INTERNATIONAL JOURNAL OF ADVANCED ROBOTIC SYSTEMS

International Journal of Advanced Robotic Systems
May-June 2020: 1–11

DOI: 10.1177/1729881420932715
journals.sagepub.com/home/arx
SAGE

A modified YOLOv3 detection method for vision-based water surface garbage capture robot

Xiali Li[1], Manjun Tian[1], Shihan Kong[2,3], Licheng Wu[1] and Junzhi Yu[2,4]

Abstract

To tackle the water surface pollution problem, a vision-based water surface garbage capture robot has been developed in our lab. In this article, we present a modified you only look once v3-based garbage detection method, allowing real-time and high-precision object detection in dynamic aquatic environments. More specifically, to improve the real-time detection performance, the detection scales of you only look once v3 are simplified from 3 to 2. Besides, to guarantee the accuracy of detection, the anchor boxes of our training data set are reclustered for replacing some of the original you only look once v3 prior anchor boxes that are not appropriate to our data set. By virtue of the proposed detection method, the capture robot has the capability of cleaning floating garbage in the field. Experimental results demonstrate that both detection speed and accuracy of the modified you only look once v3 are better than those of other object detection algorithms. The obtained results provide valuable insight into the high-speed detection and grasping of dynamic objects in complex aquatic environments autonomously and intelligently.

Keywords

Object detection, aquatic environment, garbage capture robot, modified YOLOv3, detection

Date received: 25 March 2020; accepted: 17 May 2020

Topic: Vision Systems Special Collection: Underwater Image Processing and Target Recognition
Topic Editor: Antonio Fernandez- Caballero
Associate Editor: Zhenxue Chen

Introduction

Water insecurity (quantity and quality) affects health and livelihoods. Contaminated water causes 1.7 million fatalities from treatable diseases annually. Wetlands conservatively valued at up to USD 800,000 ha/year are being lost rapidly. Poor ocean health affects access to cheap protein (i.e. fish) for approximately 275 million people and the access to 20% of the animal protein supply for 3.1 billion people in total. Meanwhile, it threatens livelihoods in the tourism and commercial fishing sectors.[1] The reserves of water resources are not only very limited but also faced with severe pollution, especially as plastic waste pollution. Therefore, the protection of water resources is one of the greatest and most sacred duties of human beings. With the

[1] School of Information Engineering, Minzu University of China, Beijing, China
[2] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4] State Key Laboratory for Turbulence and Complex Systems, Department of Mechanics and Engineering Science, Beijing Innovation Center for Engineering Science and Advanced Technology, College of Engineering, Peking University, Beijing, China

Corresponding author:
Junzhi Yu, State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; State Key Laboratory for Turbulence and Complex Systems, Department of Mechanics and Engineering Science, Beijing Innovation Center for Engineering Science and Advanced Technology, College of Engineering, Peking University, Beijing 100871, China.
Email: junzhi.yu@ia.ac.cn

development of science and technology, the robots for water surface garbage cleaning are emerged as new tools.

The researches about cleaning robot are gradually becoming prevalent,[2–9] and performances in the grasping are fascinating.[10–13] In this context, a water surface garbage cleaner robot has been developed in our lab, which can be applied to clean up water surface garbage, reduce the labor volume of sanitation workers, and improve the water ecological environment.[14] There are three primary tasks for the garbage capture robot, that is, garbage detection, garbage capture, and garbage collection. Notably, garbage detection is particularly significant in the above three parts, because it is in charge of providing reliable object location information for the robot. Besides, garbage detection is the prerequisite for the successful implementation of garbage capture and garbage collection. Consequently, an efficient detection method is highly demanded.

Object detection is not difficult for the human eye, because people can locate the object by perceiving its color, texture, and edges of the image easily. However, it is complicated for computers. With respect to the traditional object detection methods, computers execute object detection according to feature extraction like human beings. Among these methods, Haar[15] has been widely used due to both its fast feature extraction speed and ability to express the edge information of objects. Local binary pattern (LBP)[16] can better express texture information of objects. Histogram of oriented gradient (HOG)[17] uses histogram to count the edges of objects, which has satisfactory performance to express features and is widely applied. In addition, a grayscale-based detection method is applied.[18] But the aforementioned methods are based on some manual-designed features, whose shortcomings are detailed as follows. First, designing features manually is particularly troublesome. Second, there are great limitations, for example, application scenarios. Third, extracted features may be insufficient and incomplete. Fortunately, the convolutional neural network (CNN) is utilized for object detection, where all of above shortcomings have been overcome.[19–21]

With the improvement of computing power, the development of graphics processing unit processor, and the maturity of big data technology, the deep learning technology has achieved considerable performance in object detection in different fields.[22–26] For example, in the field of medicine, based on deep learning, a technique using mask regions with convolutional neural network (R-CNN) was developed for lesion detection and differentiation between benign and malignant.[27] Nowadays, the prevalent object detection methodologies are mainly divided into two categories. One is the two-stage network represented by R-CNN,[28] Fast R-CNN,[29] and Faster R-CNN,[30] all of which include a region proposal network (RPN) to get region proposals. The other is the one-stage network represented by you only look once v1 (YOLOv1),[31] YOLOv2,[32] YOLOv3,[33] and single shot detector (SSD),[34] which transform classification and detection problems into regression problems. Especially, the comprehensive performance of YOLOv3 in detection speed and accuracy is very prominent, which can achieve 57.9 average precision (AP_{50}) in 50 ms on a NVIDIA Titan X processor.[33] Hence this article employs the YOLOv3 algorithm and modifies it to detect three kinds of water surface garbage, including plastic bottles, plastic bags, and styrofoam.

The primary contributions of this article are as follows:

1. A modified YOLOv3 network is proposed to detect floating garbage. Due to that, there are three types of garbage during detection; the three-scale prediction of YOLOv3 is changed to a two-scale prediction. The two-scale can achieve 54.04 frames/s on the GeForce technology eXtreme (GTX) 1080, which guarantees the real timeness.
2. The anchor boxes of the original YOLOv3 are obtained by utilizing K-means clustering in the common object in context (COCO) data set, which is exactly appropriate to the COCO data set, but improper for our data set. Thus, all the boxes in the water surface garbage data set are reclustered to replace the original anchor boxes. The modified anchor boxes YOLOv3 can reach 91.43 mean average precision (mAP) in our data set, which compensates for the detection accuracy.

