Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-020-02843-w

ORIGINAL RESEARCH

Manipulator grabbing position detection with information fusion of color image and depth image using deep learning

Du Jiang[1] · Gongfa Li[1,3] · Ying Sun[1] · Jiabing Hu[1] · Juntong Yun[2] · Ying Liu[2]

Received: 20 July 2020 / Accepted: 12 December 2020

Abstract

In order to ensure stable gripping performance of manipulator in a dynamic environment, a target object grab setting model based on the candidate region suggestion network is established with the multi-target object and the anchor frame generation measurement strategy overcoming external environmental interference factors such as mutual interference between objects and changes in illumination. In which, the success rate of model detection is improved by adding small-scale anchor values for small area grabbing target position detection. Further, 94.3% crawl detection success rate is achieved on the multi-target detection data sets using the information fusion of color image and depth image. The methods in this paper effectively improve the model's robustness and crawl success rate.

Keywords Manipulator · Grabbing position detection · Information fusion · Deep learning

1 Introduction

At present, robot technology has been applied in more fields which requires the better stability of robot limited to its grasping performance in complex environment. Therefore, it is particularly important for the robot to achieve stable grasping in unstructured environment. In order to achieve this goal, the related researchers have spent a lot of time to solve the problem of robot grasping success rate and robustness For example, artificial design features were used to express the grabbing information in the image, or the complete three-dimensional model of the object were used to generate the grasping position. However, in the practical application of such methods, there will be some problems such as the complexity of artificial design features and the lack of a complete three-dimensional model (Billard and Kragic 2019; Fontanelli et al. 2014; Sombolestan et al. 2019; Wen et al. 2019).

For the problem, the deep learning technology represented by convolutional neural network is worth trying because of its strong fitting ability, which can obtain the function mapping between input and output data according to its successful application in computer vision. So many researchers had applied it in intelligent recognition (Jiang et al. 2019a, b, c; Sun et al. 2020a, b; Cui et al. 2020a, b; Li et al. 2019a, b, c, d, e), detection (Cai et al. 2020a, b; Li et al. 2019a, b, c, d, e; Wang et al. 2020; He et al. 2019), tracking and other fields (Li et al. 2017; Cui et al. 2020a, b; Cai et al. 2020a, b; Cheng et al. 2018; Hassan et al. 2020; Huang et al. 2020), and its accuracy and robustness had been greatly improved (Weng et al. 2020; Jiang et al. 2019c; Li et al. 2018a, b; Han et al. 2018; Nie et al. 2018). At the same time, the processing speed of deep learning models was also faster, some of which have reached real-time processing speed with the continuous development of GPU and other hardware facilities (Duan et al. 2020; Sangwan and Jain 2019; (Paolini et al. 2014); Bohg et al. 2014; Cheng et al. 2020a, b; Liao et al. 2020a, b; (Caldera et al. 2018). Although it has been proved to be effective in different classification and detection problems, it is very challenging to

✉ Du Jiang
jiangdu@wust.edu.cn

Gongfa Li
ligongfa@wust.edu.cn

[1] Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China

[2] Research Center for Biomimetic Robot and Intelligent Measurement and Control, Wuhan University of Science and Technology, Wuhan 430081, China

[3] Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China

Published online: 04 January 2021

grasp the unknown objects stably because of the different shape and posture of objects. Compared with human beings who can grasp objects immediately through coordination of eyes, brain and limbs, there are many problems in manipulator grabbing position detection with deep learning model (Tian et al. 2019; Li et al. 2019a, b, c, d, e; Ma et al. 2019; Wang et al. 2019; Liao et al. 2019). To summarize, how to realize robot autonomous grasping is vital, thus helping robots grasp objects like human beings. Therefore, it is of great significance to carry out research on robot autonomous grasping. The innovation of this paper lies in the further optimization of the deep learning framework for multimodal fusion for multi-target object stoic grasping, and the success rate of the model is enhanced by combining multiscale candidate boxes.

This article is mainly composed of 6 parts. Among them, part 2 introduces the research results of the relevant researchers in this field, and discusses the rationality of the methods adopted. The third part then explains the data set used and its pre-processing methods. Part 4 and 5 respectively introduce the methods and experimental results in this paper.

