IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 13, 2018, accepted October 30, 2018, date of publication November 15, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2880413

Design and Vision Based Autonomous Capture of Sea Organism With Absorptive Type Remotely Operated Vehicle

LI JI-YONG[1,*], ZHOU HAO[1,*], HUANG HAI[1], YANG XU[2], WAN ZHAOLIANG[1] AND WAN LEI[1]

[1]National Key Laboratory of Science and Technology of Underwater Vehicle, Harbin Engineering University, Harbin 150001, China
[2]State Key Laboratory of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Huang Hai (haihus@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grants 61633009, 51579053, 5129050, 51779052, and 51779059, in part by the Key Basic Research Project of "Shanghai Science and Technology Innovation Plan" under Grant 15JC1403300, and in part by the Field Fund of the 13th Five-Year Plan for the Equipment Pre-Research Fund under Grant 61403120301.

*Li Ji-Yong and Zhou Hao are co-first authors.

ABSTRACT Robot machine capture protein seafood like sea cucumber and seashell is expected to play a great role in food economy improvement and divers protection. In order to realize robot capture, a sea organism absorptive type remotely operated vehicle (ROV) has been designed with the pilot operation and visionbased autonomous capture modes. A novel region-based fully convolutional network with deformable convolutional networks has been developed to realize organism target recognition. The comparisons in off-line experiments have verified its advantages. In order to realize organism target following and capture control, a novel learning-based type-II fuzzy controller has been developed. Through online fuzzy rule optimization and learning, the controller can realize organism target following and capture control under image coordinate without vehicle horizontal velocity or position information in the complicated submarine environment. Field trials have been made in the Zhangzidao Island of China with the designed absorptive type ROV. The trials manifest that the designed absorptive type ROV can realize online organism target recognition, following and capture in the real submarine environment.

INDEX TERMS Remotely operated vehicle vision servo control target recognition.

I. INTRODUCTION

The worldwide demand for high protein seafood, marine drugs, health products such as seashell and sea cucumber, is growing rapidly. Each year the marine products imports of China are tens of billions of dollars with the increase of 10% [1]. For the time being, offshore natural organisms such as sea cucumbers, seashells, urchins, et al. are mostly captured by human divers. However, divers fishing not only suffers from life hazards and hypothermia diseases but also limited by time duration in the water depth of more than 20 meters. Robotic capturing, on the contrary, will both improve the operation safety level and reduce fishing cost.

In the last two decades, underwater robots have advancing dramatically and widely applied for the application of marine science, oceanic engineering, environmental exploration, and so forth [2] [3]. For example, the REMUS-100 AUV [4], Autosub AUV [5] and OceanServer Iver2 AUV [6] are applied for underwater observation and survey; H300 ROV of French ECA group, SAUVIM I-AUV (Intervention AUV) [7], Girona 500 I-AUV [8], and so forth, are developed for the field autonomous operations. For marine organism harvesting and unhurt capture, Norway has developed a submarine harvesting ROV which has realized sea urchin harvesting through remote aspiration manually Khatib *et al.* [9] have invented a multisensory humanoid robotic diver Ocean One for oceanic discovery. Ocean One is equipped with a pair of 7-DOF electrical, compliant and a torque-controlled arm and gentle hands, but it mainly depends on the operator guidance with limited observation and operation range. It is still difficult to realize agile and autonomous capture through target sensing and capture control.

The Remotely Operated Vehicles (ROVs) are commercialized and industrialized for years [10]. They are controlled from the surface with a surface control unit. Although it has been widely used for marine resources exploration and underwater platform maintenance, its operation still depends on one or more comprehensively trained operators. When the ROV confronts with complex targets and tens of DOF vehicle-manipulator systems, the operation will be very challenging [11]

On the autonomous environmental sensing and object recognition, acoustic and visual sensors are usually applied. Acoustic sonars are employed for far and middle range sensing, while visual images are for close observation [12]. Since high frequency sonars are relatively expensive for organism capture, autonomous visual observation is still essential for low cost underwater vehicles [13]

On the underwater visual recognition, some recognition approaches have been applied by using the image features [14]. García *et al.* [15] proposed a generic segmentation process for target identification and selection. Sun *et al.* [16] proposed an automatic recognition algorithm through color-based identification and shape-based identification. But the shape and color features of sea organism are very genetic to their environment such as sea grass and rock due to ecological interest [17] [18]. In order to improve the recognition accuracy between identical features under complex conditions, algorithms of deep learning based object recognition and detection have emerged since 2010. These deep learning algorithms can be divided into two categories, namely, the end-to-end algorithms and the region-proposal-based algorithms. The end-to-end algorithms show great advantage in processing speed. YOLO (You Only Look Once) [19] or SSD (Single Shot Multibox Detector) [20] can reach a processing speed of 45 fps (frames per second) per second because the skips of time-consuming region proposed step, which can reach the requirement of real-time video processing. The region-proposal-based algorithm combines the region proposal method with Convolutional Neural Networks (CNNs), such as Faster Region-based Convolutional Networks (Faster R-CNN) [21], and Region-based Fully Convolutional Networks (R-FCN) [22]. Region proposal method is used to propose Region of Interest (ROI) while a CNN is used to obtain object bounding box and its label [23]. The region-proposal-based algorithms manifest a better recognition accuracy but with a slower processing speed. R-FCN, for example, has an accuracy of 83.6% mAP (mean average precision) and a processing ability of 12 fps on PASCAL VOC data sets. However, like most of the object detection algorithms, R-FCN is also hard to accommodate geometric variations or model geometric transformations in object scale, pose, viewpoint etc. But the living marine organism shows the above-mentioned characteristics. This study will propose a novel R-FCN network with deformable convolutional networks for the high accuracy small marine organism object recognition task.

In order to realize autonomous capture, visual servo based control is essential [24]. The robot observes, measures, and tracks the target based either on the position information or on the image features directly. Myint *et al.* [25] estimated the relative pose through model-based recognition by using 1-step genetic algorithm and performs visual servo by keeping the desired pose to the target. Bonin-Font *et al.* [26] realized visual odometry tracking and intervention through color based feature detection and stereo-3D position computation. Bonin-Font *et al.* [27] used Gaussian mixture model to detect moving targets, and launched a fast compressive tracker with a Kalman prediction mechanism to locate the target position. Guo *et al.* [28] described an integration of a vision system and intervention AUV, and it realized pipe manipulation through calibration, object detection and 3D point cloud based pose estimation. On the other hand, position based visual servo depends on accurate geometric model and calibration parameters [29], which is liable to be affected by the environmental disturbances and parameter uncertainties. Image based visual servo, on the contrary, depends directly on the image feature feedback. Fornas *et al.* [30] proposed an adaptive neural network image-based visual servo controller integrated with image-based visual servo kinematic model. In fact, for the low cost vehicle without its own exact position relative to the target, image based visual servo is more robust with the points or line features Jacobian integrated into controller's feedback [31]. Moreover, in order to realize accurate visual servo and capture control, not only should the camera keep tracking the target continuously, but also should the controller realize precise approaching. This study will propose a reinforcement learning based hierarchical vision controller to present a certain degree of intelligence for the autonomous capture.

This article will focus on the autonomous capture scheme with online visual recognition and learning based visual servo on the basis of an absorptive type ROV system. The main contribution includes:

1) On the recognition of small organism object in its natural genetic environment, a novel Region-based Fully Convolutional Networks (R-FCN) with deformable convolutional network has been proposed. This network not only augments the anchor scales for spatial sampling but also helps offset target and its genetic environment

2) A novel Type-II fuzzy hybrid and intelligent image based visual servo controller with learning based particle swarm optimization (PSO) fuzzy rules optimizer without ROV velocity or position information in the horizontal plane has been proposed to integrate reinforcement learning and type-II fuzzy systems. The controller can keep the organism target in the camera field, continuously approach the target and realize stable absorption through fuzzy rules online quick iteration and optimization.

3) A novel sea organism absorptive type ROV has been developed with pilot operation and vision based autonomous capture modes. Empirical evaluations have been made in the Zhangzidao Island of Da Lian, Liao Ning province, China with the designed absorptive type ROV.

The rest of this article is organized as follows. Section 2 will describe the design and control architecture of the absorptive type sea organism capture ROV. A novel deformable convolutional network based R-FCN will be proposed in section 3 A hybrid and intelligent visual servo

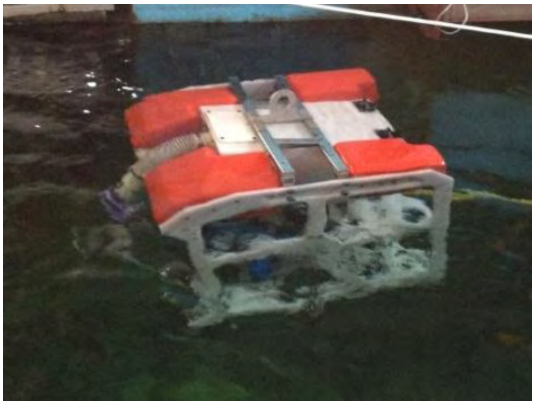


FIGURE 1. Design of sea organism absorptive type ROV.

controller will be proposed to integrate reinforcement learning and type-II fuzzy systems in section 4. Oceanic pasture experiments will be discussed and analyzed in section 5. We will make conclusions in section 6

II. ROV SYSTEM DEVELOPMENT

A. MECHANISM DESIGN

The mechanism design of sea organism absorptive type ROV is described in FIGURE 1. It is a compact-size open frame underwater vehicle designed and developed in the Harbin Engineering University for the capture of sea organism. It is 1.3m in length, 0.8m in width, and 0.9m in height, with the depth rating 100m. The ROV is weight 130kg, with the single entry capture weight 30kg. This low cost sea organism capture ROV is equipped with a magnetic compass, depth gauge, but without position or velocity sensors in the horizontal plane. Autonomous capture will be realized through the depth and direction information from visual features in the camera. In order to realize environmental perception, target recognition and visual servo, the ROV is equipped with a fisheye camera in the front of ROV for pilot operation; a wide view camera is fixed on the absorptive pipe for online recognition, and visual servo control; two spotlights with one focusing angle at 30° and the other one at 60°. These spotlights can provide enough illumination in the darkness seabed. The ROV is installed with four horizontal vector thrusters and two vertical thrusters, with electronic control system in the capsule to realize ROV motions. During the absorptive process, the sea organism will be absorbed through the absorptive pipe, while the sea water and silt will be discharged through drain pipe. The ROV will realize sea organism capture through pilot operation; online target recognition and vision based autonomous absorption control.

