{T}}$. Providing that the corresponding point to the origin of Σ_M in the left camera image coordinate system Σ_{IL} is $[{}^{IL}x_C, {}^{IL}y_C]^{\mathrm{T}}$, and that an arbitrary vector $\boldsymbol{l}$ is given by $\boldsymbol{l} = [l_1, l_2, l_3, 1]^{\mathrm{T}}$, then Eq. (6) leads to

$$ \begin{bmatrix} {}^{L}x_M \\ {}^{L}y_M \\ {}^{L}z_M \\ 1 \end{bmatrix} = \begin{bmatrix} \eta_x {}^{IL}x_C {}^{L}z_M/f \\ \eta_y {}^{IL}y_C {}^{L}z_M/f \\ l_3 \\ 1 \end{bmatrix}. \tag{7} $$

The above relation indicates the origin position ${}^{L}x_M, {}^{L}y_M$ of Σ_M in Σ_L could be determined dependently based on ${}^{L}z_M$ and ${}^{IL}x_C, {}^{IL}y_C$. When we want to calculate a flat model in 3D space from flat image in left camera flame Σ_{IL} by inverse projection, Eq. (7) means that the origin position ${}^{L}x_M, {}^{L}y_M$ in Σ_L could be determined linearly using the center position of model ${}^{IL}x_C, {}^{IL}y_C$ in Σ_{IL} and arbitrary position ${}^{L}z_M$. Figure 3 represents that the relation given by Eq. (7) could be understood from a view point of geometry as linear projection,

meaning that the point ${}^{IL}x_C$, ${}^{IL}y_C$ and ${}^{L}x_M$, ${}^{L}y_M$ are connected linearly by the distance ${}^{L}z_M$.

The assumptions that the proposed Pb3DP is based on are given: (1) A target object should be projected onto both left and right cameras images. (2) The point cloud defined in the left camera image includes the target, as shown in Fig. 1Ⓑ. Therefore, the rectangular point cloud depicted in Σ_{IL} in Fig. 3 always includes the projected target. (3) To simplify the inverse projection, the rotation around the z-axis, i.e., ${}^{L}\theta_z$ in Eq. (2) be assumed to be zero without loosing generality, since the natural light projection relation from points in Σ_M to points in Σ_{IL} is always straight without rotation around the camera depth direction. Then, Eq. (2) could be simplified into

$$
\begin{aligned}
{}^{L}\boldsymbol{T}_M(\boldsymbol{\psi}) \;=\; & \begin{bmatrix} 1 & 0 & 0 & \eta_x {}^{IL}x_C {}^{L}z_M/f \\ 0 & 1 & 0 & \eta_y {}^{IL}y_C {}^{L}z_M/f \\ 0 & 0 & 1 & {}^{L}z_M \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos^{L}\theta_x & -sin^{L}\theta_x & 0 \\ 0 & sin^{L}\theta_x & cos^{L}\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
& \begin{bmatrix} cos^{L}\theta_y & 0 & sin^{L}\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ -sin^{L}\theta_y & 0 & cos^{L}\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},
\end{aligned} \tag{8}
$$

where $\boldsymbol{\psi}$ is defined as

$$
\boldsymbol{\psi} = [{}^{L}z_M, {}^{L}\theta_x, {}^{L}\theta_y]. \tag{9}
$$

On the other hand, through comparing the current image with the initial status when the model of the target object was made, the Pb3DP method can recognize the target's orientation about ${}^{L}\theta_z$, which is a relative value that represents how many degrees the target rotated around the z_M-axis after the estimation started.