2 Related work

In fact, the detection of object grasping position is a kind of vision detection problem that the robot can grasp the position of the object in the environment with sensors, in which it is a key task based on visual information how to predict the potential grasping position from the sensor information, and map the pixel value in the image information into the real world coordinate system. Although the commonly used solution was to use machine learning to extract features which has been applied to many filed such as dealing with the sEMG (Qi et al. 2020; Li et al. 2019a, b, c, d, e; Yu et al. 2019; Huang et al. 2019; Yu et al. 2020; Luo et al. 2020), however, the most common solution is to classify the input data by convolution neural network which can extract, classify and coordinate the input sensor data using off-line training, so as to carry out the grasp configuration planning (Li et al. 2019a, b, c, d, e; Cheng et al. 2020a, b; (Ma et al. 2020; Qi et al. 2019; Lin et al. 2019; Hsiao et al. 2010). Though reasonable training, it can also achieve good results on new objects because of its strong feature generalization ability (Lei et al. 2017); Mahler et al. 2017; Agrawal et al. 2016). As a result, we proposed a method to detect the parameters of the rectangular frame based on the rectangular frame in this paper. Literature (2013) proposed a method based on search window, used to detect robots crawl the rectangle, on three-dimensional point cloud crawl in a rectangular box parameter mapping. The entire crawl testing of the rectangle is divided into two phases, the first stage by a small depth of neural network to search all possible rectangular box, and then produce some of the top rectangle, and with a more deeper nerve network from these relatively ·high ranking rectangular box, locate the top of the rectangle, the rectangle is that we crawl configuration required. The success rate of this method is 75%, and the processing time of each image was 13.5 s. Based on the similar method, the input image is preprocessed to identify candidate object regions, and then CNN classifier is applied to each region. At the same time, a structured penalty term was introduced to optimize the connection between multimode, which greatly reduced the complexity of the network. This method improves the success rate of grasping position detection to 81.8% (Wang et al. 2016). However, they needed take a lot of time to search candidate regions, which leads to poor real-time performance of the model. In order to ensure the speed of grabbing pose detection and avoid the waste of time caused by search sampling, the end-to-end detection methods began to attract people's attention. In which, the most model directly regresses the input image data to get a corresponding grabbing frame, however, whose success rate of this method is often not high. Redmon and Angelova (2014) achieved 88.0% of the capture accuracy by using the method of transfer learning, and it only takes 76 ms to process a picture, which greatly improves the accuracy and real-time performance. In reference (Kumra and Kanan 2017) , two juxtaposed residual layers were used to extract image features, and then the features of two branches were stitched together, and three layers of full connection layer were added to output grab frame parameters, which effectively improved the real-time performance of Cornell grabbing data set, and achieved a success rate of 89.2%. On this basis, matrix was introduced to quantify and classify the direction of grasping rectangle, and 93.2% success rate of grasping position detection was achieved on Cornell dataset (Guo et al. 2017). Further, a very effective grabbing algorithm in (Lin et al. 2014) was proposed without knowing the object information in advance which could gain grasping direction only by the depth information in the scene using a RGB-D camera fixed on the robot hand, and it included two steps. The first one of them is an extensive search covering all objects on the desktop, and the other was doing more detailed scanning after approaching the objects. The experiment result showed that the success rate was up to 95% for a single object on the desktop. With the continuous improvement of the capture detection network model, researchers have also begun to improve the detection accuracy from the input mode. Aiming at the problem of improving the real-time and accuracy of 3D object detection based on RGB-D image, an improved step-by-step super-pixel merging similarity sampling method is proposed in reference (Lu et al. 2016), and the multi-mode such as color, texture, contour and depth of the target is analyzed by using multi-core learning theory. The effective fusion of information can greatly reduce the number of sampling windows and reduce the operation time, at the same time, ensure the high recall rate and accurate positioning of window sampling, effectively fuse the multi-modal information of the target object, and improve the object detection accuracy. Because multi-feature fusion can effectively increase the amount of information, so that the success rate of the model can be increased (Liao et al. 2020a, b; Tian et al. 2020; Tan et al. 2019; Chen et al. 2019; Sun et al. 2020a, b).

It can be seen from the research of multimodal deep learning model that the current research mainly focuses on data preprocessing, multimodal feature fusion, model optimization and so on Li et al. 2018a, b; Schwarz et al. 2018; Lin et al. 2018; Zhang et al. 2018; Wang et al. 2015). Considering that RGB-D data based multimodal model is commonly used in the end fusion, and has achieved certain results, this paper intends to use this data feature fusion method to train the multi-target object capture detection model.

