Deep learning techniques-based perfection of multi-sensor fusion oriented human-robot interaction system for identification of dense organisms

Haiju Li¹ | Chuntang Zhang¹ | Jingwen Bo¹ | Zhongjun Ding²

¹Institute of Automation and Electronics Engineering, Qingdao University of Science and Technology, Qingdao, China
²National Deep Sea Center, Qingdao, China

Correspondence
Chuntang Zhang, No. 99 Songling Rd, Qingdao, Shandong, PR. China, 260601.
Email: zcl1999@163.com

Abstract
For detection of dense small-target organisms with indistinct features in complex background, the efficiency and accuracy of traditional target detection methods are low. Multi-sensor fusion oriented human-robot interaction (HRI) system has facilitated biologists to process and analyse data. For this, several deep learning models based on convolutional neural network (CNN) are improved and compared to study the species and density of dense organisms in deep-sea hydrothermal vent, which are fused with related environmental information given by position sensors and conductivity-temperature-depth (CTD) sensors, so as to perfect multi-sensor fusion oriented HRI system. Firstly, the authors combined different meta-architectures and different feature extractors, and obtained five object identification algorithms based on CNN. Then, they compared computational cost of feature extractors and weighed the pros and cons of each algorithm from mean detection speed, correlation coefficient and mean class-specific confidence score to confirm that Faster Region-based CNN (R-CNN)_InceptionNet is the best algorithm applicable to hydrothermal vent biological dataset. Finally, they calculated the cognitive accuracy of riocaris exoculata in dense and sparse areas, which were 88.3% and 95.9% respectively, to analyse the performance of the Faster R-CNN_InceptionNet. Results show that the proposed method can be used in the multi-sensor fusion oriented HRI system for the statistics of dense organisms in complex environments.

1 | INTRODUCTION

Scientific investigations have shown that deep-sea hydrothermal vents gather organisms driven by chemical energy in organic compounds such as methane and hydrogen sulphide, rather than photosynthesis [1]. It is of great significance to study the species, density and distribution of dense organisms in deep-sea hydrothermal vents for carrying out the comparative studies of biodiversity, exploring new energy and mineral resources and maintaining the ecological balance. However, for one thing, the complexity of the deep-sea environment, the diversity of marine meteorology and the immaturity of deep-sea detection technology make it difficult and costly to obtain data. For another, a large number of deep-sea dense biological samples and gigabytes of high-resolution video data make it not easy for biologists to process and analyse data.

Human-Robot interaction (HRI) systems are widely used in the field of deep-sea exploration. Brantner et al. introduced the control architecture of Ocean One, which enables dexterous robot manipulation to be deployed to the deep sea through cooperation between humanoid robots and human pilots [2]. Yang et al. proposed a HRI system for real-time underwater recognition of diver gestures based on deep transfer learning to facilitate human-robot collaborative tasks [3]. Hu et al. introduced covariance matrix adaptive evolution strategies (CMA-ES) to improve the safety and adaptability of robots in performing complex movement tasks [4]. These researches provided ideas for reducing the cost of deep-sea exploration and facilitating biologists to process and analyse data. Therefore, this research considers to establish a multi-sensor fusion oriented HRI system to realize the dialog between biologists and the deep-sea world. Through this

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Cognitive Computation and Systems published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Shenzhen University.
system, biologists can easily grasp various information of the hydrothermal vent biological community on the remote interface.

At present, existing multi-sensor fusion oriented HRI system only displays the hydrothermal vent biological image and marine environment information given by position sensors and conductivity-temperature-depth (CTD) sensors, such as longitude and latitude, temperature, depth, salinity and so on. In the field of deep-sea exploration, it is of practical significance to further obtain the biological information such as species, density and distribution of dense organisms. Therefore, how to quickly and accurately identify the species of dense organisms and evaluate the density of dense organisms is the focus of this research.

The traditional methods to detect the species and density of dense organisms include direct observation, environmental deoxyribonucleic acid (DNA) technology, and method of capture itself tends to destroy marine living resources and the habitat of species [6].

With the boom of deep learning, its ideas gradually seep into all walks of life, such as face and expression identification [7], daily activities monitoring [8–10], target tracking [11]. In the field of marine life exploration, many scholars have also applied this idea to automatic fish classification [12] and catfish density estimation [13]. However, these researches are focussed on single-target organisms with distinct characteristics in a simple background. For research on identification and classification of organisms in deep-sea hydrothermal vents with low-characteristics and high-density, simple object identification methods cannot meet the required speed and accuracy.