The three components ${}^{L}z_M$, ${}^{L}\theta_x$, ${}^{L}\theta_y$ of $\boldsymbol{\psi}$ are independent variables for inversely projecting the 2D model in left camera 2D image into space. Providing a set of variables in the variety of $\boldsymbol{\psi}$ be chosen and fixed, shall we describe the fixed variable as $\hat{\boldsymbol{\psi}}$ and the true pose as $\boldsymbol{\psi}$ in later contents, then the inversely projected flat model position ${}^{W}\boldsymbol{r}_{Mj}(\hat{\boldsymbol{\psi}})$ that is determined dependently by $\hat{\boldsymbol{\psi}}$ in Σ_W is derived from Eq. (6) as

$$
\begin{aligned}
{}^{W}\boldsymbol{r}_{Mj}(\hat{\boldsymbol{\psi}}) \;=\; & {}^{W}\boldsymbol{T}_M(\hat{\boldsymbol{\psi}}) {}^{L}\boldsymbol{T}_M^{-1}(\hat{\boldsymbol{\psi}}) \Big[\boldsymbol{P}^{+}(\hat{\boldsymbol{\psi}}) {}^{IL}\boldsymbol{r}_{Mj} + (\mathrm{I}_4 \\
& - \boldsymbol{P}^{+}(\hat{\boldsymbol{\psi}}) \boldsymbol{P}(\hat{\boldsymbol{\psi}})) \boldsymbol{l} \Big].
\end{aligned} \tag{10}
$$

Then, the image projected to right camera plane of flat target model, ${}^{IR}\boldsymbol{r}_{Mj}$, is also calculated using assumed $\hat{\boldsymbol{\psi}}$ as

$$
{}^{IR}\boldsymbol{r}_{Mj}(\hat{\boldsymbol{\psi}}) = \boldsymbol{P}(\hat{\boldsymbol{\psi}}) {}^{R}\boldsymbol{T}_W(\hat{\boldsymbol{\psi}}) {}^{W}\boldsymbol{r}_{Mj}(\hat{\boldsymbol{\psi}}). \tag{11}
$$

2.5 Problem conversion from pose estimation to optimization

The purpose of this subsection is to show how to convert the estimation problem of a 3D target's true pose $\boldsymbol{\psi}$ to the optimization problem. Please assume that $\hat{\boldsymbol{\psi}}$ means estimated value of $\boldsymbol{\psi}$. If a scalar function $F(\hat{\boldsymbol{\psi}})$ should satisfy that the distribution of $F(\hat{\boldsymbol{\psi}})$ has a single maximum peak F_{max} at true pose $\boldsymbol{\psi}$, and that also satisfy $F(\hat{\boldsymbol{\psi}}) = F_{max}$, then $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}$. This could be rewritten as single peak assumption, that is,

$F(\hat{\boldsymbol{\psi}}) = F_{max}$ if and only if $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi} \in L$,

where L means parameter space of $\hat{\boldsymbol{\psi}}$, and then, the problem to estimate the true pose $\boldsymbol{\psi}$ can be converted to an another problem as,

Find $\hat{\boldsymbol{\psi}}$ to maximize $F(\hat{\boldsymbol{\psi}})$ subject to $\hat{\boldsymbol{\psi}} \in L$.

It means that the estimation of true pose can be completed by optimizing $F(\hat{\boldsymbol{\psi}})$ in parameters space of $\hat{\boldsymbol{\psi}}$. Then, how to constitute a scalar function $F(\hat{\boldsymbol{\psi}})$ satisfying the single peak assumption above appears to be a next problem.

3 Real-time multi-step GA

3.1 Evaluation method

In the proposed Pb3DP method, the models with assumed pose are utilized to infer the true pose of the target object. A coincidence degree between the projected model and the target's image in the right camera captured by dual-eye cameras is evaluated by a scalar function used as fitness function in optimization GA process [18]. And the fitness value F_R calculated from right camera image is used as a numerical value to represent the coincidence degree. A higher fitness value means a higher coincidence degree between the assumed pose $\hat{\boldsymbol{\psi}}_i$ and true pose $\boldsymbol{\psi}$. Therefore, the fitness value can convert the problem of finding the true pose of the target object into finding the maximum value of fitness.