3 Preprocessing of dataset

Generally, the training Dataset used in the capture detection model contains only one object to be grasped, however, many objects often appear in the same image. In order to overcome the mutual interference between objects and realize target capture detection in unstructured environment, this paper uses multi-target data set first proposed in reference (Chu et al. 2018) as training samples, which contains two types of images: single object and multi object. The single object data set is composed of Cornell grabbing data set (Fig. 1), including 885 single object images in total on which several rectangular boxes are manually marked to represent the grabbing rectangle. As shown in Fig. 2, the upper row represents the negative sample grabbing frame, i.e., the infeasible grab expression, and the next row represents the positive sample grabbing frame, i.e., the grasping rectangle

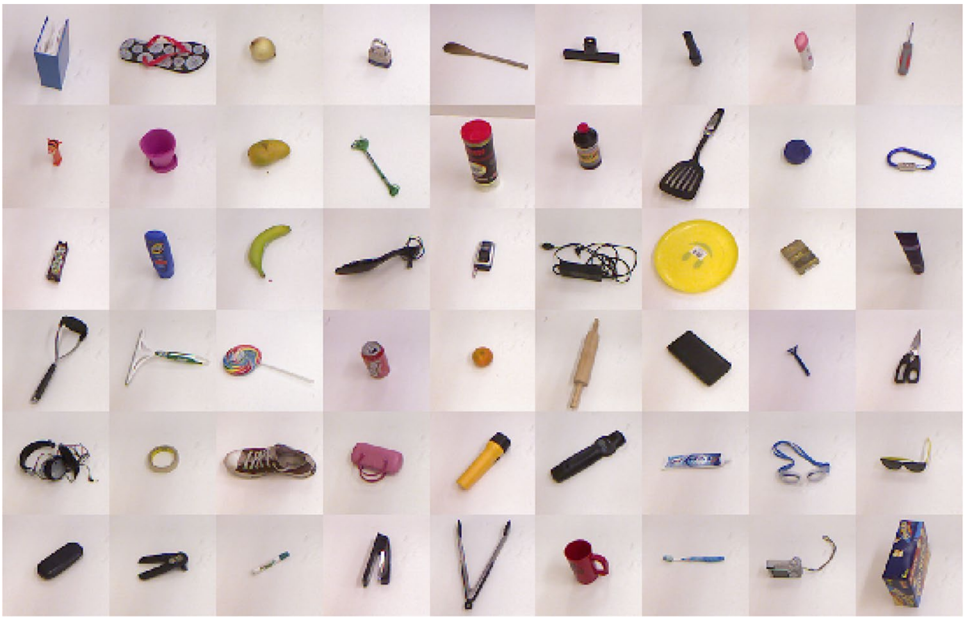


Fig. 1 Cornell grabbing dataset (Chu et al. 2018)

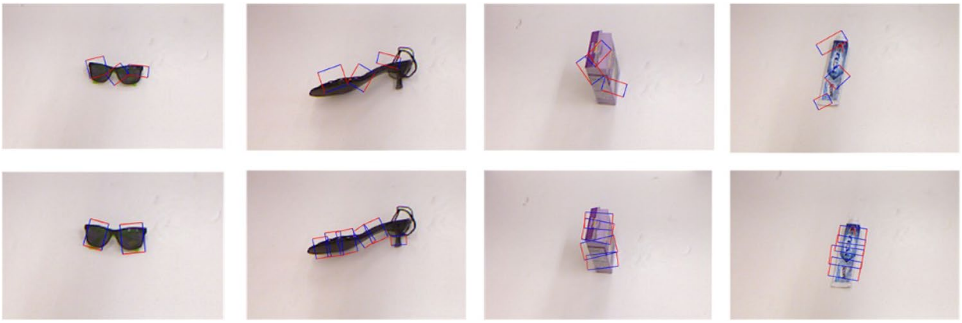


Fig. 2 Negative and positive labels of some objects (Chu et al. 2018)

expression that can be normally grasped by the manipulator. In this paper, we only extract the positive sample rectangle as learning feature for supervised learning.

In the multi object dataset (Fig. 3), a picture contains multiple target objects. Figure 3a, b respectively show the color image and depth map of the corresponding multi-objective data set, and Fig. 3c is the rectangle box for the positive sample of the corresponding object. The dataset contains 97 color images of multi-target objects from different perspectives in which each color image has its own corresponding depth map and grabbing rectangle annotation, and the depth map and color image are aligned.

In order to verify the success rate of multi-target capture detection model for new objects and the same object with different angles, this paper adopts two data set partition strategies based on image and object.