The rest of this article is organized as follows. After reviewing the related work, we will elaborate on the proposed detection method. Then, the experimental results are presented, and finally, we conclude the article.

Related works

Object detection is one of the classical problems in computer vision. Its task is to mark the position of the object in the image and to label the category of the object. Originating from the traditional artificial design feature and shallow classifier framework to the end-to-end deep learning detection framework, object detection algorithms have gradually become maturer and maturer.

There are numerous methods for object detection. With respect to the conventional object detection algorithms, a region in the image is selected by sliding window, which will be regarded as a candidate region. Then one or more features, such as Harr, HOG, LBP, or local ternary pattern, are extracted from the candidate regions. Finally, the candidate regions are classified by related classification algorithms such as Adaboost[35] and support vector machine.[36] Noticeably, in 2001, Viola's cascade + Harr scheme[37] achieved outstanding results in face detection. Furthermore, deformable part module (DPM) is a popular competitive detector. A fast category object detector based on DPM was presented[38] that divides the object into root model and part model, in which the root model is

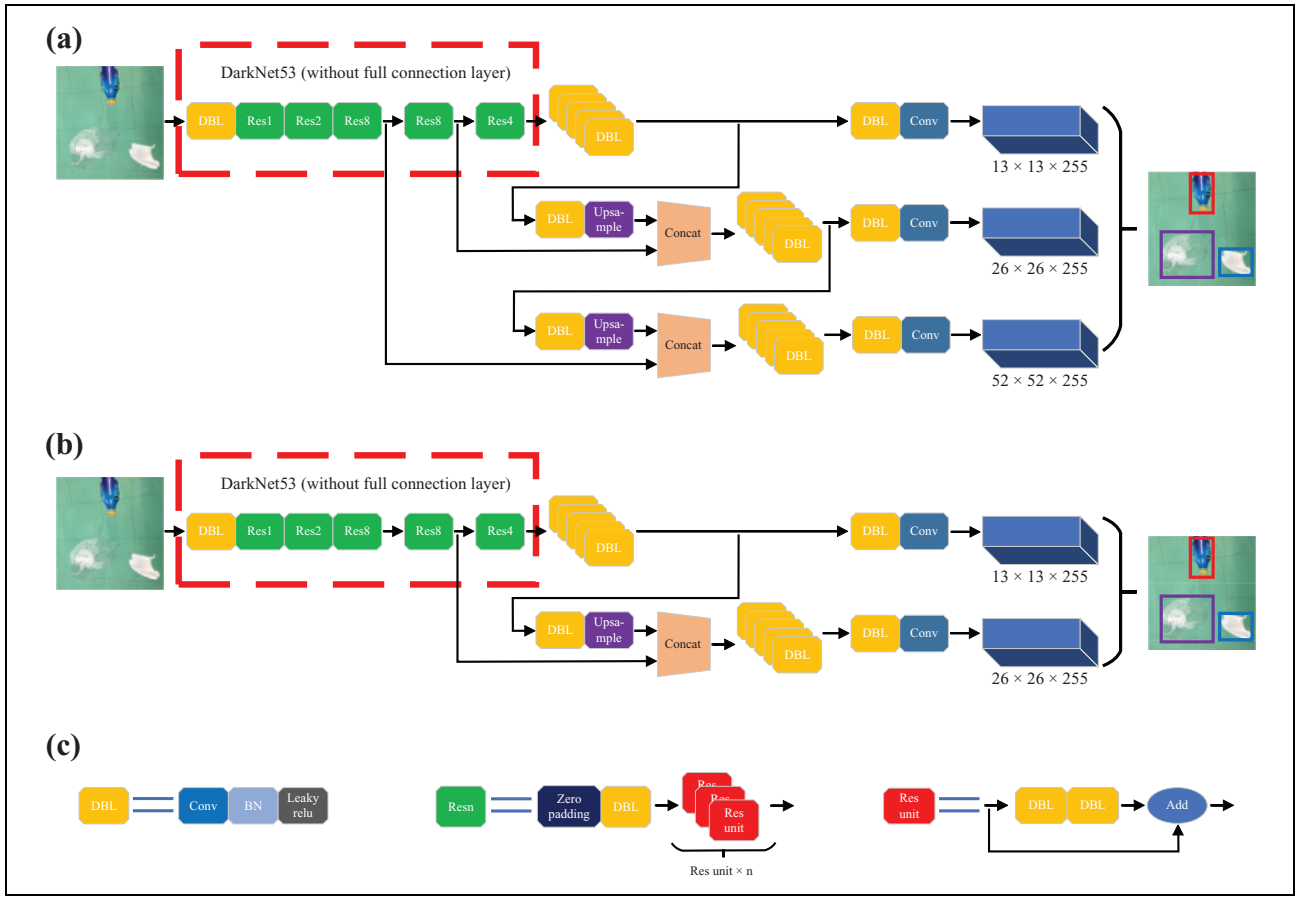

Figure 1. Network architecture. (a) The structure of the original three-scale YOLOv3. (b) The structure of two-scale YOLOv3. (c) The legends for (a) and (b). YOLOv3: you only look once v3.

equivalent to the traditional HOG feature and the part model is the template of some parts of the object. During the detection, the root model is used to locate the possible position of the object and the part model is used for further confirmation. However, the performance of DPM is rather ordinary, which cannot be adapted to the images with sharp rotations. Thus, its stability and robustness are undesirable. Additionally, the workload is relatively heavy.

Traditional methods can no longer meet the needs of production and life, for which the module of RPN and deep learning classification was put forward. A new pooling method "Sspatial pyramid pooling network" was proposed in He et al.,[39] which can produce fixed-length output despite the input image with any size, as well as the network has good robustness to object deformation. Meanwhile, the speed and accuracy have been improved unprecedentedly, whereas the primary downside is that the training process is very complex and needs a lot of storage space. R-CNN-based networks[28–30] have capability to the accurate detection, but the computational burden from RPN reduces the speed dramatically.

The RPN is not simple; as a result, the regression method based on end-to-end deep learning was born for object detection. YOLOv3 can detect objects according to three scales as shown in Figure 1(a). Its feature extraction network comes from Darknet-53, which adopts the full convolution structure and takes advantage of the residual structure as illustrated in Figure 1(c). Furthermore, it uses the strategy of multiscale fusion, which greatly improves the recall rate and accuracy of YOLOv3. By utilizing the deep and shallow features through the route layer, YOLOv3 achieves unprecedented high accuracy and speed. In particular, YOLOv3 was compared with Faster R-CNN and SSD, respectively[40,41]; the results show that YOLOv3 is in a leading position in both speed and accuracy.