The number of sampling points in the inner area is N_{in} and the outer area is N_{out}. As shown in Fig. 2, the coordinate of jth point in ith model projected into right camera image is ${}^{IR}\boldsymbol{r}_i^j$, and the evaluation value of each point in inner portion of the model (${}^{IR}\boldsymbol{r}_i^j \in S_{R,in}(\hat{\boldsymbol{\psi}})$) is $P_{R,in}({}^{IR}\boldsymbol{r}_i^j)$ calculated by Eq. (12). One of outer portion (${}^{IR}\boldsymbol{r}_i^j \in S_{R,out}(\hat{\boldsymbol{\psi}})$) is $P_{R,out}({}^{IR}\boldsymbol{r}_i^j)$ calculated by Eq. (13)

$$
P_{R,in}({}^{IR}\boldsymbol{r}_i^j) = \begin{cases} K_{R,in}, (|H_M({}^{IR}\boldsymbol{r}_i^j) - H_I({}^{IR}\boldsymbol{r}_i^j)| \le 20) \\ Q_{R,in}, (|H_M({}^{IR}\boldsymbol{r}_i^j) - H_I({}^{IR}\boldsymbol{r}_i^j)| > 20) \end{cases} \tag{12}
$$

$$
P_{R,out}({}^{IR}\boldsymbol{r}_i^j) = \begin{cases} K_{R,out}, (|H_M({}^{IR}\boldsymbol{r}_i^j) - H_I({}^{IR}\boldsymbol{r}_i^j)| \le 20) \\ Q_{R,out}, (|H_M({}^{IR}\boldsymbol{r}_i^j) - H_I({}^{IR}\boldsymbol{r}_i^j)| > 20), \end{cases} \tag{13}
$$

where

- $H_M({}^{IR}\boldsymbol{r}_i^j)$: the hue value of the projected model in right camera image at the point ${}^{IR}\boldsymbol{r}_i^j$ (jth point in ith model, lying in $S_{R,in}$).
- $H_I({}^{IR}\boldsymbol{r}_i^j)$: the hue value of right camera image at the point ${}^{IR}\boldsymbol{r}_i^j$.
- $K_{R,in}, K_{R,out}$: the evaluation value in inner area and outer area when the difference between $H_M({}^{IR}\boldsymbol{r}_i^j)$ and $H_I({}^{IR}\boldsymbol{r}_i^j)$ is less than 20 or equal.
- $Q_{R,in}, Q_{R,out}$: the evaluation value in inner area and outer area when the difference between $H_M({}^{IR}\boldsymbol{r}_i^j)$ and $H_I({}^{IR}\boldsymbol{r}_i^j)$ is larger than 20.
- N: the total number of sampling points in a model.

The fitness function can be given by the following equation:

$$F_R(\boldsymbol{\phi}) = \left\{ \sum_{{}^{IR}\boldsymbol{r}_i^j \in S_{R,in}(\hat{\boldsymbol{\psi}})} P_{R,in}({}^{IR}\boldsymbol{r}_i^j) + \sum_{{}^{IR}\boldsymbol{r}_i^j \in S_{R,out}(\hat{\boldsymbol{\psi}})} P_{R,out}({}^{IR}\boldsymbol{r}_i^j) \right\} \Big/ N. \tag{14}$$

If the projected 2D model entirely coincides with the captured target object in the right images, the fitness value calculated by Eq. (14) is designed to have a maximum value. Therefore, the fitness value distribution for all models will be shaped with a peak that represented the real pose of the target object as Fig. 4. In Fig. 4, a model of crab is employed as the target object, and it was set as the following pose:

$${}^{H}\boldsymbol{\phi}_M = [0, 0, 500, 0, 0, 0]^{\mathrm{T}}. \tag{15}$$

Figure 4a shows the image of the selected target object, and Fig. 4b shows the fitness distribution in the "$z - {}^{L}\theta_x$" plane. Here, the maximum value of fitness indicates at the point (512[mm], 5[°]), which is near to the true pose of the target object (500[mm], 0[°]). And, Fig. 4c shows the fitness distribution in the "${}^{L}\theta_x - {}^{L}\theta_y$" plane, the maximum fitness exists at the point (-2[°], -3[°]), and the result is near the true orientation (0[°], 0[°]) of the target object. Those results of the distribution of fitness value that show single dominant peak are meaning the fitness function, Eq.