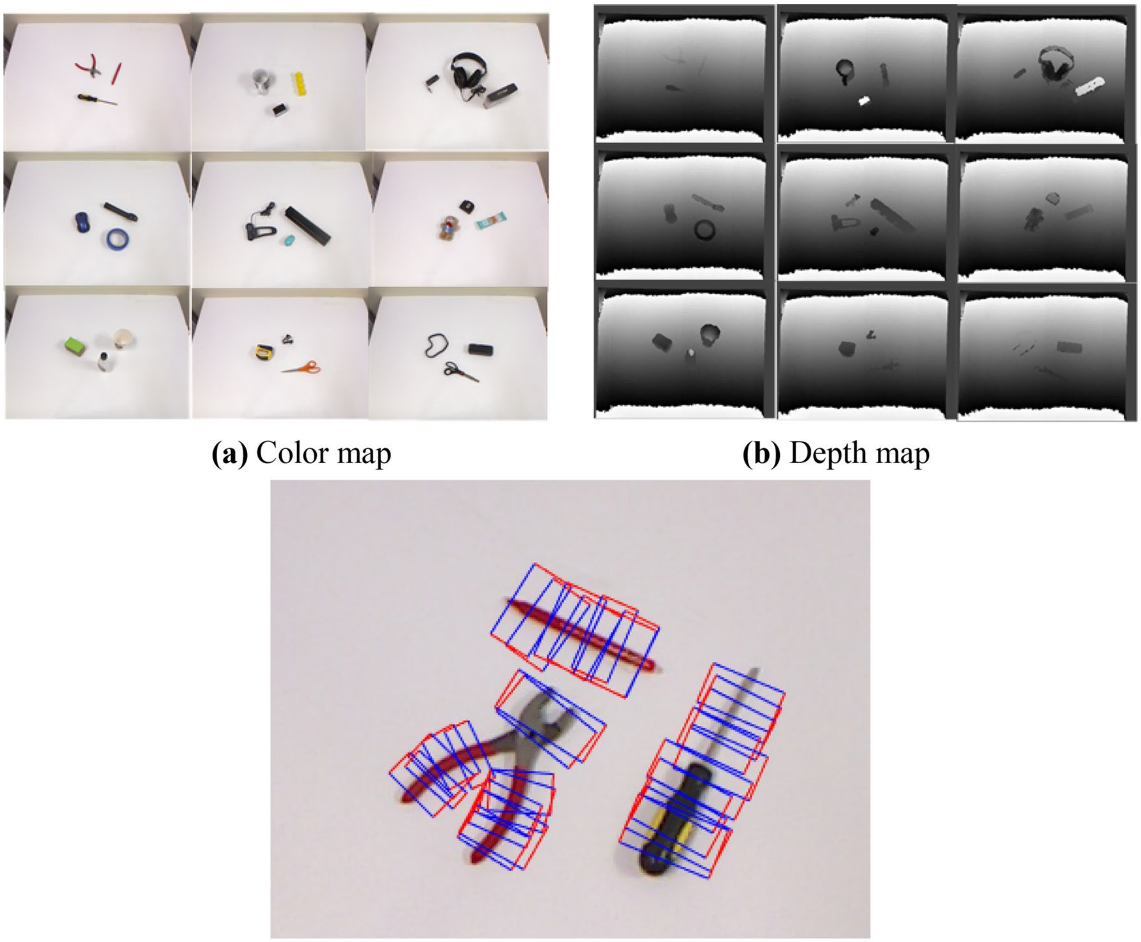
(1) Based on image

Based on the image segmentation data set, all images are randomly divided into training set and test set according to the ratio of 7:3, because the data set in this paper contains pictures of the same object from different angles. This method of data set division is to verify the detection effect of the model on the grabbing frame when the objects have been seen in different angles.

(2) Based on object

For the training set which is divided into different training sets, the training set is used to divide the training set into different test sets to ensure that the training set contains different classes of objects.

At the same time, each image in the original data set is cut according to the center whose size is unified to 300 × 300 pixels. Then, each image is processed with up-down flip, left–right symmetry and color jitter. As a result, the original is expanded to about 8 times from 982 to 7856.



(a) Color map

(b) Depth map

(c) Positive sample grab rectangle

Fig. 3 Multi object dataset. **a** Color map **b** Depth map **c** Positive sample grab rectangle