B. CONTROL ARCHITECTURE

The control architecture includes underwater electronic control system surface manipulation system (see FIGURE 2). Surface manipulation system includes graphic manipulation interface, manipulation joystick and online recognition and information processing system.

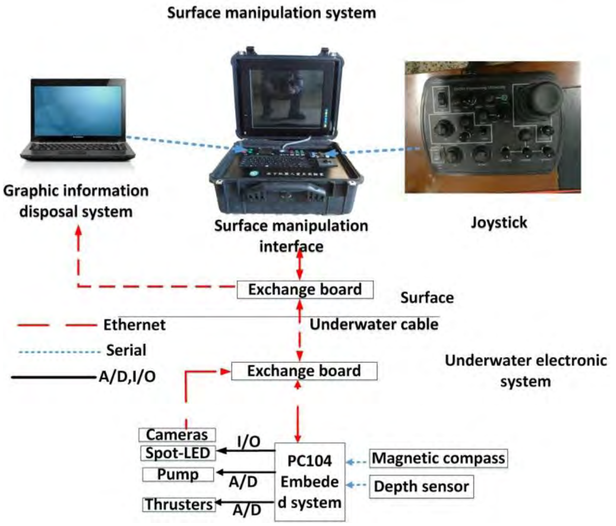


FIGURE 2. Control architecture of absorptive type ROV.

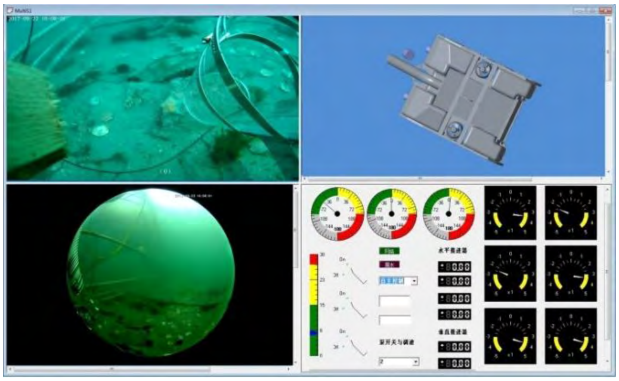


FIGURE 3. The ROV surface manipulation interface.

The ROV surface manipulation system can realize pilot manipulation and vision based autonomous capture manipulation. It includes surface manipulation interface, manipulation joystick and graphic information disposal system. The surface manipulation interface (see FIGURE 3) obtains the views from a fisheye camera, a wide-angle camera, intuitive model heading of vehicle top view model, and the vehicle status of 3 DOFs postures angles (ROV postures revolving around the x, y, and z axes respectively: roll, pitch, yaw), control mode, underwater instruments switch, diving depth, thruster commands are sent from the vehicle capsule. Since pilot operation is very challenging and skills demanding, the purpose of surface manipulation interface is to provide the pilot with a convenient, low-cost and quick training platform. Moreover pilot can operate the ROV advancing, diving, heading, transversal traveling and absorption through manipulation joystick (see FIGURE 4) Graphic information disposal system is a portable computer with ubuntu 16.04 operating system and NVIDIA GTX 1070 graphic card The system realizes online visual recognition and provides target position in the camera for autonomous capture.

The core of underwater electronic system is an embedded PC-104 system with VxWorks operation system. The system

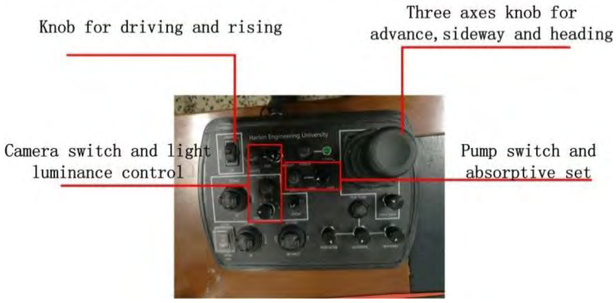

Knob for driving and rising
Three axes knob for advance, sideway and heading
Camera switch and light luminance control
Pump switch and absorptive set

FIGURE 4. Function assignment of the joystick.

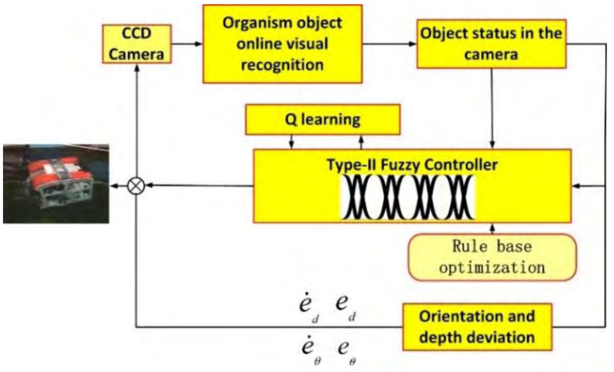

CCD Camera
Organism object online visual recognition
Object status in the camera
Q learning
Type-II Fuzzy Controller
Rule base optimization
$\dot{e}_d$ e_d
$\dot{e}_\theta$ e_θ
Orientation and depth deviation

FIGURE 5. Diagram of vision based autonomous capture control.

plays a major role in the low level control of ROV. It is in charge of sampling sensor information from magnetic compass, depth sensor, etc.; and sending specific commands to the thruster and pump from surface through Ethernet. Three operation mode including pilot remotely operation, hybrid operation mode and vision based autonomous capture mode can be realized.

1) PILOT REMOTELY OPERATION

This mode is commonly applied for typical ROV control architectures, in which a pure man-machine loop has been established mainly to serve as a remotely cruising and absorption manipulation [29]. Underwater environments and vehicle states are obtained through cameras, magnetic compass and depth sensor on the ROV. The pilot can then remotely operate the ROV advancing, diving, heading, transversal traveling and absorption through manipulation joystick PC-104 will explain logic commands to specific commands to control thrusters and pump to complete issued tasks.

2) HYBRID OPERATION MODE

In this mode, autonomous orientation and depth controls are included in the combination with pilot remote operation, which will create a more convenient interactive environment for pilot's operations. For example, effective absorption expects a constant depth during cruising, in the previous mode, pilots are distracted to maintain the depth with joystick, on the contrary, constant cruising with depth control in the hybrid operation mode can make the capture operation more convenient.

3) VISION BASED AUTONOMOUS CAPTURE MODE

The objective of vision based autonomous capture is to realize the marine organism autonomous absorption through online visual recognition and intelligent visual servo control. Through visual recognition, marine organism will be found and locked, autonomous capture will be realized through visual servo controller.

FIGURE 5 illustrates the process diagram of vision based autonomous capture. The following two sections will describe the method and process of online visual recognition and visual servo controller.

III. ONLINE VISUAL RECOGNITION

A. THE R-FCN ALGORITHM

R-FCN algorithm is a region-proposal based algorithm, which follow a two-stage object detection strategy: (1) RoIs proposal, and (2) RoIs classification and regression. Region Knob Network (RPN) [30] is a sub-network of R-FCN and is used for generating RoIs. The input feature maps, which extracted by the base convolution network, are shared between RPN and R-FCN. FIGURE 6 illustrates the overview architecture of the algorithm.

To classify RoIs, $k^2(C+1)$ layers position sensitive score maps are produced by the feature maps through convolutional computation. The $k^2(C+1)$ means to encode the output feature with C categories (and "+1" for background) and each category with k^2 score maps. The bank of k^2 score maps corresponding to a k × k spatial grid illustrates relative positions of the object. For example, when $\mathrm{k} = 3$, it means to encode 9 position information for an object category, such as top-left, top-center, …, bottom-center, bottom-right. The bounding box of RoIs regression is similar with the classification of RoIs. $4k^2$ layers position-sensitive score maps are produced in order to predict 4 coordinates of a RoI's bounding box.

After the position-sensitive score maps are generated, position-sensitive RoI pooling is proposed to encode position information into RoIs. Each RoI bounding box is divided into $k \times k$ bins corresponding to the position-sensitive score maps.

Each bin of RoI comes from pooling the corresponding position score maps. For example, the top-left bin of RoI is pooling from the corresponding area of top-left score maps (the red layers in FIGURE 6).

After the position-sensitive RoIs pooling has been finished, k^2 position-sensitive scores are generated, the k^2 position-sensitive scores then vote on the RoI. For RoIs classification, a (C + 1)-dimensional vector is generated for each RoI, and the (C + 1)-dimensional vector parameterizes the probability of each category. For RoIs' bounding box regression, a 4-dimensional vector is generated for each RoI, the 4-dimensional vector parameterizes the four coordinates of bounding box.

To train the R-FCN, the loss function on each RoI is defined as the summation of the cross-entropy loss and the

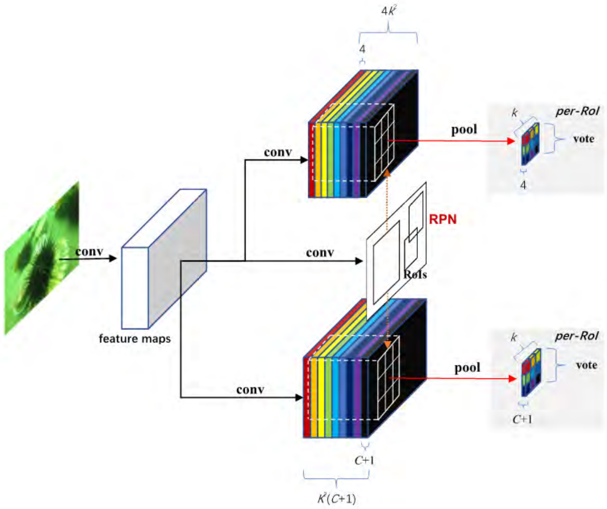


FIGURE 6. The architecture of R-FCN.