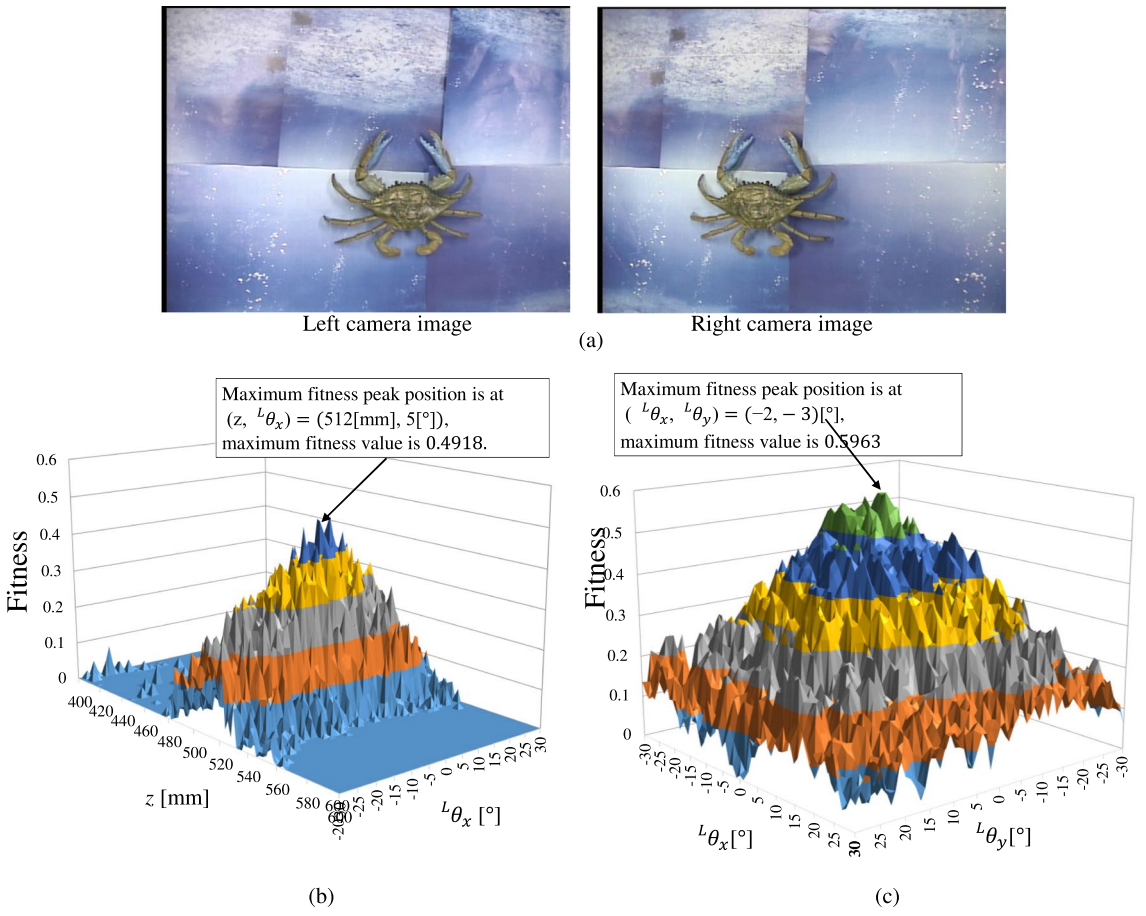


Fig. 4 Fitness distribution of the mock-up of a crab. **a** Left and right camera images; **b** fitness distribution in the z-${}^{L}\theta_x$ plane; **c** fitness distribution in the ${}^{L}\theta_x$ –${}^{L}\theta_y$ plane. In each subfigure of (**b**), (**c**), the maximum fitness value and corresponding coordinate to give the maximum value are shown in text boxes, whose values are almost same to the pose given by Eq. (15), representing 3D crab target's pose. The experimental setup is shown in Fig. 7. The coordinates are based on Σ_H

(14) can convert the target recognition and pose estimation problem into an optimization problem to find the maximum peak in fitness distribution. Since the position of the peak is located close to the position of the target object's true pose, it can show that the proposed method can estimate the 3D target pose through GA using the defined fitness equation [Eq. (14)]. This fitness function can be said as an extension of the work in [19] in which different models, including a rectangular shape surface-strips model, were evaluated using images from a single camera.