Water surface garbage capture robot

Traditional water decontamination mainly relies on manual salvage operation, which is not only difficult, risky, and inefficient but also exists some limitations. If cleaning missions are conducted on the small rivers, man-made lakes, and water amusement parks, where ships cannot voyage due to narrow areas and shallow water, dustmen will be hindered. At the same time, for some water areas with serious chemical pollution, it is not suitable for workers to clean trash. In addition, for some landscape lakes and rivers, artificial cleaning will affect the beauty of the

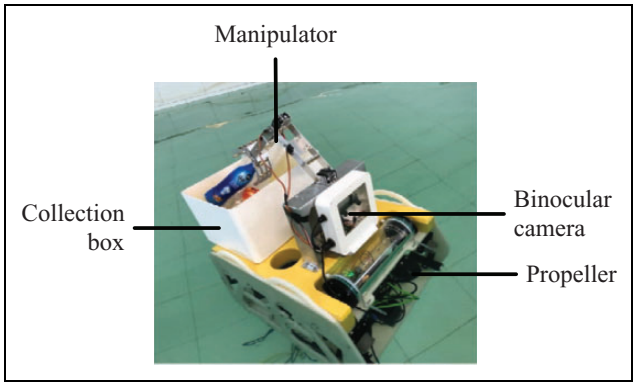

Figure 2. Configuration of the developed vision-based water surface garbage capture robot.

surrounding environment. In this context, we have designed a floating garbage capture robot system, which can identify and search the surface garbage independently according to the visual system, then accurately locate it, and finally approach and capture it.

The structure of the robot is shown in Figure 2. The robot is about 64 cm long, 49 cm wide, 47 cm high, and 25 kg weight. It is made of light and solid materials, so that it can voyage on the water surface. A binocular camera is located on the device to detect and track the object, which can provide the robot with the position information of the target. The robot can dynamically adjust the position and pose according to the grasping objects position information in real time. Once the target is locked, robot will gradually swim toward the object actuated by four propellers symmetrically mounted at the bottom. A three degrees of freedom robot manipulator is placed in front of the device, whose length is approximately 48 cm. Note that the joints can rotate from 0° to 180°, which can promote the robot to grasp and collect garbage flexibly. A replaceable garbage collection box is placed at the rear of the robot for temporarily storing. Once the box is full, it will be replaced in time.

Water surface garbage detection

In this cleaning mission, the detected objects include floating plastic bottles, plastic bags, and styrofoam. The modified YOLOv3 network is used to realize the detection. The YOLOv3 can detect and classify multiclass objects from 3 scales and 9 ranges, and its main network architecture comes from the first 52 layers of the DarkNet-53 network, which can fully extract features from images. Note that Figure 1(a) describes its architecture, where three-scale feature map will be output. However, our data are completely different from the COCO data set; moreover, the water surface situation is changeable for the actual application scenario. This characteristic results in the high real-time requirements for detection. Motivated by this demand, we transformed YOLOv3 from three-scale detection to two-scale detection by removing the last scale, which better ensures real timeness. The modified YOLOv3 network structure is shown in Figure 1(b).

YOLOv3 will scale the input image with any size to 416 × 416 pixels firstly; after one image passing through the network, YOLOv3 will output feature maps in three sizes of this image, including $13 \times 13 \times (4 + 1 + N) \times 3$, $26 \times 26 \times (4 + 1 + N) \times 3$, and $52 \times 52 \times (4 + 1 + N) \times 3$.

1. 13, 26, and 52 are the side length of the feature map; in other words, images are divided into 13 × 13, 26 × 26, and 52 × 52 cells after they go through the network.
2. $4 + 1$ represents the parameters of each predicted box $(x, y, w, h, \text{confidence})$. Note that (x, y) and (w, h) represent the center coordinate and the width and height of the box respectively, and confidence indicates the possibility of the object in this box. If there is an object in the box, the confidence value will be close to 1, otherwise it will be close to 0. As well as, N is the number of probability values. The output result will output a probability belonging to each category for each predicted object, by which the category of the object to be detected is the largest one among all probabilities. Generally, N depends on the number of object categories; in this article, N equals 3.
3. 3 indicates that there will be three sizes of bounding boxes to predict objects, while these bounding boxes are clustered from data set object boxes.

Furthermore, the loss function of YOLOv3 is defined below

$$\begin{aligned}
\mathcal{L} &= \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{obj}} [(x_i^j - \hat{x}_i^j)^2 + (y_i^j - \hat{y}_i^j)^2] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{obj}} \left[\left(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j} \right)^2 + \left(\sqrt{h_i^j} \sqrt{\hat{h}_i^j} \right)^2 \right] \\
&- \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{obj}} [\hat{c}_i^j \log(c_i^j) + (1 - \hat{c}_i^j) \log(1 - c_i^j)] \\
&- \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{noobj}} [\hat{c}_i^j \log(c_i^j) + (1 - \hat{c}_i^j) \log(1 - c_i^j)] \\
&- \sum_{i=0}^{s^2} I_{ij}^{\text{obj}} \sum_{c \in \text{class}} [\hat{p}_i^j(c) \log(p_i^j(c)) \\
&+ (1 - \hat{p}_i^j(c)) \log(1 - p_i^j(c))] \qquad (1)
\end{aligned}$$

where λ_{coord} indicates the weight of coordinate error and λ_{noobj} denotes the weight of intersection over union (IoU) error. In this experiment, λ_{coord} equals 5 and λ_{noobj} is 0.5. Next, s^2 is the number of cells ($s \times s$) of feature map.