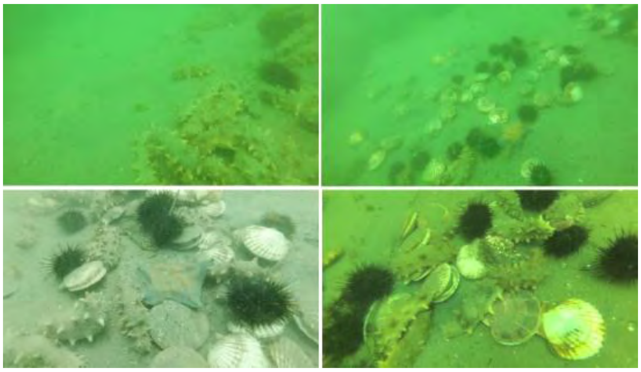


FIGURE 7. Some underwater images.

box regression loss:

$$L\left(\mathrm{s}, t_{x,y,w,h}\right) = L_{cls}\left(s_{c^*}\right) + \lambda\left[C^* > 0\right] L_{reg}(t, t^*). \quad (1)$$

Where, c^* is the ground truth label of RoI ($c^* = 0$ means background). $L_{cls}\left(s_{c^*}\right) = -\log(s_{c^*})$ is the cross-entry loss for classification, where, $s_{c^*} = e^{r_{c^*}} / \sum_{c=0}^{C} e^{r_c}$, and r_c is the score of RoI belonging to the category c. $L_{reg}\left(t, t^*\right) = R(t - t^*)$ denotes the bounding box regression loss, t represents the predicted box, t^* represents the ground truth box, R is a constant, both $L_{cls}(S_c^*)$ and $L_{reg}(t, t^*)$ were defined in [24]. λ is a constant, $\lambda = 1$. $[C^* > 0]$ is an indicator, it is equal to 0 when the ground truth label c^* is background, otherwise, the indicator equals to 1.

B. THE IMPROVED R-FCN ALGORITHM

FIGURE 7 contains some underwater images shot by the ROV. There are two problems should be overcome for the effective detection and recognition of marine organism. At First, marine organism, sea cucumber, sea urchin and scallop for example, are very small in their actual body size, which could only take up a small fraction in underwater image, in the process of cruise and search. Small object detection and recognition is a quite difficult problem for R-FCN. Secondly, the living marine organism shows varied form, varied scale and so on in different situations. However, it's hard for R-FCN to accommodate geometric variations or model geometric transformations in object scale, pose, viewpoint and so on. Therefore, the following improvements will be made on the original R-FCN algorithm in order to overcome these two problems.

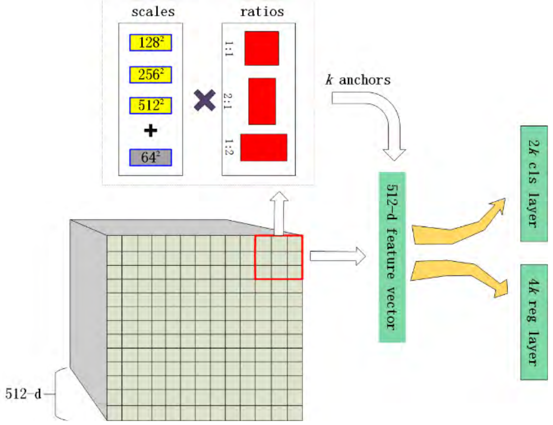


FIGURE 8. The architecture of RPN.

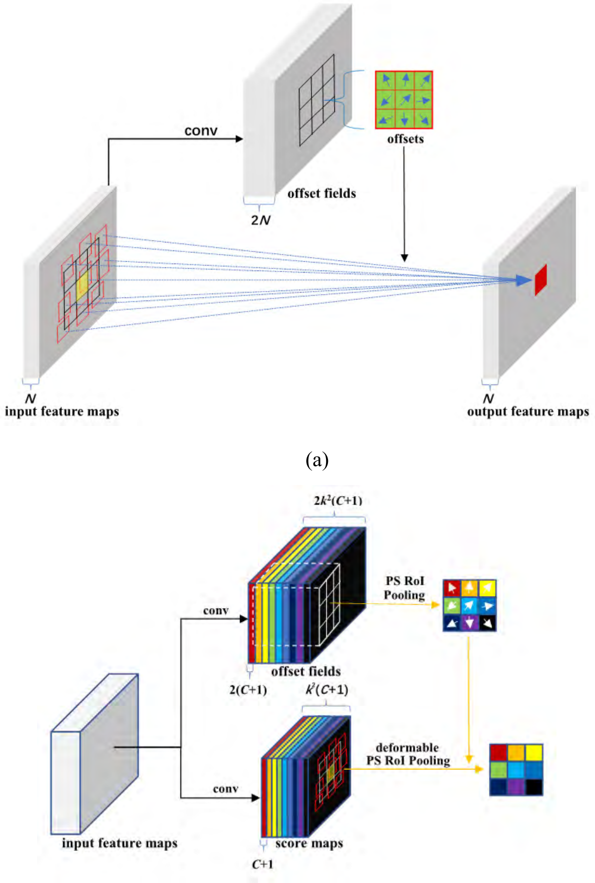
1) IMPROVEMENT IN SMALL OBJECTS DETECTION

In R-FCN algorithm, RPN is used to generate RoIs, object classification and regression is then performed on RoIs. Thus, if RPN can generate more accurate RoIs, the result of object detection and recognition will be improved. FIGURE 8 describes the architecture of RPN. As illustrated in FIGURE 8, each sliding-window generates k anchors with different scales and aspect ratios. And for each sliding-window, a 512dimension feature vector is generated through convolutional computation. So, the ''reg'' layer outputs 4k scores to encode the coordinates of k anchors, and the ''cls'' layer outputs 2k scores that estimate the foreground or background probability of each anchor (see FIGURE 8).

In the original RPN, each sliding-window location yields nine anchors with the combination of scales [128^2, 256^2, 512^2] and aspect ratios [1:1, 1:2, 2:1]. On the land, most objects have an aspect ratio close to 1:1, 1:2 or 2:1. In marine organism detection task, object likes sea urchin, sea cucumber and scallop shows similar aspect ratio as defined above. Thus, the aspect ratios of anchor reflect very strong flexibility. Moreover, most of these objects only count hundreds of pixels in image. It is improper to select 128^2 as the smallest anchor scale, and therefore the anchor scales are defined as [64^2, 128^2, 256^2, 512^2]. Thereafter, each sliding-window will produce 12 anchors and the redefined anchor scales are more inclined to generate small RoIs, therefore the accuracy of small object detection and recognition will be improved.

2) IMPROVEMENT IN LIVING ORGANISM DETECTION

Living marine organism shows varied form, especially in the changeable marine environment. For R-FCN, convolution unit samples the input feature map at fixed locations, and



(a)

(b)

FIGURE 9. Diagram of deformable convolution and deformable position-sensitive RoI pooling. (a) Deformable convolution. (b) Deformable position-sensitive RoI pooling.

pooling layer reduces the spatial resolution at a fixed ratio. So, it's hard for R-FCN to accommodate geometric variations.

The key idea of deformable convolution and deformable position-sensitive RoI pooling is to augment the spatial sampling locations in the modules with additional offsets and learn the offsets from the target tasks, without additional supervision. In the consideration with the deformable characteristic of marine organism, the two new modules are applied in R-FCN algorithm. Deformable convolution is used to replace some original convolution unit of the base network ResNet 101 [31] and deformable position-sensitive RoI pooling is used to replace the original position-sensitive RoI pooling in R-FCN.

FIGURE 9 shows the diagram of deformable convolution and deformable position-sensitive RoI pooling. As illustrated in FIGURE 9, the offset fields are obtained through applying convolutional computation over input feature maps. The channel dimension of offset fields is twice of the input feature maps, corresponding to a 2-dimension (x-axis and y-axis) offsets of each location.

We define a grid receptive field R as:

$$R=\{(-1,-1)\,,(-1,0)\,,\ldots,(0,1)\,,(1,1)\}\tag{2}$$

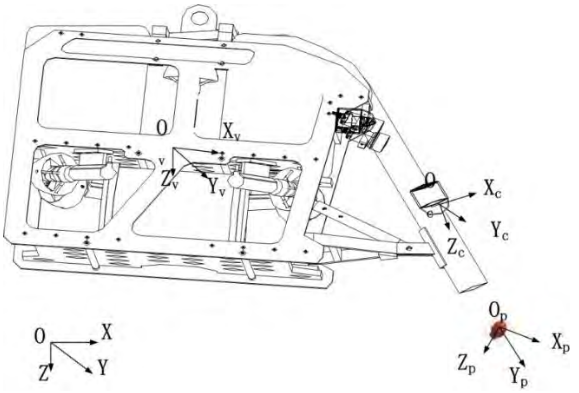


FIGURE 10. Target image vision expressed in terms coordinates.

From traditional CNN, one obtains:

$$\mathrm{O_{ut}}\,(p_0)=\sum\nolimits_{p_n\in R} w(p_n)\cdot I_n\,(p_0+p_n)\tag{3}$$

where p_0 is a point on the output feature map O_{ut}, I_n is the input feature map, p_n enumerates the locations in R.