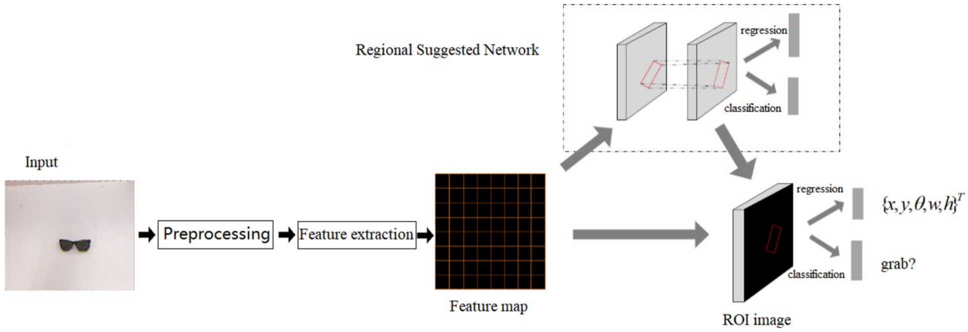


Fig. 4 Multi target object capture detection model

4 Rectangle grab frame detection model

In this paper, the convolution neural network is used to learn the features of the rectangular frame, that is, the two-stage target detection network is used to learn the location features of the rectangular frame. In order to detect the grabbing rectangle at any angle, a regression strategy of rotating rectangular box based on arbitrary angle is proposed. Taking the two-stage object detection model as the network framework, the region of interest (ROI) with angle information is generated by adding an angle variable in the region recommendation network. The ROI feature map is pooled to a uniform size. The grabbing frame is regressed through the full connection layer, and whether the candidate area is a grab frame is classified.

The multi-target grabbing detection model based on the two-stage object detection network (Fig. 4). At first, the convolution layer is used to extract the feature map from input image, which is coming to be used to generate candidate boxes oriented to any direction in region suggested network including two parallel networks (the classification layer and the regression layer) that is applied to classify the grab frame and background, and make initial regression of grasping position. Simultaneously, it is the interest pooling layer that projects candidate frames in any direction onto the feature map by maximum pooling. Finally, the classifier formed by two full connection layers is used to further regress the candidate grab frames to ensure the regression accuracy.

4.1 Generation of candidate grabbing region

In order to adapt to the task of grabbing position detection and the learning of the position information of grabbing rectangle box, a new direction variable is added in the generation of candidate grabbing region, adding six direction angles: – π/6, 0, π/6, π/3, π/2, 2π/3. As shown in Fig. 5, each anchor has three variables: size, scale and angle. When the grasping features of objects such as cups and plates in Fig. 6 are relatively small, a smaller anchoring scale was used to extract small target features in the area recommendation network. When the initial anchoring scale is (8, 16, 32), two strategies are studied: (1) changing the anchor size and using (4, 8, 16); (2) adding new anchor and using (4, 8, 16, 32).

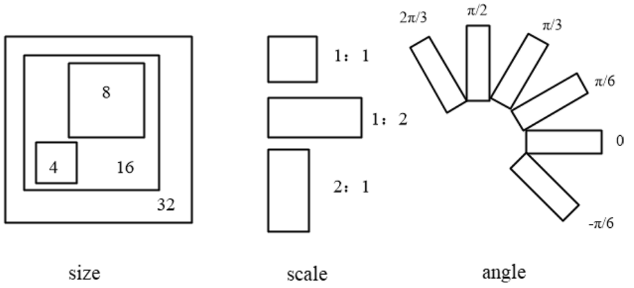


Fig. 5 Three variables corresponding to anchor point

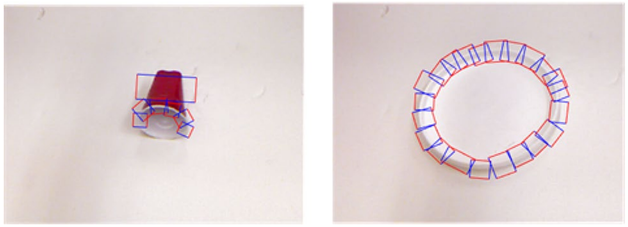


Fig. 6 Grasping characteristics of small areas

4.2 Setting of loss function

In this detection frame, there are two outputs: the score of grabbing rectangle/non grabbing rectangle and the coordinate of tilted box. The goal of learning is the mapping relationship between anchor box and real label, namely affine transformation and exponential mapping. On the feature map, the original image area corresponding to each point is classified into two categories (judging whether it is the target area), and the coordinates are generated by the region recommendation network. This method of rectangular box coordinate generation is not accurate enough. We can learn the mapping relationship between anchor and ground truth box twice to regress the final captured rectangular coordinates. Therefore, the loss function in this model, as shown in formula (1)–(5), consists of two parts: the first part represents the loss value of classification, and the latter part represents the loss value of regression function.

$$\mathrm{L}(p, l, v^{*}, v) = \mathrm{L}_{cls}(p, l) + \lambda l \mathrm{L}_{\mathrm{reg}}(v^{*}, v) \tag{1}$$

In which:

P is the probability of the classification;

$v = (v_x, v_y, v_h, v_w, v_\theta)$ is the predicted value for the grab box;

$v^{*} = (v_x^{*}, v_y^{*}, v_h^{*}, v_w^{*}, v_\theta^{*})$ is the location of the grab box for the real label;

L is the label of classification; When L = 1, the regression loss is calculated when anchor is grabbing the position, otherwise calculating the classification loss when L = 0.