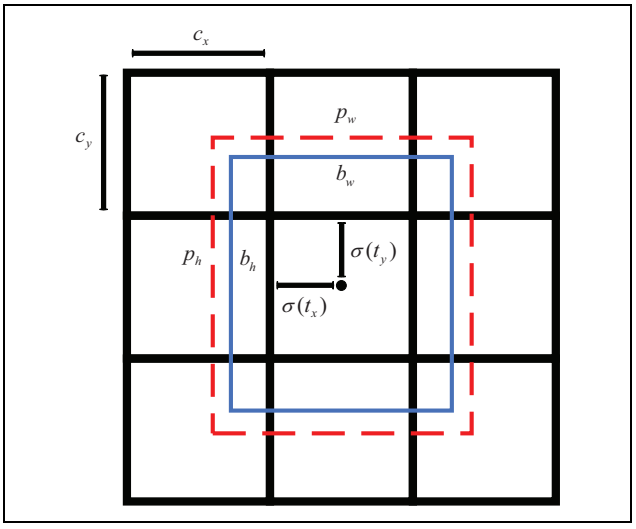

Figure 3. Prediction of bounding box. YOLOv3 will predict the width and height of the box as offsets from cluster centroids and the center coordinates of the box relative to the location of filter application using a sigmoid function. YOLOv3: you only look once v3.

Additionally, l_{ij}^{obj} is used to determine whether the jth bounding box of the i-th cell is responsible for detecting this object. If the IoU of this bounding box and ground truth is the largest, l_{ij}^{obj} is 1, otherwise, it will be 0. Similarly, l_{ij}^{noobj} means that the jth bounding box of the ith cell is not responsible for the target. Note that c_i^j is the confidence of predicting objects. Additionally, $p_i^j(c)$ denotes the probability of the class c object in the ith cell. Because of the use of binary cross-entropy loss and logistic regression for category prediction, this choice helps YOLOv3 to be applied in more complex areas and detect objects more accurately and effectively. Furthermore, it is possible to classify multiple labels on the same object as well.

YOLOv3 will predict four coordinate values (b_x, b_y, b_w, b_h) for each bounding box ultimately. The bounding boxes prediction is shown in Figure 3. The dashed line rectangular box in the graph is the preset boundary box, and the solid line rectangular box is the predicted boundary box obtained by the network prediction. The bounding box is predicted as follows

$$b_x = \sigma(t_x) + c_x \tag{2}$$

$$b_y = \sigma(t_y) + c_y \tag{3}$$

$$b_w = p_w e^{t_w} \tag{4}$$

$$b_h = p_h e^{t_h} \tag{5}$$

The actual prediction value of the network is (t_x, t_y, t_w, t_h). According to (2)–(5), the coordinates of the center point and the width height (b_x, b_y, b_w, b_h) of the predicted box can be calculated. Meanwhile, (c_x, c_y) is the offset between current cell and upper left corner cell in (2) and (3). Note that (p_w, p_h) is the width and height of the predicted boundary box on the feature map. In the aforementioned formulas, $\sigma(x)$ is a sigmoid function, whose purpose is to uniform the predicted values within the interval [0,1], to accelerate the convergence of the network.

As for obtaining the initial sizes of bounding boxes in YOLOv3, the K-means clustering method in YOLOv2 is still adopted. Clustering is the process of gathering similar things together and dividing them into different categories, which is a very important method in data analysis. Among many clustering methods, K-means is the most commonly used iterative clustering algorithm,[42–44] whose idea is to randomly select K objects as the initial clustering center, then calculate the distance between each object and each seed clustering center, and finally assign each object to the nearest clustering center. Although there are many alternatives to measure the distance, the standard of measurement is usually Euclidean distance

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

where d_{xy} denotes the linear distance (i.e. Euclidean distance) between two points $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ in n-dimensional space. The value of each cluster center is updated successively during iteration until the best clustering results are obtained. The nine prior anchor boxes of original YOLOv3 are the result of this process, which are (10×13), (16×30), (33×23), (30×61), (62×45), (59×119), (116×90), (156×198), and (373×326). However, the original prior anchor boxes of YOLOv3 are obtained by clustering the object bounding boxes from the COCO data set, which are not very reasonable for water surface garbage data set, some of which may be redundant and some of which may not be suitable at all. Hence, the object bounding boxes of water surface garbage data set should be reclustered to replace the original prior anchor boxes. The cluster results are shown in Figure 4, where reclustering results are (47×57), (62×68), (59×119), (142×104), (161×306), and (373×326).

Experiments and results

Data set preparation

Since there is no ready-made data set of water surface garbage, we need to establish our own data set of garbage. The images in the data set are mainly taken by the camera in real environment from different scenes, such as playground, floor, river bank, and so on. Only several images are downloaded from the Internet. We randomly selected some of the images and rotated them at different angles to simulate the camera shooting from different directions.

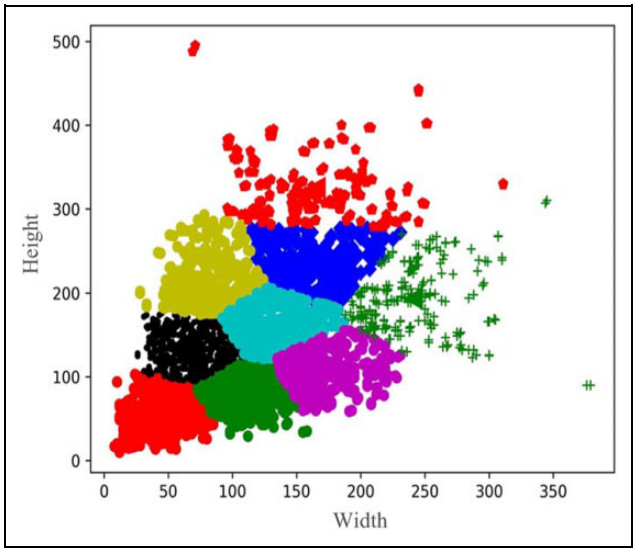

Figure 4. Clustering result. The width and height information of boxes in our data set are clustered as abscissa and longitudinal coordinates, and small icons with different colors and shapes represent different clustering categories.

Besides, we adjusted their brightness to simulate various illumination conditions. The data augmentation process is shown in Figure 5. Next, we used the LabelImg to annotate the images; we circle the target objects in the image with rectangular boxes. The coordinates of the upper left corner and lower right corner of the rectangular boxes will be recorded in the extensible markup language file. Finally, they are made into visual object classes format data set. The number of final training images is 1204, and the number of test images is 301. The number of objects of each class in the data set and the specific experimental parameters are listed in Table 1. The area distribution histogram of the object bounding box in the data set is shown in Figure 6.