Convolutional neural network (CNN) [14] is one of the highly-concerned algorithms in deep learning [15]. Compared with traditional object detection algorithm [16–18], those algorithms based on CNN have a stronger ability to extract features, and have higher speed and accuracy in processing complex scenes. There are two kinds of object detection algorithms based on CNN. One are region-based detectors, such as Region-based Convolutional Neural Network (R-CNN) [19], Fast R-CNN [20], Faster R-CNN [21], Region-based Fully Convolutional Network (R-FCN) [22], which have a low error rate and a low recognition omission rate. But its detection speed is slow, so it can not satisfy real-time detection scenes. The other are region-free detectors, such as You Only Look Once (YOLOv1) [23], YOLOv2 [24], YOLOv3 [25] and Single Shot Multi-box Detector (SSD) [26], which have fast recognition speed and can meet real-time performance. It can be applied to embedded mobile devices, but is slightly inferior to the former in accuracy. In these two kinds of convolution-based meta-architectures [27], the basic network, which is used to extract the high-level features of the input image, such as Visual Geometry Group Network (VGG16) [28], Residual Network (ResNet) [29], Networks for Mobile Vision Applications (MobileNet) [25] and Inception Network (InceptionNet) [26], is great important. Because the number of parameters and the types of layers directly affect the speed and accuracy of the detector.

Therefore, several deep learning models based on convolutional neural networks are improved and compared to study the species and density of dense organisms in deep-sea hydrothermal vents under complex backgrounds. In order to better serve biologists, based on the existing HRI system that fuses environment information given by position sensors and CTD sensors, biological information is further fused into the HRI interaction system. In this way, a multi-sensor fusion oriented human-robot interactive intelligent system for automatic identification of dense organisms is perfected. First, a hydrothermal vent biological dataset is established from scratch. Under TensorFlow framework, the parameters of the dataset are trained and adjusted to obtain a biological detection model with good speed and accuracy. Then the model is used in multi-sensor fusion oriented HRI system to identify biological species and count biomass. This research realizes large-scale, high-density, real-time detection of dense organisms, with low cost, simple operation, high working efficiency, and strong update ability.

2  |  METHODOLOGY

2.1  |  Meta-architectures

2.1.1  |  SSD

SSD utilises region proposals [23] in the regression-based detection process, to set up default bounding boxes [26] called anchors in Faster R-CNN, which improves the detection effect compared to YOLOv1 with hypothesised bounding boxes [23]. The SSD is improved on the basic network VGG16. Improvements include replacing fully connected layer 6 (FC6) and FC7 of VGG16 with convolution layer 6 (Conv6) and Conv7 respectively, and adding several auxiliary structures [28] behind the basic network. And convolutional layers have been added to obtain more feature maps, namely Con4_3, Con5, Conv8_2, Conv9_2, Conv10_2, Conv11_2, for detection. The core idea of the SSD is to take intensive sampling at different positions of the image with different scales and aspect ratios uniformly, then extract features with CNN to obtain confidence scores for classification and localization information for bounding-box regression.

2.1.2  |  Faster R-CNN

Faster R-CNN (see Figure 1) uses region proposal network (RPN) instead of the selective search method in Fast R-CNN [20] to extract prediction bounding boxes. RPN [14] is a fully convolutional network (FCN), which can simultaneously predict bounding boxes and conditional class probabilities on each
grid. First, faster R-CNN extracts the features of the whole image [21], and then inputs the feature values to RPN to obtain prediction bounding boxes. Then region of interest (RoI) pooling layer [20] is used to get fixed default bounding boxes. Finally, the default bounding boxes are classified and position regressed.

2.1.3 | R-FCN

To solve the problem of incomplete weight sharing and insufficient position sensitivity, R-FCN is proposed based on Faster R-CNN and FCN [22]. It uses FCN to extract features of the target object to generate feature maps. Through convolutions, these feature maps further generate region proposal network (RPN) and position-sensitive score maps. The RPN is used to generate a series of regions of interests (RoIs). The position-sensitive score network compares the feature maps generated by FCN with the RoIs output by RPN. Then, non-maximum suppression (NMS) is used to remove duplication and vote on mean class-specific confidence score to locate and classify the target objects.