In the deformable convolution, the regular grid R is augmented with offsets $\Delta p_n|n = 1,\ldots,N$, where $N = |R|$. Then, equation (3) becomes:

$$\mathrm{O_{ut}}\,(p_0)=\sum\nolimits_{p_n\in R} w(p_n)\cdot I_n(p_0+\Delta p_n+p_n)\tag{4}$$

As illustrated in FIGURE 9, deformable position-sensitive RoI pooling is similar with deformable convolution. At first, $2k^2(C+1)$ layers position-sensitive offset fields are generated through convolutional computation. Then, position sensitive ROI pooling on offset fields generates $k \times k$ offsets for $k \times k$ bins of the RoI. Finally, the position-sensitive RoI poolings score maps with additional offset in the location of each bin generates $k \times k$ output features.

IV. INTELLIGENT VISUAL SERVO CONTROLLER

A. KINEMATICS CONTROLLER FOR THE IMAGE CONFIGURATIONS

Vision servo control integrates visual information feedback with robotic orientation and position control. Since the low cost absorptive ROV does not equip with the Doppler Velocity Log (DVL) like position sensor, its visual based control is based on the estimation of displacement and orientation of feature pixels. Vision based capture control can be expressed from FIGURE 10 in terms of image coordinates. The objective of vision based autonomous capture control is to control the ROV movement towards the recognized target, until the target enters into the absorption range of the pipe to realize quick absorption. During the ROV moving towards the recognized organism target, the target feature should be kept within the CCD image plane.

There four coordinate frames have been set in FIGURE 10, namely the global frame $\sum$O-XYZ, the vehicle frame $\sum O_v$-$X_vY_vZ_v$, the camera frame $\sum O_c$-$X_cY_cZ_c$, and the target frame $\sum O_p$-$X_pY_pZ_p$. If we set the vector ($x_t(t)$, $y_t(t)$,

$z_t(t))^T$ as the position of feature point in the vehicle frame, vector $(x_t^c\,(t)\,, y_t^c\,(t))^T$ as the feature point projection on the image plane, one has:

$$\begin{bmatrix} x_t^c(t) \\ y_t^c(t) \\ 1 \end{bmatrix} = \frac{1}{z^c(t)} \boldsymbol{MT} \begin{bmatrix} x_t(t) \\ y_t(t) \\ z_t(t) \\ 1 \end{bmatrix}, \tag{5}$$

where $z^c(t)$ is the feature point depth in the camera frame, $\boldsymbol{M}$ is a 3×4 matrix determined by the intrinsic parameters of the camera.

$$\boldsymbol{M} = \begin{pmatrix} \alpha_c & -\alpha_c \cot \varphi_c & u_c & 0 \\ 0 & \gamma_c / \sin \varphi_c & v_c & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where α_c and γ_c are the scalar vectors of the u_c and v_c axis in the image plane. u_c and v_c represent the position of the principle point of the camera. φ_c represents the angle between the two axes [32]. $\boldsymbol{T}$ is the homogeneous transformation matrix of the vehicle frame with respect to the camera frame,

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{d}_t \\ 0 & 1 \end{bmatrix}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are the rotation and translation components of the transformation matrix. Thus (4) can be rewritten as:

$$\begin{bmatrix} x_t^c(t) \\ y_t^c(t) \end{bmatrix} = \frac{1}{z^c(t)} \boldsymbol{M}_p \boldsymbol{T} \begin{bmatrix} x_t(t) \\ y_t(t) \\ z_t(t) \\ 1 \end{bmatrix}, \tag{6}$$

where $\boldsymbol{M}_p$ is the 2×4 sub matrix of matrix $\boldsymbol{M}$.

$$\boldsymbol{M}_p = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \end{bmatrix}$$

If $\boldsymbol{m}_i^T$ denotes the ith row vector of matrix $\boldsymbol{M}$, the depth of the feature point is:

$$z^c(t) = \boldsymbol{m}_3^T \boldsymbol{T} \begin{bmatrix} x_t(t) \\ y_t(t) \\ z_t(t) \\ 1 \end{bmatrix}. \tag{7}$$

Since monocular ranging often involves distance deviation, this paper will propose a goal-oriented visual servo scheme in this section. From (6) and (7), the camera can provide deviation information of depth and orientation for the vehicle. If we set $\square_d$ denotes desired position of □, the expected orientation and depth of the target feature point can be expressed as:

$$\theta_d = arctg \frac{y_{td}(t)}{x_{td}(t)} \quad \text{and } z_d^c(t) = \boldsymbol{m}_3^T \boldsymbol{T} \begin{bmatrix} x_{td}(t) \\ y_{td}(t) \\ z_{td}(t) \\ 1 \end{bmatrix}.$$

The orientation and depth deviation can be expressed as:

$$\begin{cases} e_\theta(t) = \theta_d(t) - R(\theta)\theta(t) \\ e_d(t) = z_d^c(t) - z^c(t), \end{cases} \tag{8}$$

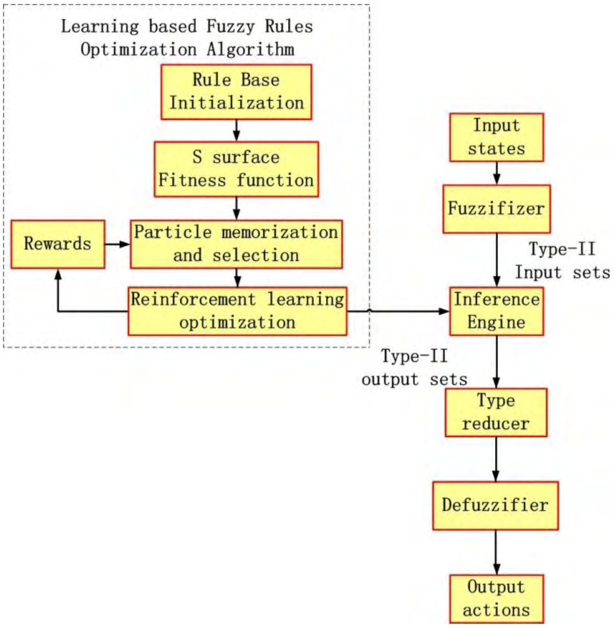


FIGURE 11. Diagram of Type-II Fuzzy Learning Controller.

where $R(\theta)$ is the rotation function of orientation angle of the vehicle frame with respect to camera frame. Therefore, the vision based control can be realized through orientation and distance control.

B. BLEARNING BASED TYPE-II FUZZY CONTROLLER

The learning based type-II fuzzy approaching controller should not only keep the visual feature within the image, but also enable the pipe approaching the target. The advantage of this controller is to autonomously improve control performance through the action-selection policy.

Moreover, for low cost vehicle without horizontal velocity or position sensor, it is important to keep the target feature within the image and enable the pipe approaching the target. In compare with type-I fuzzy systems, type-II fuzzy system can help minimize uncertain effects of unknown environment. The following will integrate nonlinear controller for ROV stable and accurate motions control and type-II fuzzy learning to ensure the target insight and continuously closing. A novel learning based PSO fuzzy rules optimizer has been developed for uncertain environmental and dynamic change and factors during capturing process (See FIGURE 11)

From FIGURE 12, states $\boldsymbol{St} = \{pe_1\, pe_2 \ldots pe_n\}$ are defined as a set of position states and errors of target in the field of camera. The 2048×1536 pixels CCD image plane can be divided into 32×24 discrete-grid with each grid contain 64×64 pixels. FIGURE 12 can illustrate current and expected states of depth and orientation of the pipe, the field of camera include desirable states area, approaching states area, dangerous states area, safe states area, et al.

1) TYPE-II FUZZY CONTROLLER

The type-II fuzzy inference system is designed to construct action rules and decide action output ranges according to

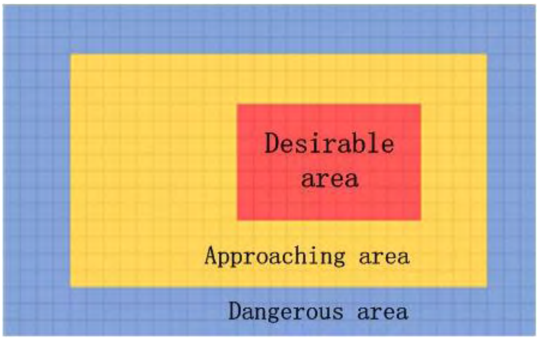


FIGURE 12. Discrete grid of 2048×1536 CCD image plane.

current and expected states of depth and orientation of the pipe. This system contains a type-reducer and a normal defuzzfier. The type-reducer maps a type-II fuzzy set into a type-I fuzzy one, while the defuzzfier transforms the fuzzy output into the crisp one. The proposed type-II fuzzy set is defined as following:

$$\tilde{\boldsymbol{D}}\{((pet(t), \mu(t)), \mu_{\tilde{D}}(pe(t), u(t)))|\forall pe(t) \in \boldsymbol{St}, \forall \boldsymbol{J}_s \supseteq [0, 1]\} \tag{9}$$

where $\mu_{\tilde{D}}(pe(t), u(t)) \in [0, 1]$, $pe(t)$ represents the input state element of the target in the field of camera, $\boldsymbol{St}$ denotes the state set of $pe(t)$, $\boldsymbol{J}_s$ represents the membership of $pe(t)$ in the $\boldsymbol{St}$.