For regressions of border positions, a smooth loss is defined as:

$$\mathrm{L}_{\mathrm{loc}}(t^{*}, v) = \sum_{i \in \{x,y,w,h,\theta\}} smooth_{L_1}(t_i^{*} - v_i) \tag{2}$$

$$smooth_{L_1}(x) = \begin{array}{ll} 0.5x^2 & if|x| \ < 1 \\ |x| - 0.5 & otherwise \end{array} \tag{3}$$

$$\begin{aligned} v_x &= \frac{x - x_a}{w_a}, v_y = \frac{y - y_a}{h_a}, \\ v_h &= \log\frac{h}{h_a}, v_w = \log\frac{w}{w_a}, v_\theta = \theta - \theta_a + k, \end{aligned} \tag{4}$$

$$\begin{aligned} v_x^{*} &= \frac{x^{*} - x_a}{w_a}, v_y^{*} = \frac{y^{*} - y_a}{h_a}, \\ v_h^{*} &= \log\frac{h^{*}}{h_a}, v_w^{*} = \log\frac{w^{*}}{w_a}, v_\theta^{*} = \theta^{*} - \theta_a + k, \end{aligned} \tag{5}$$

In which, x, x_a, x^{*}, respectively represents the prediction rectangle, anchor box, and real label coordinate box.

4.3 Fusion strategy of depth information and color information

Based on the model, four different input modes are used to train and test the model. A total of 7856 images of the expanded multi-target data set are used to verify the performance of the grabbing rectangle detection model by using color image, depth map, RGD three channel image and color-depth fusion as input. Among the above four types of data, the color image and RGD are three channels, the depth image is a single channel, and the color depth is integrated into two parallel three-channel inputs. The structure of grabbing pose detection network based on color image (Fig. 7a). The input is 300 × 300 three channel RGB image, which is transformed into 224 × 224 size by preprocessing network and sent to feature extraction network. The network is composed of 14 convolution layers, 14 activation layers and 4 pooling layers, and the kernel of convolution layer is the kernel of convolution layer_ Size = 3, pad = 1, stripe = 1, all pooling layer kernel_ size = 2, pad = 0, stride = 2. As shown in Fig. 7b, for the depth image, the input layer channel needs to be changed to 1.

Because of the grasp position detection, the color texture information of the object and the relative relationship of the position need to be considering. Compared with the study of single mode information, fusion of color and depth information can provide a richer learning features, so this paper used two ways to learn the multi-modal RGB-D characteristics. As shown in Fig. 8, the three-channel deep image grayscale processing is then replaced by the B channel in the RGB image information, and the three-channel RGD information is entered into the network.

In order to unify the depth value and color channel value, the depth value obtained by z-coordinate is normalized to 0–255 interval, and the data visualization results of three modes are finally obtained, as shown in Fig. 9, where (a) is the color RGB three channels, (b) is the depth image graying, and (c) is the RGB image after color depth fusion.

In Fig. 10, two parallel neural networks are used to extract the convolution features of the three channel color image and the three channel depth image respectively, and the pooled features are spliced on the full connection layer for the frame regression and classification of the subsequent candidate grabbing frames.

5 Analysis of experimental results

In this section, we will verify the network training model with different strategies, and verify the theoretical grasping success rate of the model according to the success capture evaluation mechanism.