2.2 | Feature extractor

2.2.1 | ResNet

The infrastructure of the ResNet model is the residual learning block [29]. The residual block, \( x_{i+1} \), is divided into direct mapping \( h(x_i) \), and residual mapping \( F(x_i, w_i) \), to extract features from shallow layer to deep layer [31]. The residual mapping is inserted between the basic convolutional network. \( 1 \times 1 \) convolutions are added to the direct mapping section when the residual blocks need to increase or decrease the dimension. Therefore, ResNet solves the problems of slow speed and easy gradient explosion or gradient disappearance in training process of depth CNN model.

2.2.2 | InceptionNet

In order to reduce computation and avoid overfitting, InceptionNet is proposed. A sparse network structure is designed, which can enhance the performance of neural network, and ensure the efficiency of computing resources as well [31]. Inception module is a vital infrastructure in InceptionNet that can be stacked multiple times (see Figure 2). It has a total of four branches, each of which uses a \( 1 \times 1 \) convolution kernel to integrate the information of different feature channels. It also includes \( 3 \times 3, 5 \times 5 \) convolution kernels and a max pooling layer to ensure the diversity of features.

2.2.3 | MobileNet

MobileNet is a lightweight and low-latency network, which can effectively enhance the real-time performance of object detection. To solve the problem of low computational efficiency and large number of parameters, depthwise separable convolution [30] and two hyper-parameters [30] are introduced in MobileNet. The depthwise separable convolutions factorise a standard convolution into a depthwise convolution and a \( 1 \times 1 \) pointwise convolution [30]. A single convolution kernel is applied to each input channel to obtain the depth of input channel. Then \( 1 \times 1 \) convolution kernel is used to calculate the depthwise convolution [30]. Assuming that \( M \) is the number of feature map input channels, \( N \) is the number of feature map output channels, and \( D_k \times D_k \) is the size of the convolution kernel. The ratio of the computation cost of the depthwise separable convolution and the standard convolution is \( 1/N + 1/D_k^2 \). Two hyper-parameters [30], width multiplier(\( \alpha \)) and resolution multiplier(\( \beta \)), are introduced. The width multiplier is used to reduce the number of channels at each layer, that is, the number of channels in the input layer changes from \( M \) to \( \alpha M \), and the number of
Channels in the output layer changes from $N$ to $aN$. The resolution multiplier is used to reduce the resolution of the images, acting on the feature map and reducing the size of the input and output feature map. It acts on the feature map and changes the size of the feature map from $D_F \times D_F$ to $\beta h D_F \times \beta h D_F$. 

**FIGURE 3** The structure of the perfect multi-sensor oriented human-robot interaction system

**FIGURE 4** Hydrothermal vent biological sample set. (a) Global map of the manned submersible sailing around deep-sea hydrothermal vent. (b) Enlarged map of black smoker of the deep-sea hydrothermal vent. (c) The creature with the largest proportion in the hydrothermal vent biological community: *rimicaris exoculata*. (d) The creature with a small proportion in the hydrothermal vent biological community: *actiniaria*. 
TABLE 1  The computational cost of feature extractors

| Feature extractor | Parameters(M) | FLOPs(G) |
|-------------------|---------------|----------|
| MobileNetV1       | 4.20          | 0.557    |
| InceptionNetV3    | 23.20         | 4.880    |
| ResNet101         | 44.55         | 7.260    |

Abbreviation: FLOPs, floating points of operations.

3  MULTI-SENSOR FUSION ORIENTED HRI SYSTEM

The deep learning technology based on CNN is applied to the HRI system with single function to perfect the multi-sensor fusion oriented HRI system for identification of dense organisms in deep-sea hydrothermal vents. Operators control underwater camera through console to obtain biological images. Then these images are input into the trained model for detection. At the same time, multi-sensor device is equipped with a position sensor and a CTD sensor to obtain corresponding environmental information. The structure of the perfect multi-sensor oriented HRI system is shown in Figure 3.