A Gaussian primary membership function can be expressed as:

$$\mu_{\tilde{\boldsymbol{D}}_{ij}} = \exp[-\frac{1}{2}\left(\frac{pe(t)_{ij} - m_{ij}}{\sigma_{ij}}\right)^2], \tag{10}$$

where m_{ij} is uncertain mean $m_{ij} \in \left[m_{ij}^1, m_{ij}^2\right]$, σ_{ij} is fixed standard deviation, $\mu_{\tilde{\boldsymbol{D}}_{ij}}(t)$ denotes the membership degree, which is a bounded set $\mu_{\tilde{\boldsymbol{D}}_{ij}} \in \left[\underline{\mu}_{\tilde{\boldsymbol{D}}_{ij}}, \bar{\mu}_{\tilde{\boldsymbol{D}}_{ij}}\right]$, where $\underline{\mu}_{\tilde{\boldsymbol{D}}_{ij}}$ and $\bar{\mu}_{\tilde{\boldsymbol{D}}_{ij}}$ are the lower and upper bound respectively:

$$\bar{\mu}_{\tilde{\boldsymbol{D}}_{ij}} = \begin{cases} \mu_{\tilde{\boldsymbol{D}}_{ij}}\left(m_{ij}^1, \sigma_{ij}, pe(t)_{ij}\right), & pe(t)_{ij} < m_{ij}^1 \\ 1, & m_{ij}^1 \le pe(t)_{ij} \le m_{ij}^2 \\ \mu_{\tilde{\boldsymbol{D}}_{ij}}\left(m_{ij}^2, \sigma_{ij}, pe(t)_{ij}\right), & pe(t)_{ij} > m_{ij}^2 \end{cases}$$

and

$$\mu_{\tilde{\boldsymbol{D}}_{ij}} = \begin{cases} \underline{\mu}_{\tilde{\boldsymbol{D}}_{ij}}\left(m_{ij}^2, \sigma_{ij}, pe(t)_{ij}\right) & pe(t)_{ij} \le \dfrac{m_{ij}^1 + m_{ij}^2}{2} \\ \mu_{\tilde{\boldsymbol{D}}_{ij}}\left(m_{ij}^1, \sigma_{ij}, pe(t)_{ij}\right) & pe(t)_{ij} > \dfrac{m_{ij}^1 + m_{ij}^2}{2}. \end{cases} \tag{11}$$

The fuzzy operation is implemented through algebraic product operation. We compute the firing strength corresponding with the ith rule:

$$\boldsymbol{F}_i = \begin{cases} \left[\underline{f}_i, \bar{f}_i\right] \\ \underline{f}_i = \prod_{j=1}^{ni} \underline{\mu}_{\tilde{\boldsymbol{D}}_{ij}}, \bar{f}_i = \prod_{j=1}^{ni} \bar{\mu}_{\tilde{\boldsymbol{D}}_{ij}}, \end{cases} \tag{12}$$

where $\underline{j}_i$ and j_i are the lower and upper firing strength respectively. Therefore the left most point $\boldsymbol{u}_L$ and the right most point $\boldsymbol{u}_R$ can be expressed as:

$$\boldsymbol{u}_L = \frac{\sum_{i=1}^{Lc} \bar{f}_i a_i + \sum_{i=Lc+1}^{mr} f_i a_i}{\sum_{i=1}^{Lc} \bar{f}_i + \sum_{i=Lc+1}^{mr} \underline{f}_i},$$
$$\boldsymbol{u}_R = \frac{\sum_{i=1}^{Rc} \underline{f}_i a_i + \sum_{i=Rc+1}^{mr} \bar{f}_i a_i}{\sum_{i=1}^{Rc} \underline{f}_i + \sum_{i=Rc+1}^{mr} \bar{f}_i}, \tag{13}$$

where mr is the number of the rules, Lc and Rc are the left and right crossover points respectively. Thus the defuzzified output is the average of $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$:

$$\boldsymbol{u}_o = \frac{(\boldsymbol{u}_R + \boldsymbol{u}_L)}{2}. \tag{14}$$

2) LEARNING BASED FUZZY RULES OPTIMIZATION ALGORITHM

In this paper, the purpose of fuzzy rules is to choose the ROV action to approach the target for absorption according to current and expected status of depth and orientation of the ROV pipe. The rules can be expressed as:

Rules: If pe_1(t) is $\boldsymbol{St}_1$ and ,…, pe_n(t) is $\boldsymbol{St}_n$, then u_1(t) is a_1(t) and,…, u_m (t) is a_m(t)

Where action set $\boldsymbol{A}(t) = \left\{a_1(t)\ a_2(t)\ \ldots\ a_m(t)\right\}$ represents a set of actions for the vehicle. The vehicle will execute an action such as advance, sideway, heading and diving motions through state evaluation in order to stabilize its pose, keep the target in the camera, control the ROV approach the target and be ready for absorption.

The state evaluation function will predict return through state-action pairs. In the consideration with different field environment during the capture process, manual selection with fixed fuzzy rules is difficult to be universally feasible. This study will optimize fuzzy rules through the following improved particle swarm optimization algorithm.

Improved particle swarm optimization (PSO) algorithm has been applied as an intelligent evolutionary algorithm to train and optimize fuzzy rules. In the PSO algorithm [33], each particle searches and adjusts its status to find the optimal solution according to the experiences of the particle and its neighbors. In other words, each primarily proposed singleton-type fuzzy rule is purposed as a particle candidate so that the action can be optimize through improved PSO algorithm.

Since the PSO algorithm is likely to be affected by the initial state, the inertial weight ω has been adopted to balance local search and global search ability in the improved PSO algorithm. The improved PSO is defined as:

$$\begin{cases} v_k^c(t+1) = \omega v_k^c(t) + c_1 \times r() \times \left(P_{best_k^d} - x_k^c(t)\right) \\ + c_2 \times r() \times \left(x_k^c(t) - P_{worst_k^d}\right) + \lambda_3 \\ \times r() \times \left(G_{best_k^d} - x_k^c(t)\right) \\ x_k^c(t+1) = x_k^c(t) + v_k^c(t+1), \end{cases} \tag{15}$$

where $v_k^c(t)$ is the kth particle current speed, $k = 1, \ldots, k_p$, k_p is the particle population size, c_1 and c_2 are constant acceleration coefficients, $P_{best_k^d}$ denotes the best previous position for the kth particle, $G_{best_k^d}$ denotes the best previous position of all the particles in the swarm, $P_{worst_k^d}$ denotes the worst previous position for the kth particle, $x_k^c(t)$ denotes current position for the kth particle, $r()$ denotes the random number between one and zero. Moreover, the weight ω is obtained from:

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{N_{\max}} \times N_c. \tag{16}$$

where $N_{\max}$ is the maximum iteration number, N_c is the current iteration number.

The improved PSO fuzzy rule optimization algorithm consists of five major stages: initialization, fitness function determination, particle memorization and selection. The optimization process is described in the following:

I. Initialization: Before proceeding improved PSO optimization, fuzzy rules of actions vs input state have been randomly generated. The population of rule particle size is set to be 15, and the dimension of the fuzzy rule is 4, corresponding to 4 DOFs motions of ROV.

II. Fitness function determination: For each trial of fuzzy rule, the fitness function is very important to decide the best actions for the controller. Since the designed learning based type-II fuzzy approaching controller should realize the accurate and stable absorption capture online, fuzzy rules can not be literately optimized from each step of control errors, which not only will reduce the controller convergence speed, but also may involve adverse effect. Therefore, to select an effective and intelligent fitness function is important to realize fuzzy rules quick iteration and optimization in each step length of control command release. In order to realize quick target approaching and absorption maintenance, the control actions are expected to be loosely large when the deviation is relatively large, to be strictly small when the deviation is relatively small. Sigmoid function (see FIGURE 13) of equation (21) is one of these functions.

$$u = \frac{2}{1 + \exp(-k_p e - k_d \dot{e})} - 1, \tag{17}$$

where k_p and k_d are the proportional and derivative gains.

$$FIT = \sum_{i=1}^{m} \left| \frac{k_e(|e_i| + |\dot{e}_i|)}{k_u(\frac{2}{1+\exp(-k_p e_i - k_d \dot{e}_i)} - 1) - a_i} \right|, \tag{18}$$

where k_e and k_u are adjustable parameters, e_i and a_i are the error of a certain direction and corresponding actions of fuzzy rule.

III. Particle memorization and selection: each rule particle will be evaluated through the memorization of its own fitness value and the selection of the maximum one as $P_{best_k^d}$, the maximum rule vector is obtained as:

Therefore, the fitness function is defined as:

$$P_{best} = \left[P_{best_1^d}, P_{best_2^d}, \ldots, P_{best_{15}^d} \right]. \tag{19}$$

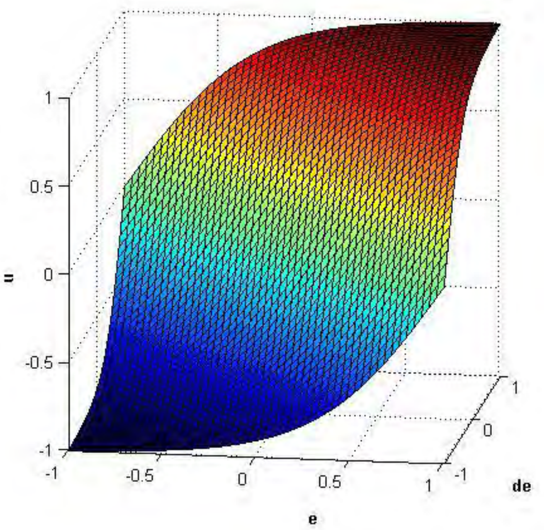


FIGURE 13. Sigmoid function.

IV. Learning based particles further optimization: On the other hand, some uncertain factors except environmental effects will carry out great effects on the vehicle. For example, the capturing weight change could cause disturbance and dynamic change during the capture process. Since the Q-learning algorithm [34] is very effective in improving the controller's robustness, it is applied to further optimize the fuzzy rule actions.