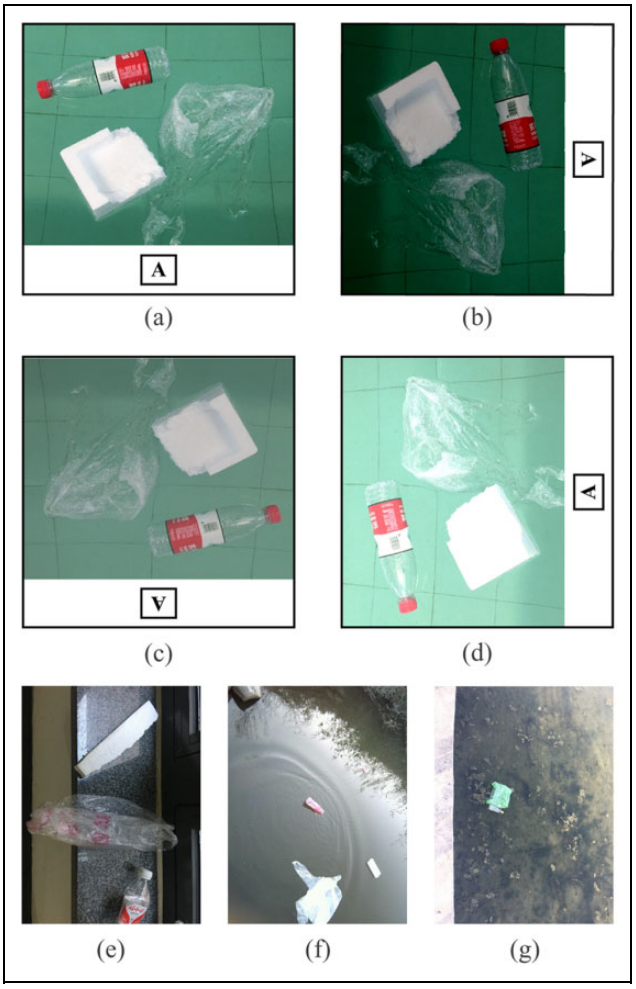

Figure 5. Data set images. Note that (a) comes from shooting, the other three are obtained by rotating (a) with different angles. The state of *A* represents the direction of the image. In addition, the brightness conditions of the images are different, so as to better simulate the different illumination conditions similar to the real environments. In addition, the background of (a), (e), (f), and (g) comes from different scenes, respectively, thus enhancing the robustness of detection.

Detection experiment

In this experiment, we detected three types of water surface garbage: plastic bottle, plastic bag, and styrofoam on the GTX 1080. Then, we measured frames per second, mAP, and AP for each category of the three-scale YOLOv3 (YOLOv3-3S), two-scale YOLOv3 (YOLOv3-2S), and modified the anchor boxes of the two-scale YOLOv3 (YOLOv3-2SMA: modified the anchor boxes of two-scale YOLOv3), respectively. In addition, we trained the SSD network, Faster R-CNN with visual geometry group (VGG16) backbone (Faster R-CNN [FR]16), and Faster R-CNN with Res101 backbone (FR101) in our own data set. The final quantitative results are listed in Table 2 and the detection performance of YOLOv3-2SMA is shown in Figure 7. Note that some of the images come from the water surface scenes and some from other scenes. Different surroundings enhance the robustness of the detection method, so that the robot can be adapt to complex environments.

The results show that the YOLOv3-2SMA is remarkable in all aspects in this experimental scenario. After removing one scales of YOLOv3, the YOLOv3-2S is slightly inferior to the YOLOv3-3S in terms of accuracy, but the speed of YOLOv3-2S is apparently superior to the YOLOv3-3S. Therefore, YOLOv3-2S can better meet the real-time requirements significantly. Meanwhile, some original anchor boxes of YOLOv3-3S are replaced with the anchor boxes reclustered from our own data set so that YOLOv3-2SMA can achieve the same accuracy of YOLOv3-3S.

SSD network also adopts the idea of regression. The network architecture is based on VGG16,[45] which is composed of 13 convolutional layers, 5 pooling layers, and 3

Table 1. Experimental parameters.

Number of training set images	Number of bottles in training set	Number of bags in training set	Number of styrofoam in training set
1204	1390	1192	1225
Number of test set images	Number of bottles in test set	Number of bags in test set	Number of styrofoam in test set
301	275	191	204
GPU	Size of input images	Learning rate	Batch
GTX 1080 × 4	416 × 416	0.001	64

GPU: graphics processing unit.

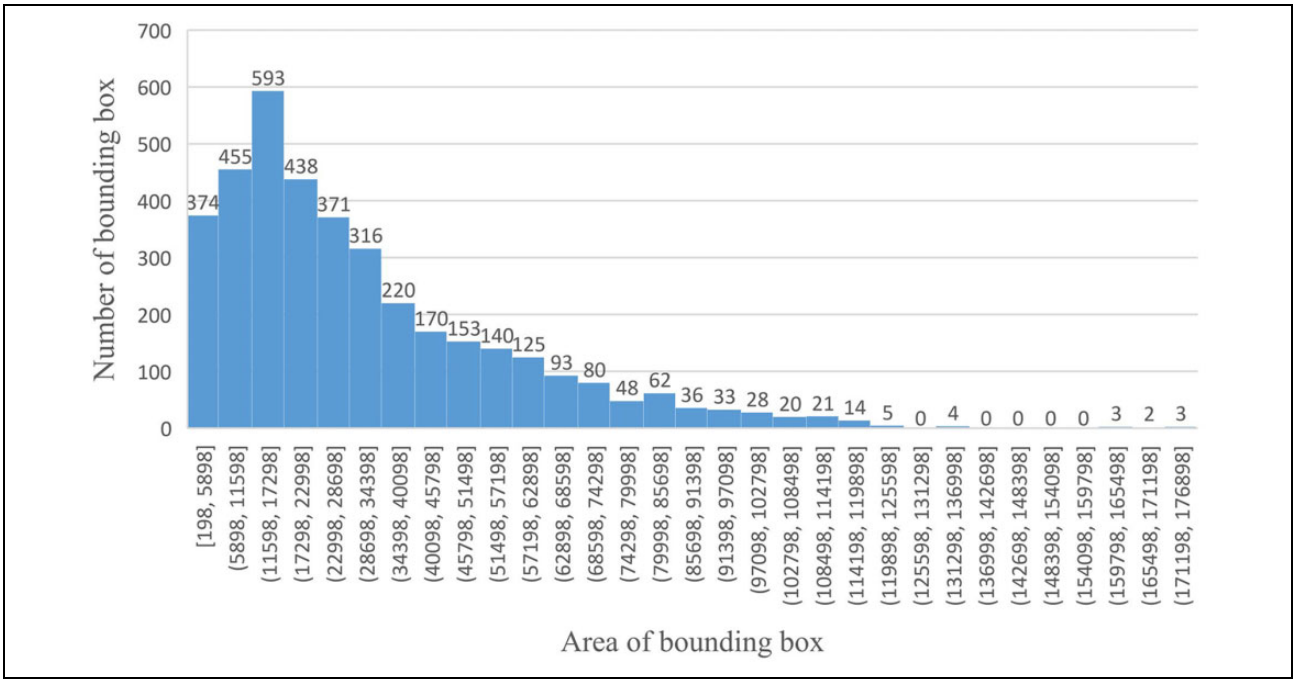

Figure 6. Size distribution of object bounding box in training set.