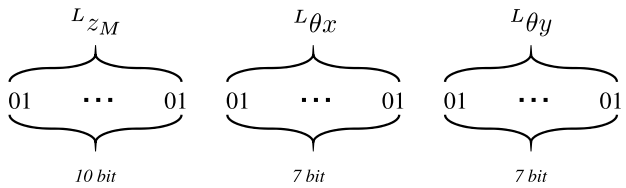


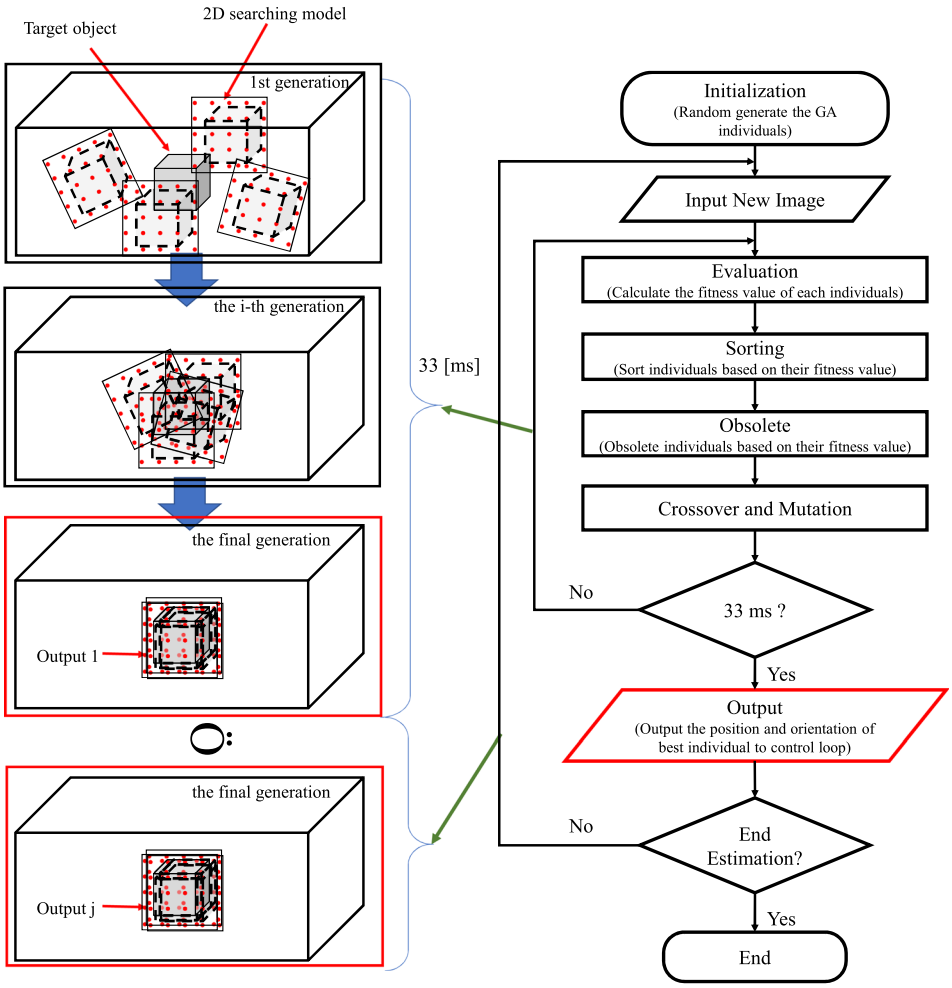
Fig. 5 Gene information

3.2 Real-time multi-step GA

Calculating all possible poses of the target object for making a fitness distribution like Fig. 4 is time-consuming, contradicting real-time pose estimation. Therefore, in Pb3DP, we employed Real-time Multi-step GA (RM-GA) to satisfy the real-time recognition in the image frame fresh rate of 30 [fps]. The reason why we choose RM-GA is that the selected poses by RM-GA are just calculated for the GA's optimization [20].