The Q-learning algorithm is purposed to predict the and optimized rule output from the mapping of the state and action pairs of current vehicle pose & position state and rule action pairs. According to Q-learning algorithm, the action a_k will be updated as follows:

$$\begin{aligned} Q_{t+1}(pe(t), a(t)) &= Q_t(pe(t), a(t)) \\ &+ \alpha \left[r(t+1) + \gamma Q_{best}(pe(t+1)) - Q_t(pe(t), a(t)) \right] \end{aligned} \tag{20}$$

where $r(t+1)$ is the immediate reinforcement reward, α and γ are the discount parameter and learning rate respectively, $\gamma \in [0, 1]$, $Q_{best}(pe(t+1))$ is the best estimation of $Q_t(pe(t+1))$ value. The Q value will be updated as:

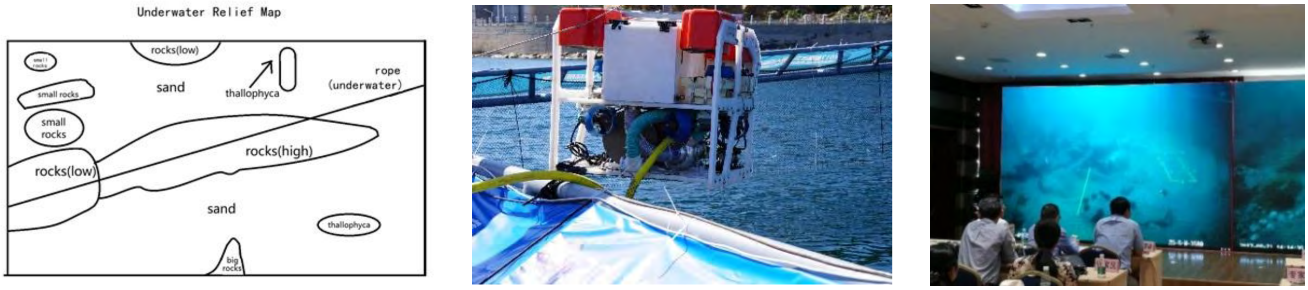
$$\Delta Q = r(t+1) + \gamma Q^*(pe(t+1)) - Q(pe(t), a(t)). \tag{21}$$

Through type-II fuzzy operation, Q values are updated at each control time step, the expected Q value output for each rule and action is:

$$Q(pe(t+1)) = \frac{1}{2} \sum_{i=1}^{mr} \left(\frac{f_i(pe(t))}{\sum_{j=1}^{mr} \underline{f}_j(pe(t))} + \frac{\bar{f}_i(pe(t))}{\sum_{j=1}^{mr} \bar{f}_j(pe(t))} \right) q_i(t), \tag{22}$$

where $q_i(t+1) = q_i(t) + \varepsilon \Delta q_i(t)$ $i = 1, \ldots, mr$, ε is the learning rate.

$$\Delta q_i(t) = \frac{1}{2} \Delta Q \left(\frac{f_i(pe(t))}{\sum_{j=1}^{mr} \underline{f}_j(pe(t))} + \frac{\bar{f}_i(pe(t))}{\sum_{j=1}^{mr} \bar{f}_j(pe(t))} \right). \tag{23}$$



(a)

(b)

(c)

FIGURE 14. URPC introduction. (a) Terrain of the net cage. (b) Robot entry of the net cage. (c) Large screen of the contest.

TABLE 1. Performace comparisons between the original R-FCN and improved R-FCN on the data sets of marine organism.

Methods	mAP@0.5 / 0.7	AP@0.5 / 0.7		
		Sea cucumber	Sea urchin	Scallop
R-FCN	90.23 / 80.04	90.48 / 80.34	89.80 / 79.51	90.42 / 80.28
improved R-FCN	90.65 / 81.24	90.83 / 81.49	90.49 / 81.03	90.62 / 81.20

TABLE 2. Detection comparisons on the multi scal marine organism between the original R-FCN and the improved R-FCN.

Methods	mAP@0.5 / @0.7		
	Small	Medium	Lager
R-FCN	35.45 / 22.40	76.86 / 59.73	89.98 / 85.84
improved R-FCN	42.94 / 29.78	79.33 / 71.24	90.27 / 89.92

Objects are divided into three categories according to the bounding box area, i.e. small: area < 28^2 pixel, medium: 28^2 < area < 49^2, large: 49^2 < area.

Based on equation (23), each particle is further modified. and steps I, II, III and IV are repeated iteratively until G_{best} is obviously improved. The particle with best fitness of P_{best} is the global best one G_{best^d}. Therefore the best fuzzy rule fitness value is obtained to deal with environmental disturbance and pose change during capturing process.

V. FIELD TRIALS IN ZHANGZIDAO ISLAND

From September 19th to September 23th, the first Underwater Robot Picking Contest (URPC) was launched in Zhangzidao Island sponsored by National Natural Science Foundation of China (NSFC). This objective of this contest is to promote underwater robot autonomous perception and manipulation for the organism target machines agilely capture. The major items in the contest include Object recognition off-line on the computer and online recognition with robot; and autonomous perception and manipulation capture. 16 teams from different universities and companies participated this contest. In order to launch this contest, the contest organizers, Dalian University of Technology and Zhangzidao Group Corporation, have constructed a 15m×15m (length × width) net cage in the offshore pasture field and laid a great number of sea organisms such as sea cucumbers, sea urchins and scallops (see FIGURE 14). Moreover, the organizers have provided a great many submarine organism photo samples for the training of objects off-line recognition.

A. COMPARISONS OF OFF-LINE VISUAL RECOGNITION

The marine organism detection and recognition model have been trained from the data sets provided from the URPC organizing committee. There are totally 12882 labeled underwater images in the data sets, the data annotation files sets are provided in PASCLA VOC format. The model is trained on an Ubuntu 16.04 computer with an NVIDIA GTX 1070 graphic card (8 GB graphic memory).

The mean average precision (mAP) scores are applied in (Everingham *et al.* [35] 2010) for the evaluation of the model. And the greater the mAP scores, the better the model. In the off-line experiments, mAP scores are applied to as an Intersection-over-union (IoU) thresholds valued at 0.5 and 0.7. When the IoU threshold is set as 0.7, the evaluation criteria will be more strict.

In the off-line experiments (see FIGURE 15), the outputs from the original R-FCN and the improved R-FCN are visualized in FIGURE 15. The evaluation results of two models are illustrated in TABLE 1 and TABLE 2.

In TABLE 1, the effects of the improved R-FCN are evaluated on the living marine organism detection. Although the living marine organism shows varied forms, scales and

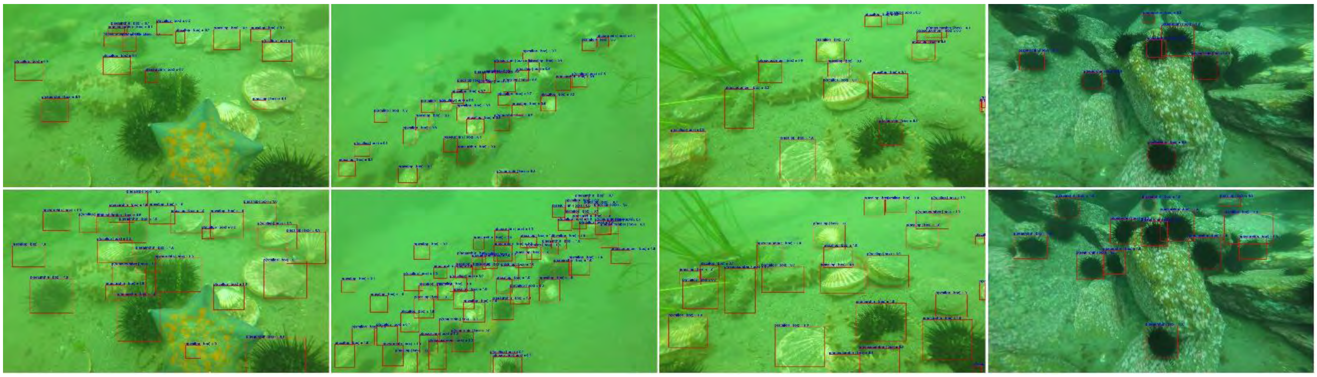


FIGURE 15. R-FCN (top) vs. improved R-FCN (bottom) on marine organism recognition task.

so on in different situations, the improved R-FCN algorithm which take deformable convolution and deformable position sensitive RoI pooling into consideration, manifests the detection accuracy improvement in different categories of living marine organism.

In TABLE 2, the effects of the improved R-FCN are evaluated in small object detection. From TABLE 2, the improved R-FCN manifests detection accuracy improvement on different scales, especially in small scale object. The mAP@0.5 scores are 35.45% and 42.94% respectively, from the comparisons between the original R-FCN and the improved R-FCN on small scale object detection Moreover, the mAP@0.7 scores are 22.40% and 29.78% respectively, from the comparisons between the original R-FCN and the improved R-FCN in small scale object detection.

Therefore, the improved R-FCN algorithm manifests improvement on the living marine organism and small scale objects detections

B. ONLINE VISUAL RECOGNITION

In compare with off-line recognition, Online visual recognition means the realization of recognition during vehicle cruising. It requires the detection and recognition algorithm being more robust during the vehicle moving and shaking. Since the capsule space of underwater vehicle is compact while the graphics card is too large to integrate into the capsule. So, an external laptop with GTX 1070ti was used for object detection.

In the online recognition contest, the evaluation criterion is described as follow:

$$error = \sum_{i=1}^{3} w_i\,(t_i - g_i)^2 \,/ \sum_{i=1}^{3} g_i, \quad \sum_{i=1}^{3} w_i = 1, w_i \geq 0, \tag{24}$$

where t_1, t_2 and t_3 are the numbers of detected object, i.e. sea cucumber, sea urchin and scallop; g_1, g_2 and g_3 are the numbers of labeled object, i.e. sea cucumber, sea urchin and scallop, which are labeled by the URPC organizing committee; $w_1 w_2$ and w_3 are the weights of different categories. The smaller the *error* scores, the better the performance of the online recognition algorithm.

TABLE 3. Performace of the online visual recognition contest.

Categories	Sea cucumber	Sea urchin	scallop
predicted	0	9	52
labeled	0	14	57

TABLE III shows the results of online visual recognition contest. FIGURE.16 reflects some video frames of the online recognition results. From FIGURE.16, the improved R-FCN algorithm of this article manifests good performance during online recognition process.