Table 2. Detection results on data set.

Network	FPS	AP_{bottle}	AP_{bag}	$AP_{styrofoam}$	mAP
YOLOv3-2SMA	**54.1562**	**88.94**	**93.78**	**91.56**	**91.43**
YOLOv3-3S	48.4432	88.00	93.76	91.82	91.19
YOLOv3-2S	54.0394	89.26	92.03	91.09	90.79
SSD	—	85.45	84.92	87.50	85.96
FR16	17.2034	80.22	87.62	85.94	85.79
FR101	13.4442	79.73	84.05	88.76	84.18

FPS: frames per second; YOLOv3: you only look once v3; SSD: single shot detector; AP: average precision; mAP: mean average precision.

fully connected layers. In the ImageNet Image Classification and Location Challenge in 2016, this model achieved excellent results.[34] Moreover, SSD can detect objects from multiple scales, whose performance is outstanding. Experiments show that SSD network is inferior to the YOLOv3 network in all aspects. It is close to YOLOv3 in the detection of bottles, but it is far worse in the detection of bags and styrofoam.

Faster R-CNN-based networks are weaker than YOLOv3 in both speed and accuracy, especially in speed. Faster R-CNN takes nine anchor boxes from each pixel of feature map, resulting in that the computation for each anchor box classification is heavy. Furthermore, the feature extraction network of Faster R-CNN is not more sufficient than that of YOLOv3. Besides, YOLOv3 can predict objects from three-scale feature maps, whereas Faster R-CNN only predicts from one.

Field experiment

In addition, we conducted field experiments to verify whether the robot can be adapted to different environments and observed its states in complex surroundings to judge whether it has the ability to solve practical problems. The field experiments are shown in Figures 8 and 9.

The workflow of the robot is illustrated in Figure 10. When the device launches, it will cruise on the water surface with a random path. At this moment, the detection module will start to detect whether there is a target object in the visual range. When more than one garbage is detected in the visual range, we will first select the garbage closest to the center coordinate as the target. Because it is

Figure 7. Detection results.

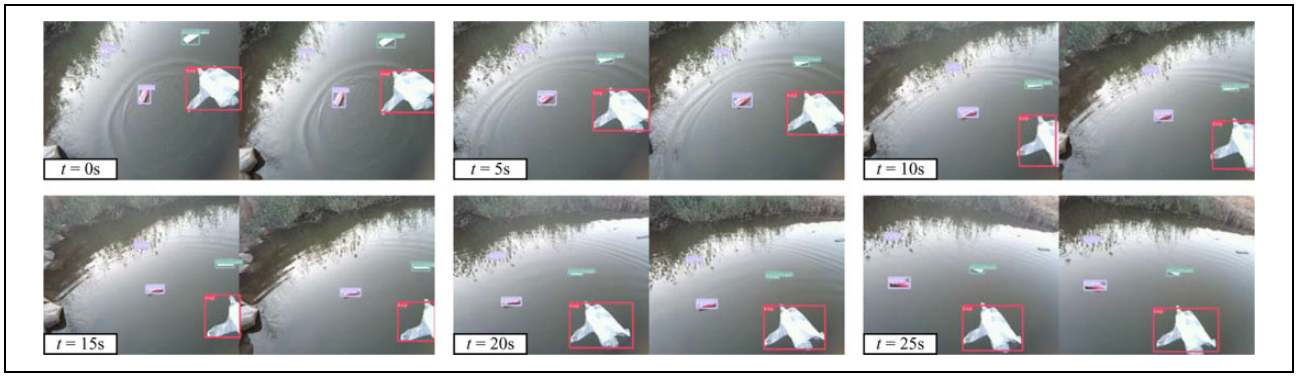

Figure 8. The detection of field experiment.

Figure 9. Scenario of a field experiment. When the target object was detected, the robot approached to the object gradually. Once the robot approached into the graspable distance range, it implemented grasping.

nearest the center of the field of vision, which the possibility of box loss is less when the robot approaches to the target. If a target is found, the robot will immediately adjust its pose and accelerate to the target based on the detected target information. Until the distance between robot and the target is proper for grasping, the manipulator will grasp the target. After the target is successfully captured, it will return to the detection module for redetection.

In the field environment, the robot detected the objects accurately and in real time. According to the detected object information, it then approached the object and implemented the grasping successfully.

Discussion

The experimental results demonstrated that YOLOv3-2SMA outperformed the other detection methods. High-

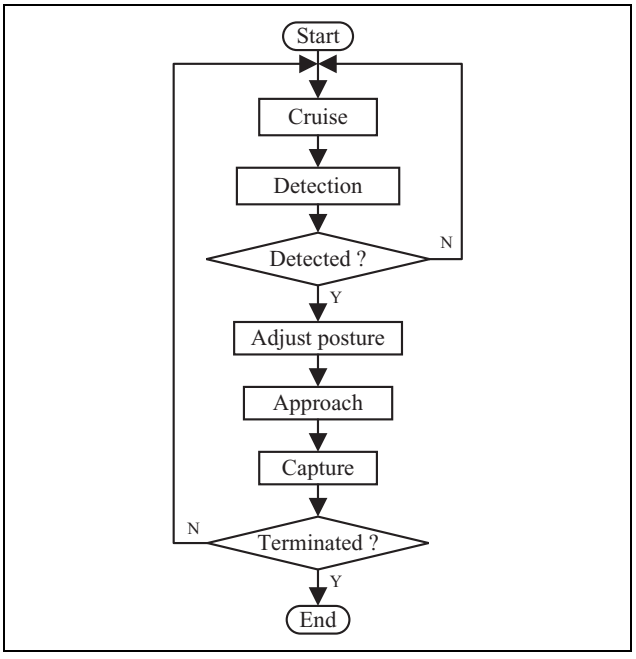

Figure 10. Workflow chart for the garbage capture robot.

speed detection can process the image in real time and provide the object information for the robot in time in the changeable and complex environment. Even though Faster R-CNN has made a significant breakthrough in accuracy, it cannot achieve real-time performance due to the computational burden from the two-stage network. YOLOv3 stands out in detection networks in terms of speed, owing to YOLOv3-2S is one scale less than YOLOv3-3S, which means reducing computation load and resulting in a significant increase in detection speed. YOLOv3-2S can realize the detection task faster.