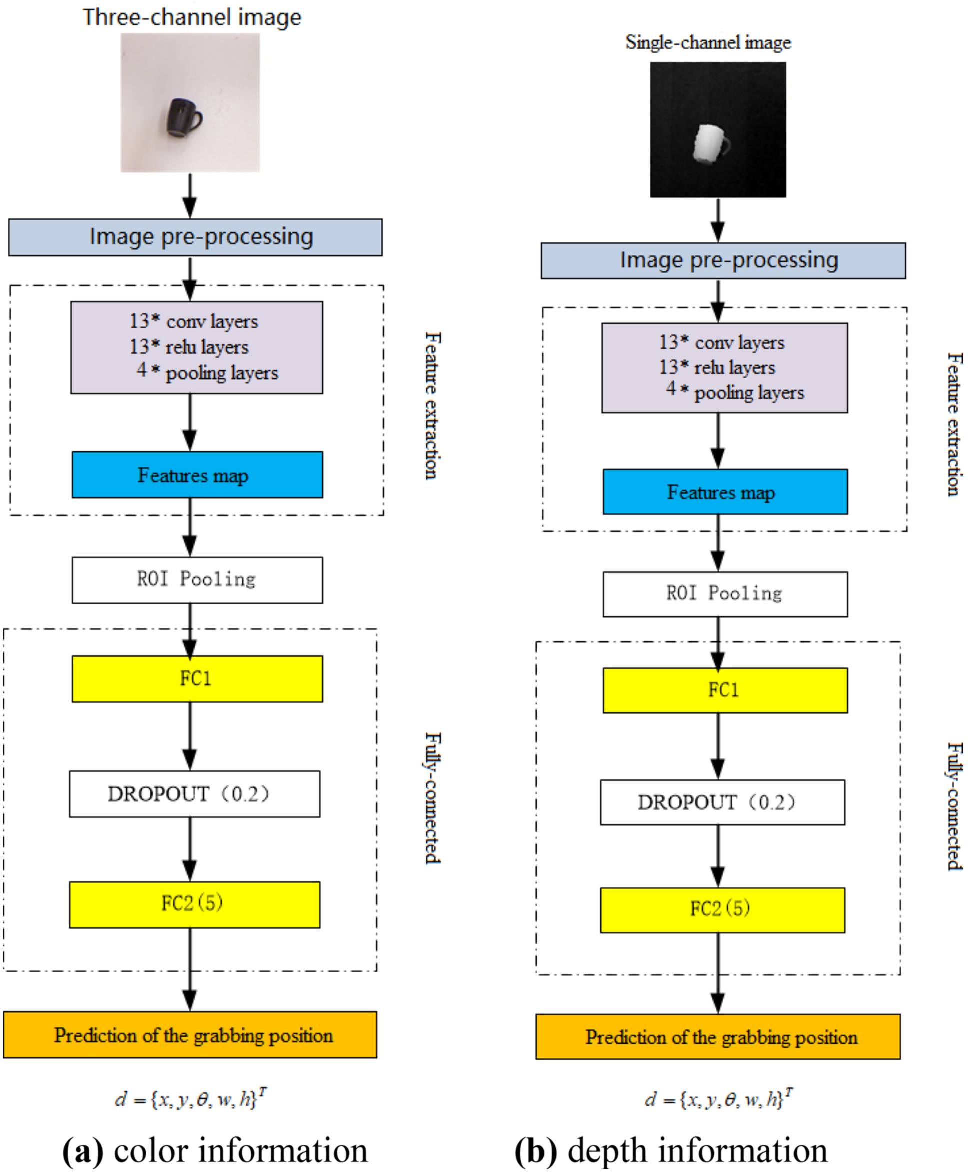
Because the successful prediction of the grasping rectangle is the key factor for the grasping stably with robot, the grasping angle and the area of the grasping area are generally used as the evaluation criteria. In the paper, the predicted grabbing rectangle is G and the real label rectangle is G*. If a predicted grabbing rectangle satisfies the following two conditions, it is considered that the predicted rectangle can be captured successfully.

(1) The angle difference between the predicted grabbing frame g and the real label is less than 300;

(2) The score value is less than 0.25.

$$score = \frac{area(G \cap G *)}{area(G \cap G *} \tag{6}$$

Fig. 7 Single modal sample input. **a** color information **b** depth information



5.1 Analysis of prediction results of grasping model

The multi-target test set is input into the grabbing detection model to verify the model detection effect. As shown in Table 1, in order to verify the influence of the size of the proposed network anchor on the model, the anchor scales were respectively set to (8, 16, 32), (4, 8, 16) and (4, 8, 16, 32).The training samples are all three channel color images, using VGG16 as a feature extraction network, in which, the training and test samples are segmented by multi-objective image set. For multi object images, firstly, all the grabbing rectangles with probability greater than 0.7 are predicted by the capture detection model, and then the highest probability grabbing rectangle corresponding to each category is obtained by background segmentation of color information as the top1 grabbing rectangle of the category, and the first five probability grabbing rectangles corresponding to each category are taken as the top5 grabbing rectangles of the category. From the first group and the second group of experimental data, the detection success rate of the model on top1 and top5 is improved with a smaller scale anchor value. From the second and third group of comparative experimental data, the retention of large-scale anchor value is conducive to the improvement of model grabbing detection success rate.

In order to verify the influence of different feature extraction networks on the success rate of model detection, the classic networks such as ResNet-50, MobileNet and VGG_16 were used to test the effect of different feature extraction networks. The training samples are all three channel color images, and anchor is set to 4, 8, 16, 32. The capture detection success rate (Table 2), and the model Top1