Different from the previous works that calculate six variables to infer an assumed pose [21], in the proposed Pb3DP method, three variables can determine the pose of a 2D model in 3D space according to Eq. (9). It means that only three variables are needed to compute simultaneously, which improved the time–response performance of RM-GA by reducing the calculating time. In the RM-GA used in Pb3DP, each chromosome includes 24 bits for searching three parameters: ten for the position and fourteen for orientation, as shown in Fig. 5. Figure 6 shows the flowchart of the Real-time Multi-step GA, and the recognition process in 3D space is presented on the left. Here, a 2D searching

Fig. 6 Flowchart of the RM-GA



model in 3D space represents a GA individual. The GA's operation is conducted in the sequence as evaluation, sorting, obsolete, crossover, and mutation. Several 2D searching models representing different relative poses converge to the target object through the GA evolution process. The 2D searching model that represents the true pose with the highest fitness value calculated by Eq. (14) is output for every 33 [ms]. Then, the model with highest fitness is directly transferred to the next generation as the initial model to evaluate the next new images.

4 Experiments

The experimental result in this paper presents the visual servoing performance by the Pb3DP method. During the experiment, the target object changed its position and orientation with time. The hand–eye robot recognizes the target object's pose in real time and keeps an assigned pose relationship to control the robot's hand to track the trajectory of the target object.

4.1 Experimental environment

The experimental environment can be referred to Fig. 7. Two manipulators are employed in this experiment; both are the PA-10 robot arm manufactured by Mitsubishi Heavy Industries. Two cameras are mounted on the VS-robot's end-effector and connected to a host computer (CPU: Intel i7-3770, 3.40[GHz]). The layout of these two cameras forms a binocular vision configuration. The resolution of dynamic images is 640 × 480 [pixel], and the frame frequency of stereo cameras is set as 30[fps]. TC-robot controls the trajectory of the target object, and the VS-robot conducts the visual servoing towards the target object through the pose estimation.

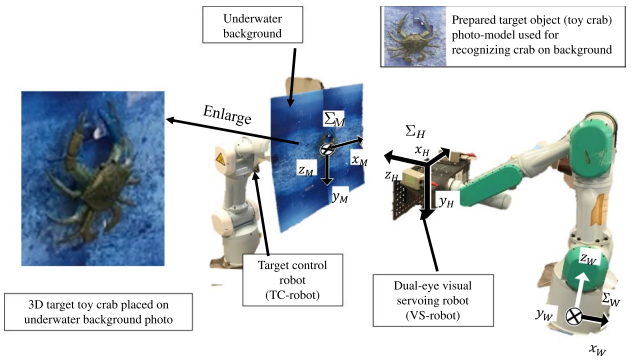


Fig. 7 Experimental layout. The motion of the target animal, crab, is given by TC-robot, and the VS-robot moves to keep desired relative pose of the VS-robot against the crab attached on a panel with sea bottom backdrop whose motion is given by TC-robot. World coordinate system Σ_W, hand coordinate system Σ_H, and target coordinate system Σ_M are depicted in the figure

Because, this experiment is examining the visual servoing performance of Pb3DP. Therefore, the experimental results are represented in ${}^{W}\boldsymbol{T}_H({}^{W}\hat{\boldsymbol{\psi}}_H(t))$, ${}^{W}\hat{\boldsymbol{\psi}}_H = [{}^{W}\hat{x}_H(t), {}^{W}\hat{y}_H(t), {}^{W}\hat{z}_H(t), {}^{W}\hat{\varepsilon}_{1H}(t), {}^{W}\hat{\varepsilon}_{2H}(t), {}^{W}\hat{\varepsilon}_{3H}(t)]$

to show if the trajectory of the robot's hand represented by Σ_H can match with the target's given by Σ_M that moves with motion of TC-robot in Fig. 7, or not. During the experiment, the target object moves along with the desired relative pose between Σ_H and Σ_M represented by Σ_W is given by a defined trajectory, which is shown in Fig. 8a–f as black solid lines. To the position change, the target object moves along the axis of $k_M, (k = x, y, z)$ separately and the position changes can be referred to step(1) ~ step (6) in Fig. 8. Meanwhile, the change of target's orientation happened after the position changes were completed. The target rotated itself around the axis of $k_M, (k = x, y, z)$, and those rotation can be referred to step (7) ~ step (12) in Fig. 8. The movement range of position change is 100[mm] in each direction, and the degree of orientation change is set as 20[deg], which can be transformed into quaternion as "0.173." The time of each step is marked as (1)–(12) in the top of Fig. 8. The target object changed its position and orientation with time.