High-precision detection can help the robots complete tasks more accurately, reliably, and stably. The performances of SSD and Faster R-CNN in accuracy are far less than the YOLOv3 network. The basic frameworks of SSD and Faster R-CNN are VGG16, VGG19, and Res101, while the basic network of YOLOv3 is Darknet-53. Note that Darknet-53 performs on par with state-of-the-art classifiers but with fewer floating-point operations and more speed. Darknet-53 is better than ResNet-101 and 1.5× faster. Darknet-53 has similar performance to ResNet-152 and is 2× faster. Darknet-53 also achieves the highest measured floating-point operations per second.[33] Because YOLOV3-2S lacks one-scale prediction results, the accuracy will inevitably be affected. YOLOv3 is remarkable in precision and speed, so when it is improved, the rising space will not be very conspicuous. Thus, our goal is to achieve high speed while maintaining the original precision at least. The anchor boxes after reclustering will be more appropriate to the garbage data set; therefore, the accuracy of YOLOv3-2SMA can be compensated. In conclusion, YOLOv3-2SMA possesses a better performance than the others in both the speed and the accuracy closely related to the capability of autonomous cleaner robot.

Based on the aforementioned results, the proposed algorithm outperforms previous studies and is suitable for autonomous cleaner robot. However, for a better robustness in the complex aquatic environments, a more adequate data set is indispensable.

Conclusion and future work

A real-time and high-precision detection method is extremely critical for the successful grasping of water surface garbage robot; however, conventional detection methods cannot meet the requirements at the same time. For a comprehensive consideration, a modified YOLOv3 network is developed in this study. First, we transform three-scale detection into two-scale detection for a high speed. Second, we adjust the prior anchor boxes according to K-means clustering over our own data set to compensate for the detection accuracy. Compared with other detection networks during the experiment, YOLOv3-2SMA can achieve 91.43 mAP in 18.47 ms on GTX 1080, which ensures the real-time and accurate garbage detection. Furthermore, field experiments reveal that this method can be applied to the robot in complex aquatic environments. The high-speed and high-precision detection of YOLOv3-2SMA provides overwhelming visual support for the robot's cleaning tasks, which make the water surface garbage cleaning autonomous and intelligent. Meanwhile, it can greatly improve the efficiency of the waste cleaning industry and make outstanding contributions to protecting the water surface environment. Additionally, it will reduce the investment of human resources and ensure the personal safety of cleaners.

The next step is to obtain more complex scene image data, so as to make its detection more robust and accurate. Besides, studies indicate that plastic garbage poses a serious threat to marine life, human health, and the economy.[46–51] In the future, we will extend the domains of garbage cleaning missions from water surface to underwater for widespread use.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Natural Science Foundation of China under Grant 61725305, Grant U1909206, Grant 61873291, and Grant 61773416 and in part by the Minzu University of China 111 Project.

ORCID iDs

Xiali Li https://orcid.org/0000-0001-7950-6204

Licheng Wu https://orcid.org/0000-0001-5739-634X
Junzhi Yu https://orcid.org/0000-0002-6347-572X