The position sensor gives latitude and longitude information and biological coordinate information. The CTD sensor gives information about temperature, salinity, and depth of the deep-sea hydrothermal vent. The species, quantity, mean confidence score and mean detection speed of target organisms are given by deep learning techniques-based image identification system. Using serial communication, these external information are input to the computer through the serial interface of the universal asynchronous receiver/transmitter (UART) of the single-chip microcomputer. The decentralised Kalman filtering technology is used to realize the fusion of multi-sensor information. And these information are displayed on the display interface of the computer for biologists to process and analyse the data.

4  EXPERIMENTS

This paper combines different meta-architectures [19~26] and different feature extractors[28~31] to cf. and improve several kinds of deep learning models. The experimental steps are as follows: 1) Data preparation: obtaining and preprocessing source data and establishing a hydrothermal vent biological data set from scratch; 2) Model training: training five the hydrothermal vent biological detection model through Object Detection API; 3) Object detection: the images in the test set are input into five exported models for object detection; 4) Algorithm comparison: five typical algorithms based on the CNN, namely Faster R-CNN_InceptionNet, Faster R-CNN_ResNet, R-FCN_ResNet, SSD-MobileNet and SSD_InceptionNet are compared in terms of mean detection speed, correlation coefficient and mean class-specific confidence score. And it is confirmed that Faster R-CNN_InceptionNet is more suitable for the hydrothermal vent biological dataset; 5) Performance analyzes of Faster R-CNN_InceptionNet: 10 biological pictures in dense areas and ten biological pictures in sparse areas are selected to calculate the cognitive accuracy of dense *rimicaris exoculata*. For this, the performance of Faster R-CNN_InceptionNet for the detection of dense small-target organisms with indistinct features in complex background is analysed.

4.1  Data preparation

The manned submersible, a kind of underwater operation device with comprehensive detection capability, navigated slowly around the deep-sea hydrothermal vent (see Figure 4a,b) at a temperature of about 275°C and photographed a large number of videos. OpenCV was used to convert video into images at the frame rate of one frame per second. A total of 14,486 images were obtained, all of which were used as the data source of this paper. Most of the organisms in images are *rimicaris exoculata* (see Figure 4c), and a few others, such as *actiniaria* (see Figure 4d).

All the images were divided into training set, verification set and test set according to the ratio of 6.4:1.6:2. The images in training set and verification set were used for model training. The images in the test set were used to test the performance of the model. *Rimicaris exoculata*, *actiniaria* and other organisms were labelled in these images in training set and verification set, then the labelled images were converted to XML format files. The key to measuring the quality of the model is to label all morphological features of organisms. The XML format file need to be converted to CSV format file, and then to the TFRecord format file in order to make the hydrothermal vent biological dataset. The CSV format file records the names of the images, the species of the organisms, and the coordinates of the ground truth boxes.

4.2  Model training

The training of the model is to pre-learn the morphological features of organisms, and correct the weights of the feature vectors as well as. The training process is as follows: (1) Propagating the original images forward in the convolutional network to obtain the basic image features; (2) Extracting multi-scale feature maps and selecting default boxes with different aspect ratios at various positions in these map; (3) Computing coordinate offset and class-special confidence
score of each default box; (4) computing the total loss and (5) Back-propagating the total loss and adjusting the weight of each network layer.

It is difficult to measure the quality of a detection algorithm, because in addition to the structure of the algorithm itself, there are many other factors that affect its speed and accuracy. For example, different parameter settings, different definition methods of loss function and different feature extraction networks, such as, ResNet [29], MobileNet [30], InceptionNet [31] etc. The parameters and floating points of operations (FLOPs) of feature extractors used in this paper are shown in Table 1. It can be seen that MobileNetV3 is a lightweight model with small parameters and FLOPs, InceptionNetV3 is in the middle of these two aspects, while ResNet101 has the highest space complexity and time complexity.

Therefore, in order to further compare the advantages of different algorithms, a unified implementation for training model and biological detection was proposed in the TensorFlow framework, and the same parameters such as training iteration steps, learning rate, batch size were set. In addition, the loss function includes softmax loss for classification and smooth L1 loss for bounding-box regression. The total loss is the weighted sum of the errors of classification and regression. Softmax loss, a loss function for multiple classifications, is composed of softmax layer and cross-entropy loss. The softmax layer maps the output of each node to the interval (0, 1), forming the probability distribution with sum of 1. Cross entropy is used to determine how close the actual output is to the expected output. On the one hand, smooth L1 loss can prevent excessive gradient value from damaging the network parameters when the difference between the predicted box and the truth box is too large in the early stage of training. On the other hand, smooth L1 loss can ensure that the gradient value is small enough to continue to converge when the difference between the predicted box and the truth box is small in the later stage of training.