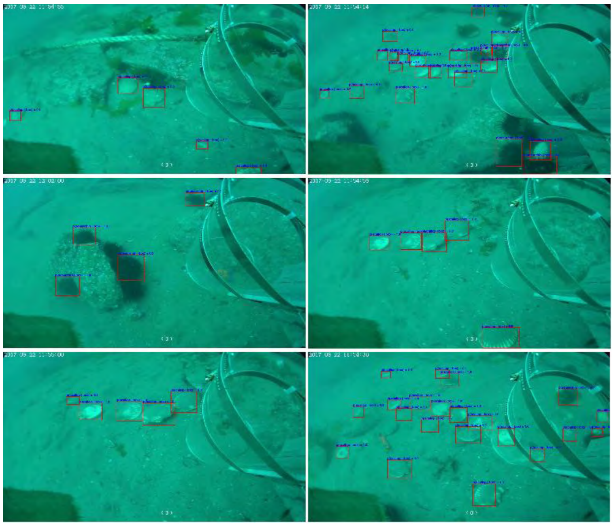


FIGURE 16. Some video frames in the online recognition.

C. VISUAL SERVO CONTROL FOR AUTONOMOUS CAPTURE

The autonomous capture process is manifested from FIGURE 17 to FIGURE 21. FIGURE 19 and FIGURE 20 describe two target following scenarios. The authors have compared the experimental results of target following and absorption process with and without learning based fuzzy rules optimization algorithm of section IV. FIGURE 20 describes the target switching process from disturbance and blocks. The relationship between pixel distance and real distance is obtained from the underwater fixed height calibration and measurement.

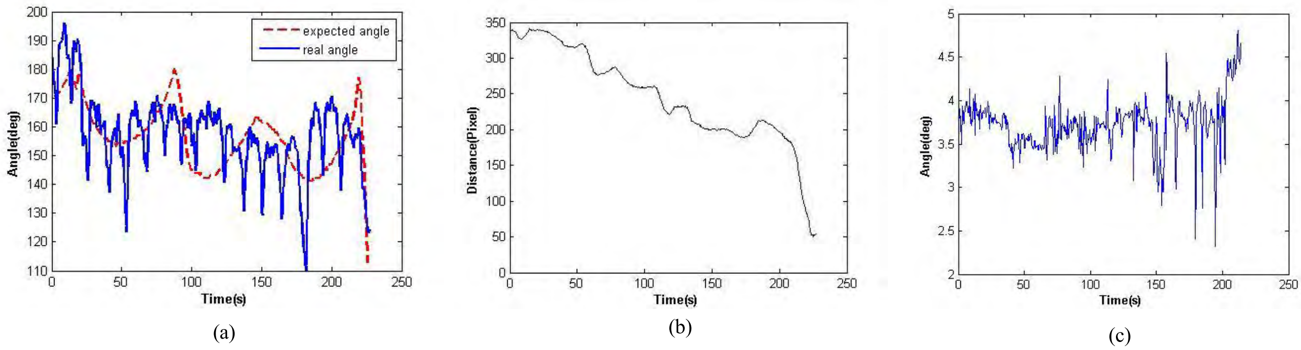


FIGURE 17. Autonomous capture process of scenario I without learning based fuzzy rules optimization algorithm. (a) Angle comparisons. (b) Pixel distance between target and pipe during capture. (c) Vehicle rolling state.

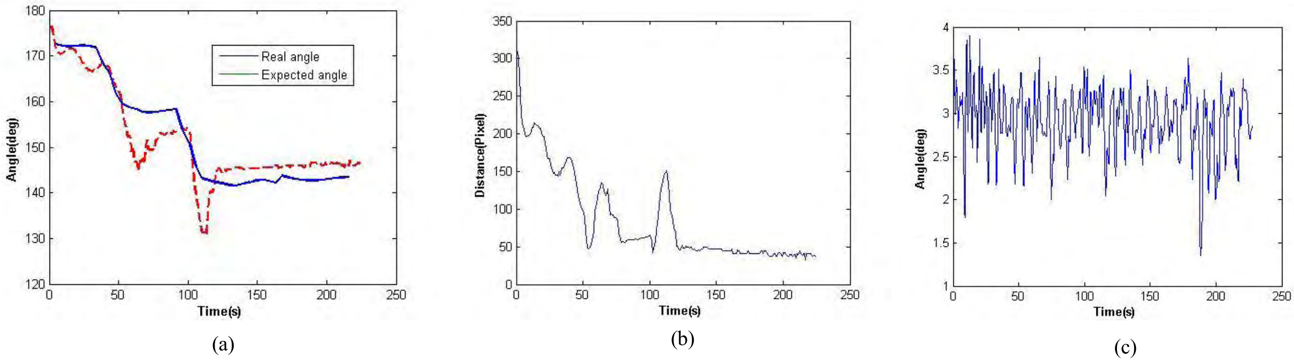


FIGURE 18. Autonomous capture process of scenario I with learning based fuzzy rules optimization algorithm. (a) Angle comparisons. (b) Pixel distance between target and pipe during capture. (c) Vehicle rolling state.

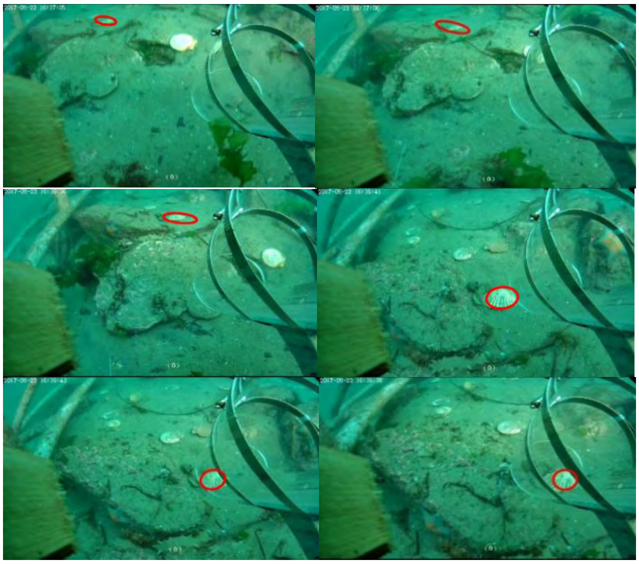


FIGURE 19. Video frames in the autonomous capture process scenario I.

FIGURE 17 (b) and FIGURE 18 (b) describe the pixel distance between target and pipe during capture of FIGURE 19, while FIGURE 17 (c) and FIGURE 18 (c) describe the ROV rolling states during capture of FIGURE 19. From Figure 17 and 18, although the ROV can realize target following and absorption through type-II fuzzy visual servo

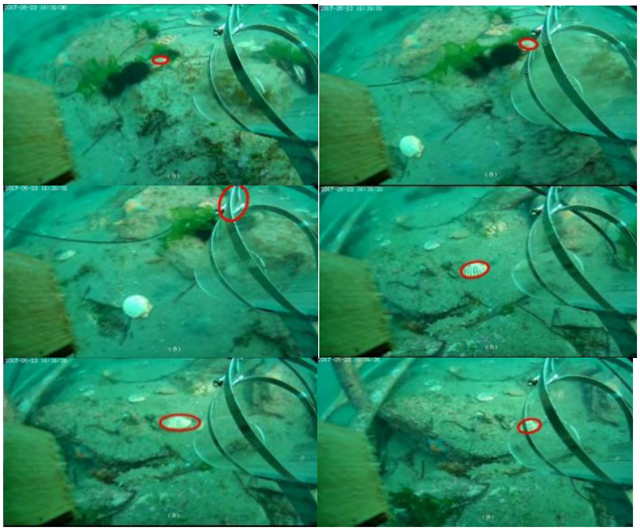
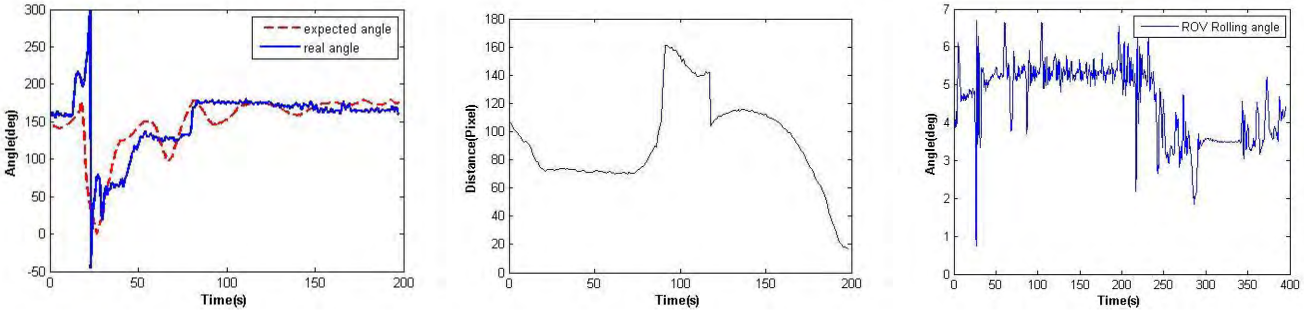


FIGURE 20. Video frames in the autonomous capture process scenario II.

controller, environmental disturbance and rolling change during the capture could cause the vehicle shaking and partly unstable during the capture process. Although the type II fuzzy controller is advantaged in the control of nonlinear process and external disturbances, the target following still experienced some fluctuation process. With learning based fuzzy rules optimization algorithm of Section IV, the



(a) (b) (c)

FIGURE 21. Autonomous capture process of scenario II with learning based fuzzy rules optimization algorithm. (a) Angle comparisons (b) Pixel distance between target and pipe during capture (c) Vehicle rolling state.

following and absorption process can be more stable and convergence.

Furthermore, experiments in FIGURE 20 and 21 describe another process, in which the environmental disturbance and target block cause the target switch. But robot can still realize capture with learning based fuzzy rules optimization algorithm of Section IV. In the autonomous capture process, the proposed learning based type-II fuzzy controller can realize target following in complicated submarine environment. The learning based fuzzy rules optimization algorithm can optimize following rules and adapt to the environment. Although the expected angle is disturbed by robot shaking and image distortion, the robot can still realize target capture.