References

1. Kins P and Gupta J. Perspective: a healthy planet for healthy people. *Global Sustain* 2019; 2: 1–9.
2. Laschi C, Mazzolai B, and Cianchetti M. Soft robotics: technologies and systems pushing the boundaries of robot abilities. *Sci Robot* 2016; 1: eaah3690.
3. Albitar H, Dandan K, Ananiev A, et al. Underwater robotics: surface cleaning technics, adhesion and locomotion systems. *Int J Adv Robot Syst* 2016; 13: 7.
4. Yuan F, Hu S, Sun H, et al. Design of cleaning robot for swimming pools. In: *Proceedings 2011 international conference on management science and industrial engineering (MSIE)*, Harbin, China, 8–11 January 2011, pp. 1175–1178. Harbin: IEEE.
5. Bai J, Lian S, Liu Z, et al. Deep learning based robot for automatically picking up garbage on the grass. *IEEE Trans Consum Electron* 2018; 64(3): 382–389.
6. Kim J, Mishra AK, Limosani R, et al. Control strategies for cleaning robots in domestic applications: a comprehensive review. *Int J Adv Robot Syst* 2019; 16(4): 1–21.
7. Li Z, Li Z, Li Y, et al. Development of the self-adaptive pipeline cleaning robot. *Adv Mater Res* 2010; 97–101: 4482–4486.
8. Xu M, Karuppusamy NS, and Kang B. A novel design to improve the cooperative ability of the multi-cleaning robot in the unknown environment. *Adv Sci Lett* 2017; 23(10): 9557–9560.
9. Prabakaran V, Elara MR, Pathmakumar T, et al. Floor cleaning robot with reconfigurable mechanism. *Autom Constr* 2018; 91: 155–165.
10. Mahler J, Matl M, Satish V, et al. Learning ambidextrous robot grasping policies. *Sci Robot* 2019; 4(26): eaau4984.
11. Mahler J, Pokorny FT, Hou B, et al. Dex-Net 1.0: a cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In: *Proceedings IEEE international conference on robotics and automation (ICRA)*, Stockholm, Sweden, 16–21 May 2016, pp. 1957–1964. Stockholm: IEEE.
12. Mahler J, Liang J, Niyaz S, et al. Dex-Net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, March 2017.
13. Mahler J, Matl M, Liu X, et al. Dex-Net 3.0: computing robust robot vacuum suction grasp targets in point clouds using a new analytic model and deep learning. *arXiv preprint arXiv:1709.06670*, September 2017.
14. Kong S, Tian M, Qiu C, et al. IWSCR: an intelligent water surface cleaner robot for collecting floating garbage. *IEEE Trans Syst Man Cybern Syst* 2020; 1–11. DOI: 10.1109/TSMC.2019.2961687.
15. Whitehill J and Omlin CW. Haar features for FACS AU recognition. In: *Proceedings. IEEE international conference on automatic face and gesture recognition (FGR06)*, Southampton, UK, 11–12 April 2006, pp. 97–101. Southampton: IEEE.
16. Ojala T, Pietikäinen M, and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 2002; 24(7): 971–987.
17. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, San Diego, California, USA, 20–26 June 2005, pp. 886–893. San Diego: IEEE.
18. Ma Y, Li Q, Zhou Y, et al. A surface defects inspection method based on multidirectional gray-level fluctuation. *Int J Adv Robot Syst* 2017; 14(3) 1–17.
19. Szarvas M, Yoshizawa A, Yamamoto M, et al. Pedestrian detection with convolutional neural networks. In: *Proceedings IEEE intelligent vehicles symposium (IV'05)*, Las Vegas, Nevada, USA, 6–8 August 2005, pp. 224–229. Las Vegas: IEEE.
20. Chen F and Jahanshahi MR. NB-CNN: deep learning-based crack detection using convolutional neural network and naïve Bayes data fusion. *IEEE Trans Ind Electron* 2018; 65(5): 4392–4400.
21. Kagaya H, Aizawa K, and Ogawa M. Food detection and recognition using convolutional neural network. In: *Proceedings of the 22nd ACM international conference on multimedia*, Orlando, Florida, USA, 18–19 June 2014, pp. 1085–1088. Orlando: ACM.
22. Kyathanahally SP, Döring A, and Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magn Reson Med* 2018; 80(3): 851–863.
23. Dias PA, Tabb A, and Medeiros H. Apple flower detection using deep convolutional networks. *Comput Ind* 2018; 99: 17–28.
24. Kellenberger B, Marcos D, and Tuia D. Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning. *Remote Sens Environ* 2018; 216: 139–153.
25. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging* 2017; 30: 427–441.
26. Choi H. Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions. *Q J Nucl Med Mol Imag* 2018; 52(2): 109–118.
27. Chiao J, Chen K, Liao K, et al. Detection and classification the breast tumors using mask R-CNN on sonograms. *Medicine (Baltimore)* 2019; 98(19): e15200.
28. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings 27th IEEE international conference on computer vision and pattern recognition (CVPR)*, Columbus, Ohio, USA, 24–27 June 2014, pp. 580–587. Columbus: IEEE.
29. Girshick R. Fast R-CNN. In: *Proceedings IEEE international conference on computer vision (ICCV)*, Santiago, Chile, 11–18 December 2015, pp. 1440–1448. Santiago: IEEE.

30. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proceedings 29th international conference neural information processing sysems (NIPS)*, Montreal, Canada, 11–12 December 2015, pp. 91–99. Montreal: NIPS.
31. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, realtime object detection. In: *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)*, Las Vegas, Nevada, USA, 26 June–1 July 2016, pp. 779–788. Las Vegas: IEEE.
32. Redmon J and Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)*, Honolulu, Hawaii, USA, 21–26 July 2017, pp. 7263–7271. Honolulu: IEEE.
33. Redmon J and Farhadi A. YOLOv3: an incremental improvement. *arXiv preprint arXiv:1804.02767*, April 2018.
34. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: *European conference on computer vision (ECCV)*, Amsterdam, The Netherlands, 8–16 October 2016, pp. 21–37. Amsterdam: ECCV.
35. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; 55(1): 119–139.
36. Vapnik V. *The nature of statistical learning theory*. Berlin: Springer Science & Business Media, 2013.
37. Viola P and Jones MJ. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)*, Kauai, Hawaii, USA, 8–14 December 2001, pp. 511–518. Kauai: IEEE.
38. Ali A, Olaleye OG, and Bayoumi M. Fast region-based DPM object detection for autonomous vehicles. In: *Proceedings of the IEEE 59th midwest symposium on circuits and systems (MWSCAS)*, Abu Dhabi, United Arab Emirates, 16–19 October 2016, pp. 1–4. Abu Dhabi: IEEE.
39. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2015; 37(9): 1904–1916.
40. Benjdira B, Khursheed T, Koubaa A, et al. Car detection using unmanned aserial vehicles: comparison between faster R-CNN and YOLOv3. In: *Proceedings 1st international conference on unmanned vehicle systems, Oman (UVS)*, Muscat, Oman, 5–7 February 2019, pp. 1–6. Muscat: IEEE.
41. Park JH, Hwang H, Moon J, et al. Automated identification of cephalometric landmarks: part 1—comparisons between the latest deep-learning methods YOLOV3 and SSD. *Angle Orthod* 2019; 89(6): 903–909.
42. Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett* 2010; 31(8): 651–666.
43. Alsabti K, Ranka S, and Singh V. An efficient K-means clustering algorithm. *Turk J Electr Eng Comput Sci*, 1997. https://surface.syr.edu/eecs/43
44. Sculley D. Web-scale K-means clustering. In: *Proceedings 19th. International conference world wide web ACM (WWW'10)*, Raleigh, North Carolina, USA, 26–30 April 2010, pp. 1177–1178. Raleigh, North Carolina: WWW.
45. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, September 2014.
46. Schneider F, Parsons S, Clift S, et al. Collected marine litter—a growing waste challenge. *Mar Pollut Bull* 2018; 128: 162–174.
47. Perkins S. Plastic waste taints the ocean floors. *Nature*. DOI: 10.1038/nature.2014.16581, December 2014. https://www.nature.com/news/plastic-waste-taints-the-ocean-floors-1.16581
48. Xue Z, Liu J, Wu Z, et al. Development and path planning of a novel unmanned surface vehicle system and its application to exploitation of Qarhan Salt Lake. *Sci China Inf Sci* 2019; 62: 084202:1–084202:3.
49. Yu J, Li X, Pang L, et al. Design and attitude control of a novel robotic jellyfish capable of 3D motion. *Sci China Inf Sci* 2019; 62: 194201:1–194201:3.
50. Ding J, Jiang F, Li J, et al. Microplastics in the coral reef systems from Xisha islands of South China Sea. *Environ Sci Technol* 2019; 53(14): 8036–8046.
51. Bergmann M, Tekman MB, and Gutow L. Marine litter: sea change for plastic pollution. *Nature* 2017; 544: 297.