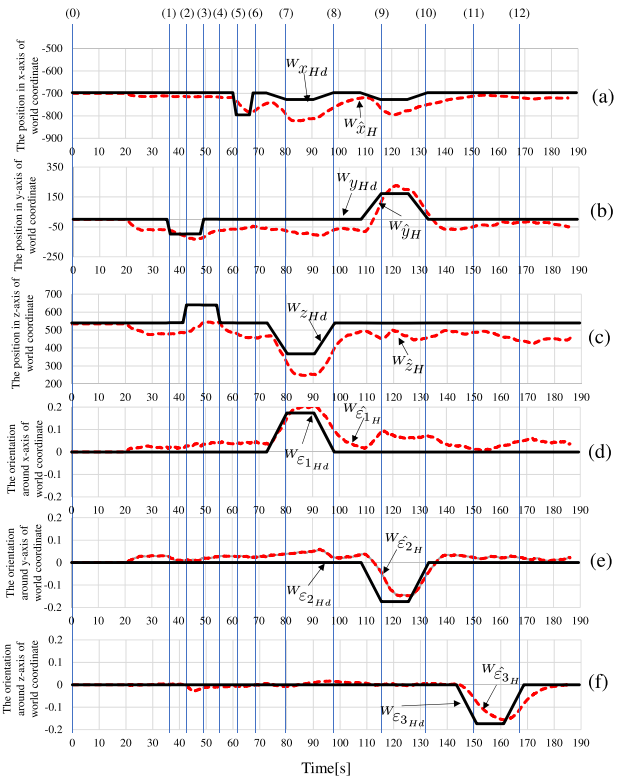


Fig. 8 The experiment result. The trajectory of target object's movement is shown as black lines in (**a**–**f**), and the movement of robot's hand is shown as red lines

Meanwhile, a pose relationship between the robot hand and the target object is kept as Eq. (16)

$$^{Hd}\boldsymbol{T}_M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 500[\text{mm}] \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{16}$$

Equation (16) also represents the initial pose of TC-robot and VS-robot in step (0) shown in Fig. 8, and corresponds to Eq. (15).

4.2 Experimental result

The trajectories of robot's hand are shown as $^{W}\hat{k}_H, (k = x, y, z)$ in Fig. 8, and the trajectories of target object's movements are shown as black solid lines as $^{W}k_H, (k = x, y, z)$. As a result, the Pb3DP method can recognize the pose of the target object in real time and guide the robot's hand to track the movement of the target object in time keep $^{H}\boldsymbol{T}_M$ to same as Eq. (16). The position visual servoing results are shown as Fig. 8a–c, and the orientation visual servoing results are shown in Fig. 8d–f. To the position tracking result, the movement of the robot's hand can follow the trajectory of the target object, but there is some error between the desired hand position $^{W}k_H(k = x, y, z)$ and the real hand position $^{W}\hat{k}_H(k = x, y, z)$. The reason is considered as time delay of pose estimation and mechanical motion time delay caused by motor's time delay.

Concerning the orientation tracking results, as shown in Fig. 8, the robot hand can track the trajectory of the target object with some time delays and some errors during the experiment, because the recognition of orientation recognition result is always changing, while the pose recognition is being conducted by RM-GA, which is providing a new estimation result every 33[ms]. According to the experimental results, we can see that the PA-10 can recognize and track the target object in real time by employing the proposed Pb3DP method, while the 6 degrees of pose elements are changed in turns that showed that 6 degrees of pose could be measured in real time and VS-robot could be controllable to the pose changing of the TC-robot shown in Fig. 7.

5 Conclusion

In this paper, the Pb3DP stereo-vision system is introduced. It was verified that the proposed method can make the robot's hand–eye track the 3D target object without a pre-defined model. It means that the 3D pose of an arbitrary target can be estimated in real time by its 2D model. Meanwhile, the RM-GA ensured the feasibility and robustness of recognition towards target objects and also the visual servoing robot could be controlled to the devised relative pose shown by the moving target.