VI. CONCLUSION

This study has designed a novel sea organism absorptive type ROV with pilot operation and vision based autonomous capture modes. In order to realize online recognition of target organism, a novel region-based fully convolutional network with deformable convolutional network has been developed to realize organism target recognition. From off-line and online experiment, the recognition accuracy can reach 90.23 mAP (see TABLE 2). A novel learning based type-II fuzzy controller has been developed to realize organism target following and capture control. Through particle optimization and learning, the learning based fuzzy rules optimization algorithm could optimize the following process under vehicle dynamic change and external disturbance, particularly when the target blocked. The controller can realize organism target following and capture control under image coordinate without vehicle velocity or position information in the horizontal plane under complicated submarine environment. Empirical trials manifest the designed absorptive type ROV can realize online organism target recognition, following and capture in the real submarine environment

REFERENCES

[1] G. Krause. *Trip Report for the November 2014 Marketing Mission to China*. Accessed: Nov. 2014. [Online]. Available: http://www.pscha.org

[2] K. He, R. Wang, D. Tao, J. Cheng, and W. Liu, "Color transfer pulse-coupled neural networks for underwater robotic visual systems," *IEEE Access*, vol. 6, pp. 32850–32860, 2018.

[3] P. Ridao, M. Carreras, D. Ribas, and R. Garcia, "Visual inspection of hydroelectric dams using an autonomous underwater vehicle," *J. Field Robot.*, vol. 27, no. 6, pp. 759–778, 2010.

[4] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, Oct. 2010, pp. 4396–4403.

[5] M. E. Furlong, D. Paxton, P. Stevenson, M. Pebody, S. D. McPhail, and J. Perrett, "Autosub Long Range: A long range deep diving AUV for ocean monitoring," in *Proc. IEEE/OES Auto. Underwater Vehicles*, Southampton, U.K., Sep. 2012, pp. 1–7.

[6] C. M. Clark *et al.*, "Tracking and following a tagged leopard shark with an autonomous underwater vehicle," *J. Field Robot.*, vol. 30, no. 3, pp. 309–322, 2013.

[7] G. Marani, S. K. Choi, and S. K. Choi, "Underwater autonomous manipulation for intervention missions AUVs," *Ocean Eng.*, vol. 36, no. 1, pp. 15–23, 2009.

[8] E. Simetti, G. Casalino, S. Torelli, A. Sperindé, and A. Turetta, "Floating underwater manipulation: Developed control methodology and experimental validation within the TRIDENT project," *J. Field Robot.*, vol. 31, no. 3, pp. 364–385, 2014.

[9] O. Khatib *et al.*, "Ocean one: A robotic avatar for oceanic discovery," *IEEE Robot. Automat. Mag.*, vol. 23, no. 4, pp. 20–29, Dec. 2016.

[10] R. Bogue, "Underwater robots: A review of technologies and applications," *Ind. Robot, Int. J.*, vol. 42, no. 3, pp. 186–191, 2015.

[11] J. Zhang, W. Li, J. Yu, X. Feng, Q. Zhang, and G. Chen, "Study of manipulator operations maneuvered by a ROV in virtual environments," *Ocean Eng.*, vol. 142, pp. 292–302, Sep. 2017.

[12] G. Marani and J. Yuh, *Introduction to Autonomous Manipulation*. Berlin, Germany: Springer, 2014.

[13] D. Lee, G. Kim, D. Kim, H. Myung, and H.-T. Choi, "Vision-based object detection and tracking for autonomous navigation of underwater robots," *Ocean Eng.*, vol. 48, pp. 59–68, Jul. 2012.

[14] A. Elibol, J. Kim, N. Gracias, and R. Garcia, "Efficient image mosaicing for multi-robot visual underwater mapping," *Pattern Recognit. Lett.*, vol. 46, pp. 20–26, Sep. 2014.

[15] J. C. García, J. J. Fernández, P. J. Sanz, and R. Marín, "Increasing autonomy within underwater intervention scenarios: The user interface approach," in *Proc. IEEE Int. Syst. Conf.*, San Diego, CA, USA, Apr. 2010, pp. 71–75.

[16] F. Sun, J. Yu, S. Chen, and D. Xu, "Active visual tracking of free-swimming robotic fish based on automatic recognition," in *Proc. 11th World Congr. Intell. Control Automat.*, Shenyang, China, Jun./Jul. 2014, pp. 2879–2884.

[17] A. Burguera, F. Bonin-Font, J. L. Lisani, A. B. Petro, and G. Oliver, "Towards automatic visual sea grass detection in underwater areas of ecological interest," in *Proc. 21st IEEE Int. Conf. Emerg. Technol. Factory Automat.*, Berlin, Germany, Sep. 2016, pp. 1–4.

[18] J. Li, R. M. Eustice, and M. Johnson-Roberson, "High-level visual features for underwater place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2015, pp. 3652–3659.

[19] *International Conference on Robotics and Automation*, Washington State Convention Center, Seattle, WA, USA, May 2015, pp. 3652–3659.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Veges, NV, USA, Jun. 2016, pp. 779–788.

[21] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[23] J. Dai, L. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 379–387.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[25] Y. Wang, S. Jiang, B. Chen, and H. Wu, "Trajectory tracking control of underwater vehicle-manipulator system using discrete time delay estimation," *IEEE Access*, vol. 5, pp. 7435–7443, Jun. 2017.

[26] M. Myint, K. Yonemori, A. Yanou, M. Minami, and S. Ishiyama, "Visual-servo-based autonomous docking system for underwater vehicle using dual-eyes camera 3D-pose tracking," in *Proc. IEEE/SICE Int. Symp. System Integr. (SII)*, Nagoya, Japan, Dec. 2015, pp. 989–994.

[27] F. Bonin-Font, G. Oliver, S. Wirth, M. Massot, P. L. Negre, and J.-P. Beltran, "Visual sensing for autonomous underwater exploration and intervention tasks," *Ocean Eng.*, vol. 93, pp. 25–44, Jan. 2015.

[28] S. Guo, S. Pan, L. Shi, P. Guo, Y. He, and K. Tang, "Visual detection and tracking system for a spherical amphibious robot," *Sensors*, vol. 17, no. 4, p. 870, 2017, doi: 10.3390/s17040870.

[29] D. L. Rizzini, F. Kallasi, J. Aleotti, F. Oleari, and S. Caselli, "Integration of a stereo vision system into an autonomous underwater vehicle for pipe manipulation tasks," *Comput. Elect. Eng.*, vol. 58, pp. 560–571, Feb. 2017, doi: 10.1016/j.compeleceng.

[30] D. Fornas *et al.*, "Fitting primitive shapes in point clouds: A practical approach to improve autonomous underwater grasp specification of unknown objects," *J. Exp. Theor. Artif. Intell.*, vol. 28, nos. 1–2, pp. 369–384, 2016, doi: 10.1080/0952813X.2015.1046274.

[31] J. Gao, A. A. Proctor, Y. Shi, and C. Bradley, "Hierarchical model predictive image-based visual servoing of underwater vehicles with adaptive neural network dynamic control," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2323–2334, Oct. 2016.

[32] X. Liang, H. Wang, Y.-H. Liu, W. Chen, and J. Zhao, "A unified design method for adaptive visual tracking control of robots with eye-in-hand/fixed camera configuration," *Automatica*, vol. 59, pp. 97–105, Sep. 2015.

[33] Y.-H. Liu, H. Wang, C. Wang, and K. K. Lam, "Uncalibrated visual servoing of robots using a depth-independent interaction matrix," *IEEE Trans. Robot.*, vol. 22, no. 4, pp. 804–817, Aug. 2006.

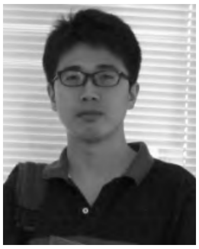
[34] K. Sedlaczek and P. Eberhard, "Using augmented Lagrangian particle swarm optimization for constrained problems in engineering," *Structural Multidisciplinary Optim.*, vol. 32, no. 4, pp. 277–286, 2006.

[35] M. Everingham, L. D. Fauw, K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) chanllenge," *IJCV*, vol. 6, no. 88, pp. 303–338, 2010.

LI JI-YONG received the B.S. degree from the Ship Building and Oceanic Engineering College Harbin Engineering University, Harbin, China, in 2015, where he is currently pursuing the Ph.D degree with the National Key Laboratory of Science and Technology of Underwater Vehicle His research interests include underwater robot and control.

ZHOU HAO received the B.S. degree from the Ship Building and Oceanic Engineering College Harbin Engineering University, Harbin, China, in 2016, where he is currently pursuing the master's degree with the National Key Laboratory of Science and Technology of Underwater Vehicle His research interests include underwater robot and visual recognition.

HUANG HAI received the B.S. and Ph.D. degrees in mechanical engineering from the Harbin Institute of Technology, Harbin, China, in 2001 and 2008, respectively. He is currently an Associate Professor and a Ph.D. Candidate Supervisor with the National Key Laboratory of Science and Technology of Underwater Vehicle, Harbin Engineering University, Harbin. His current research interests include underwater vehicle and autonomous operation.



YANG XU received the B.S. degree in automation from China Ocean University in 2009 and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences. His current research interests include underwater image processing and pattern recognition.

WAN ZHAOLIANG received the B.S. degree from the Ship Building and Oceanic Engineering College Harbin Engineering University, Harbin, China, where he is currently pursuing the master's degree with the National Key Laboratory of Science and Technology of Underwater Vehicle His research interests include underwater robot and target following.

WAN LEI is currently a Professor and a Ph.D. Candidate Supervisor with the National Key Laboratory of Science and Technology of Underwater Vehicle, Harbin Engineering University, Harbin, China. He has authored nearly 100 articles in underwater vehicles. His current research interests include underwater vehicle and autonomous control.