The specific experimental steps are as follows: first, a configuration file, label_map, was prepared, which recorded the names of three types of target objects to be identified, namely *rimicaris exoculata*, *actiniaria* and complex background. Then, combined with VOC dataset, the hydrothermal vent biological detection model was trained through the Object Detection API, which was an object recognition system based on the TensorFlow framework launched by Google. The learning rate was set to 0.004 and the decay rate was set to 0.95 for 200,000 iterations [26]. The maximum size of the default box was set to 0.95 and the minimum size was 0.35. And, It was of great importance to set the path of the TFRecord format file of the dataset and the path of the label_map file. Finally, a new directory, train_dir, was created to save models and logs. And the train.py file under the legacy folder was run to start training model.

### 4.3 Object detection and algorithm comparison

After the training, the five models were exported. 1515 images of the test set were input into these exported models for object detection, then the predicted species and number of target organisms were obtained. In addition, the authors manually counted the true number of organisms in the test set for comparative analysis.

These algorithms were compared from the aspects of mean detection speed, correlation coefficient and mean class-specific confidence score. Correlation coefficient refers to the degree of linear correlation between the predicted number set and the real number set. The mean class-specific confidence score is the average value of the product of the conditional class probabilities and the intersection over union (iou) [23]:

\[
\frac{1}{N} \sum_{j=0}^{N} \Pr(Class_i|Object) \times \Pr(Object) \times \text{IOU}_{pred}^{truth} \\
= \frac{1}{N} \sum_{j=0}^{N} \Pr(Class_i) \times \text{IOU}_{truth}^{pred}
\]

Where, \(N\) represents the number of class \(i\) appearing in the predicted box. \(\Pr(Class_i|Object)\) represents the possibility of being class \(i\) on the premise that a object organism is detected. The class-specific confidence score is normalized between 0 and 1, and the threshold is set as 0.5. \(\text{IOU}_{truth}^{pred}\) represents the relation between the predicted box and the ground truth box:

| Model                  | Mean detection speed (s/per image) | Correlation coefficients of *rimicaris exoculata* (>0.9000) | Confidence score of *rimicaris exoculata* (>0.5000) | Confidence score of *actiniaria* (>0.5000) |
|------------------------|-----------------------------------|-------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------|
| Faster R-CNN_InceptionNet | 3.0550                            | 0.9563                                                      | 0.9089                                              | 0.9505                                       |
| Faster R-CNN_ResNet     | 10.2218                           | 0.9630                                                      | 0.9126                                              | 0.9231                                       |
| R-FCN_ResNet            | 13.9801                           | 0.9723                                                      | 0.8763                                              | 0.8781                                       |
| SSD_InceptionNet        | 3.1506                            | 0.9025                                                      | 0.9187                                              | 0.8827                                       |
| SSD_MobileNet           | 1.7887                            | 0.9139                                                      | 0.8964                                              | 0.9039                                       |

Abbreviations: R-CNN, region-based convolutional neural network; SSD, single shot multi-box detector.
\[ \text{ROI}_{\text{pred}} = \frac{\text{Area of Overlap}_{\text{pred}}}{\text{Area of Union}_{\text{pred}}} \]  

Where, the numerator represents the overlap area of the prediction box and the ground truth box, and the denominator represents the union area of them (see Figure 5).

In Table 2, In terms of mean detection speed, SSD_MobileNet is 1.7887 s/per image, which is the fastest, R-FCN_ResNet is 13.9801 s/per image, which is the slowest, and Faster R-CNN_InceptionNet and SSD_InceptionNet are around 3 s/per image. The mean class-specific confidence scores of each algorithm is greater than 0.85, and the Faster R-CNN_InceptionNet and the Faster R-CNN_ResNet both have the best performance of greater than 0.9. As for the correlation coefficient of *rimicaris exoculata*, the region-based detector is about 0.9, which was lower overall than the region-free detector, and the region-free detectors is all above 0.95.