References

1. Allen PK, Timcenko A, Yoshimi B, Michelman P (1993) Automated tracking and grasping of a moving object with a robotic hand-eye system. IEEE Trans Robot Autom 9(2):152–165
2. Chaumette F (1998) Potential problems of stability and convergence in image-based and position-based visual servoing. The confluence of vision and control. Springer, London, pp 66–78
3. Chaumette F, Malis E (2000) 2 1/2 D visual servoing: a possible solution to improve image-based and position-based visual servoings. Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065) 1: 630-635
4. Tian H, Kou Y, Phyu K W, Funakobo R. and Minami M (2018) Visual Servoing to Arbitrary Target by Using Photo-Model Definition. In The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec): pp 2A1-M17
5. Tian H, Kou Y, Minami M (2019) Visual servoing to arbitrary target with photo-model-based recognition method. 24th International Symposium on Artificial Life and Robotics: pp 950-955
6. Tian H, Kou Y, Li X, Minami M (2020) Real-time pose tracking of 3D targets by photo-model-based stereo-vision. J Adv Mech Design Syst Manuf 14(4):JAMDSM0057
7. Tian H, Kou Y, Kawakami T, Takahashi R, Minami M (2019) Photo-model-based stereo-vision 3D perception for marine creatures catching by ROV. Oceans 2019:1–6
8. Barbosa GB, Da Silva E C, Leite AC (2021) Robust Image-based Visual Servoing for Autonomous Row Crop Following with Wheeled Mobile Robots. In 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE): pp 1047-1053
9. Sharma RS, Shukla S, Beher L, Subramanian VK (2020) Position-based visual servoing of a mobile robot with an automatic extrinsic calibration scheme. Robotica 38(5):831–844
10. He Z, Wu C, Zhang S, Zhao X (2018) Moment-based 2.5-D visual servoing for textureless planar part grasping. IEEE Transactions on Industrial Electronics 66(10): pp 7821-7830
11. Zeng X, Gao Y, Hou S, Peng S (2015) Real-time multi-scale tracking via online RGB-D multiple instance learning. J Softw 10(11):1235–1244
12. Susperregi L, Martínez-Otzeta JM, Ansuategui A, Ibarguren A, Sierra B (2013) RGB-D, laser and thermal sensor fusion for people following in a mobile robot. Int J Adv Robot Syst 10(6):271
13. Kawakami T, Takahashi R, Tian H, Kou Y, Minami M (2020) Real-time Spatial Recognition by Underwater Stereo Vision. 25th International Symposium on Artificial Life and Robotics: pp 837-842
14. Gupta M, Yin Q, Nayar SK (2013) Structured light in sunlight. In Proceedings of the IEEE International Conference on Computer Vision: pp 545-552
15. Yejun K, Hongzhi T, Mamoru M (2020) A Realtime 3D Pose Estimation Method towards Arbitrary Target with Stereo Vision. 25th International Symposium on Artificial Life and Robotics: 831-836
16. Lwin KN, 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. J Intell Robot Syst 96(2):245–266

17. Tian H, Kou Y, Li X, Minami M (2020) Real-time pose tracking of 3D targets by photo-model-based stereo-vision. J Adv Mech Design Sys Manuf 14(4):JAMDSM0057
18. Minami M, Zhu J, Miura M (2003) Real-time Evolutionary Recognition of Human with Adaptation to Environmental Condition, The Second International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS), Proceeding CD-ROM PS010103
19. Song W, Minami M, Mae Y, Aoyagi S (2007) On-line evolutionary head pose measurement by feedforward stereo model matching. In Proceedings 2007 IEEE International Conference on Robotics and Automation: pp 4394-4400
20. Myint M, Lwin KN, Mukada N, Yamada D, Matsuno T, Toda Y, Minami M (2019) Experimental verification of turbidity tolerance of stereo-vision-based 3D pose estimation system. J Marine Sci Technol 24(3):756–779
21. Kou Y, Tian H, Minami M, Matsuno T (2018) Improved eye-vergence visual servoing system in longitudinal direction with RM-GA. Artif Life Robot 23(1):131–139