In Figure 6, the total losses of five model have gradually decreased and tended to a stable value for iterations. It is well known that the total loss in the training process is inversely proportional to the robustness of the model, that is, the smaller the total loss, the better the robustness. The convergence value of the total loss of the region-free detectors is between 2.5 and 3. The total loss of region-based detectors converge below 0.7, and Faster R-CNN_ResNet has the best convergence.
The region-based detector has a high precision and a low speed, while the region-free detector has a faster cognitive speed, but its precision is slightly lower than the former. The Faster R-CNN_InceptionNet, which has good performance in both precision and speed, is determined as the best algorithm applicable to the hydrothermal vent biological dataset.

4.4 | Performance analyzes of faster R-CNN_InceptionNet

After comparing the detection effect of the five algorithms from different aspects, the performance of the Faster R-CNN_InceptionNet model need to be analysed. 10 biological pictures in dense areas and 10 biological pictures in sparse areas were used for testing to compute the recognition accuracy of *rimicaris exoculata*. The formula of accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where, $TP$ means True Positive where a *rimicaris exoculata* is predicted as a *rimicaris exoculata*. $TN$ means True Negative where a *actiniaria* is predicted as a *actiniaria*. $FP$ means False Positive where a *rimicaris exoculata* is predicted as a *actiniaria*. $FN$ means False Negative where a *actiniaria* is predicted as a *rimicaris exoculata*. 

**FIGURE 7** Recognition effects of the faster region-based convolutional neural network_InceptionNet model. (a) and (c) Display diagram of the recognition effect of *actiniaria* in the sparse area. (b) Display diagram of the recognition effect of *rimicaris exoculata* in the sparse area. (d) Display diagram of the recognition effect of *rimicaris exoculata* in dense area.
In Table 3 and Table 4, the cognitive accuracy of *rimicaris exoculata* in sparse areas is 95.9%, and that is 88.3% in dense areas, the former is 7.6% higher than the latter. The *rimicaris exoculata* in the sparse area are scattered independently in the hydrothermal vent, and their contour is clear, as shown in Figure 7b. The *rimicaris exoculata* in dense areas are almost all stacked together, and their contour is more complex and fuzzy, as shown in Figure 7d, which results in lower recognition rate and generalisation ability [19]. In Figure 7, the prediction box accurately circles the object organisms, and the class-specific confidence scores are all >0.8, which means the localization error is small. With the exception of a few organisms with no obvious morphological features, these organisms are accurately detected without false background detection.

5 | CONCLUSIONS

In this paper, by comparing and improving the five deep learning model based on CNN, it is determined that Faster R-CNN_InceptionNet is more suitable for the detection of dense small-target organisms with indistinct features in complex background. Experiments show that the proposed method can automatically detect the species and quantity of dense organisms with higher speed and accuracy. And it is feasible and of realistic value that the improved multi-sensor fusion oriented HRI system is used to help biologists analyse and maintain the ecological balance of deep-sea hydrothermal vents. In terms of the identification and classification of dense small-targets, further research directions include but are not limited to: 1) On the premise of ensuring the detection speed and accuracy, reducing the amount of calculations and parameters to compress the size of the model; 2) Using convolutional neural depth network on a large scale or fusing context features with biological features to enhance feature expression capability, so as to improve the detection speed and accuracy of dense small-targets; 3) Transfer learning and video tracking of small-target objects.

ACKNOWLEDGMENT

The authors thank the National Deep Sea Centre, Qingdao, China for providing us the necessary deep-sea biological data and the significant supporting facilities to carry out this study.

ORCID

Chuntang Zhang https://orcid.org/0000-0001-7553-0653

REFERENCES

1. Ooka, H., McGlynn, S.E., Nakamura, R.: Electrochemistry at deep-sea hydrothermal vents: utilization of the thermodynamic driving force towards the autotrophic origin of life. Chem. Electro. Chem. 6(5), 1316–1323 (2019).
2. Brantner, G., Khatib, O.: Controlling ocean one: human-robot collaboration for deep-sea manipulation. J. Field Robot. 4, 28–51 (2020).
3. Xia, Y., Sattar, J.: Visual diver recognition for underwater human-robot collaboration. In: International Conference on Robotics and Automation. Department of Computer Science and Engineering, The University of Minnesota, Minneapolis. arXiv:1809.10201, (2018) https://arxiv.org/abs/1809.10201.
4. Hu, Y., et al.: Evolution strategies learning with variable impedance control for grasping under uncertainty. IEEE Trans. Ind. Electron. 66(10), 7788–7799 (2018).
5. Lauren, M.S., et al.: Quantification of environmental DNA (eDNA) shedding and decay rates for three marine fish. Environ. Sci. Technol. 50(19), 10456–10464 (2016).
6. Schramm, K.D., et al.: A comparison of stereo-BRUVs and stereo-ROV techniques for sampling shallow water fish communities on and off pipelines. Mar. Environ. Res. 162, 105198 (2020).
7. Zhou, J., et al.: Improved softmax loss for deep learning based face and expression recognition. Cognit. Comput. Syst. 1(4), 97–102 (2019).
8. Qi, W., Andrea, A.: A multimodal wearable system for continuous and real-time breathing pattern monitoring during daily activity. IEEE J. Biomed. Health Inform. 24(8), 2199–2207 (2020).
9. Qi, W., Su, H., Aliverti, A.: A smartphone-based adaptive recognition and real-time monitoring system for human activities. IEEE Trans. Hum. Mach. Syst. 50(5), 414–423 (2020).
10. Qi, W., Su, H., Yang, C.: A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone. Sensors. 19(17), 3731 (2019).
11. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: siamese CNN for robust target association. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, pp. 418–425 (2016).
12. Chhabra, H.S., Srivastava, A.K., Nijhawan, R.: A hybrid deep learning approach for automatic fish classification. Emerg. Trends Inform. Technol. 605, 427–436 (2020).
13. Donatello, C., Pierre, G., Carlo, S.: Catfish density estimation by aerial images analysis and deep learning. In: Proceedings of the ACM Symposium on Applied Computing, pp. 1111–1114, New York, US (2019). Part F147772.
14. Bahrani, K., et al.: Convolutional neural network for reconstruction of 77’ like images from 3 T MRI using appearance and anatomical features. In: Proceedings of the First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016. Athens (2016).
15. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521(7553), 436–444 (2015).
16. Nagasato, D., et al.: Deep neural network-based method for detecting central retinal vein occlusion using ultrawide-field fundus ophthalmoscopy. J. Ophthalmol. 2018, 1–6 (2018).
17. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2(2), 121–167 (1998).
18. Belenky, A., Ansari, N.: On deterministic packet marking. Comput. Network. 51(10), 2677–2700 (2007).
19. Lenc, K., Vedaldi, A.: ‘RCNN minus R’, in arXiv:1506.06981, (2015). https://arxiv.org/abs/1506.06981.
20. R. Girshick, “Fast R-CNN,” 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, Santiago, Chile, (2015) https://doi.org/10.1109/ICCV.2015.169.
21. Ren, S., et al.: ‘Faster R-CNN: towards real-time object detection with region proposal networks’. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017).
22. Dai, J., et al.: R-FCN: object detection via region-based fully convolutional networks. In: D.L. Daniel, U. von Luxburg, R. Garnett, & et al. (Eds.), Advances in Neural Information Processing Systems, pp. 379–387. Curran Associates, New York (2016).
23. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. Las Vegas, NV, USA, (2016) https://doi.org/10.1109/CVPR.2016.91.
24. J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525. Honolulu, HI, (2017) https://doi.org/10.1109/cvpr.2017.690.
25. Redmon, J., Farhadi, A.: ‘YOLOv3: An incremental improvement’, in arXiv:1804.02767, 2018, [online] Available: https://arxiv.org/abs/1804.02767

26. Liu, W., et al.: SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol. 9905. Springer, Cham (2016) https://doi.org/10.1007/978-3-319-46448-0_2

27. Huang, J., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu (2017)

28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for largescale image recognition. In: International Conference on Learning Representations. San Diego (2015)

29. He, K., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, pp. 770–778 (2016)

30. Howard, A.G., Zhu, M., Chen, B., et al.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, in arXiv:1704.04861, (2017), https://arxiv.org/abs/1704.04861

31. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston, pp. 1–9 (2015)

How to cite this article: Li H, Zhang C, Bo J, Ding Z. Deep learning techniques-based perfection of multisensor fusion oriented human-robot interaction system for identification of dense organisms. Cogn. Comput. Syst. 2021;1–10. https://doi.org/10.1049/ccs2.12010