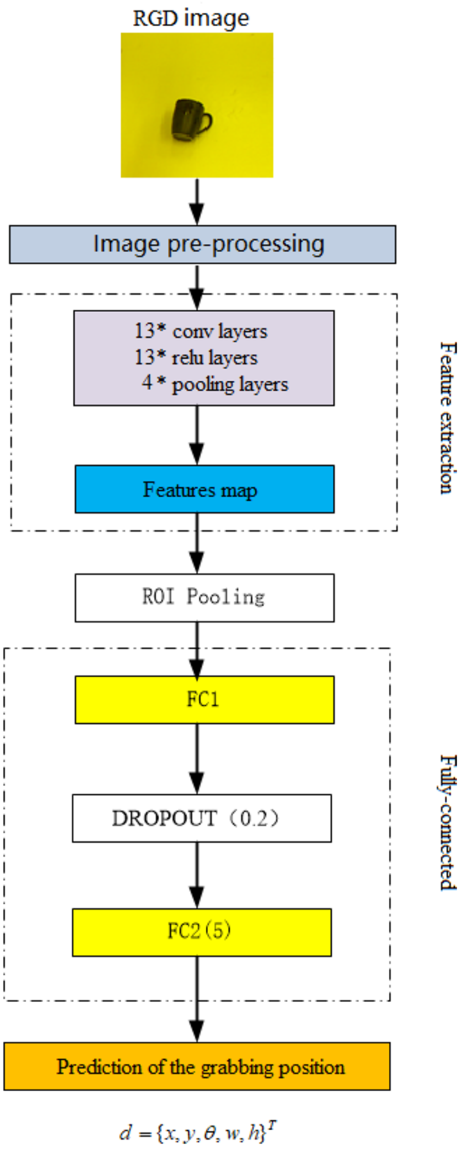


Fig. 8 Multimodal Fusion Strategy based on RGD Information

is taken as the final capture success rate. From the experimental results, we can see that resnet-50 has a better detection effect for the grab model, and its success rate on image wise and object wise is the highest.

In order to verify the influence of different modal input information on the success rate, four kinds of information (color, depth, RGD and depth-color multimodal) were used as input for feature learning, and test set Top1 grabbing rectangle box is taken as the final grabbing power. Anchor is set to (4, 8, 16, 32) in four kinds of grab detection models, and resnet-50 is used as feature extraction framework. The success rate of grab detection is obtained on the whole test set, as shown in Table 2. Five different types of objects are selected to test the detection effect of different models. The test set of RGD information and depth color multimodal model is three channel color image information, and the top-1 grabbing rectangle detection of five types of objects is shown in Fig. 11.

From the Top1 prediction of different modes, the rectangle detection model based on color and depth multi-mode is relatively better. The prediction effect of objects such as scissors through single-mode deep grasp detection model is not ideal, and the prediction effect of objects such as water cup through single-mode color grab detection model is not ideal, but based on color, depth and multi-mode, the prediction effect of the model is not ideal. The results show that the depth multi-modal grabbing rectangle detection model has better prediction results for these two kinds of objects.

It can be seen from Table 3 and Fig. 12 that the single channel depth information as a training sample has the worst result to the model. The model with color and depth input at the same time has a higher detection success rate than the model with color or depth information alone, which indicates that the model after the fusion of color and depth information has more powerful learning ability for features, and the model with color and depth input at the same time is easier convergence. At the same time, compared with the method of directly replacing B-channel data with depth

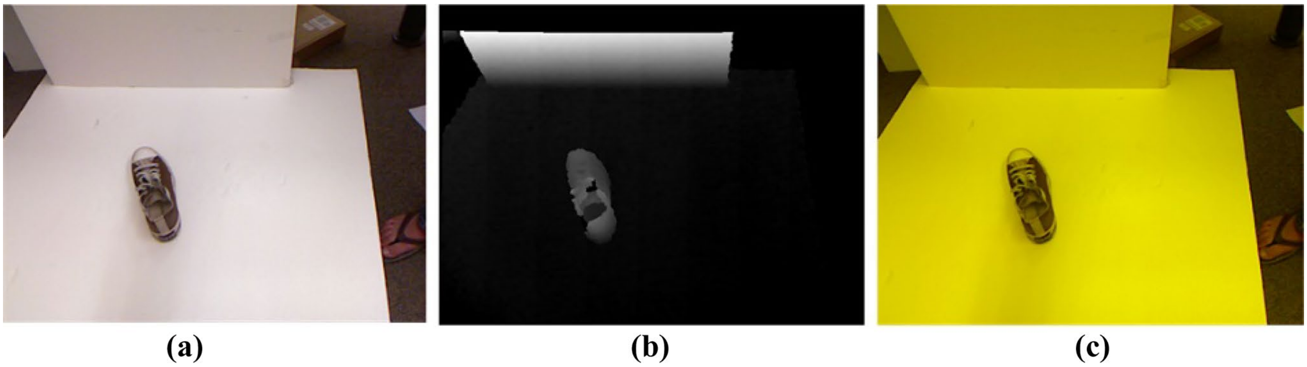


Fig. 9 Visualization of three different modal data

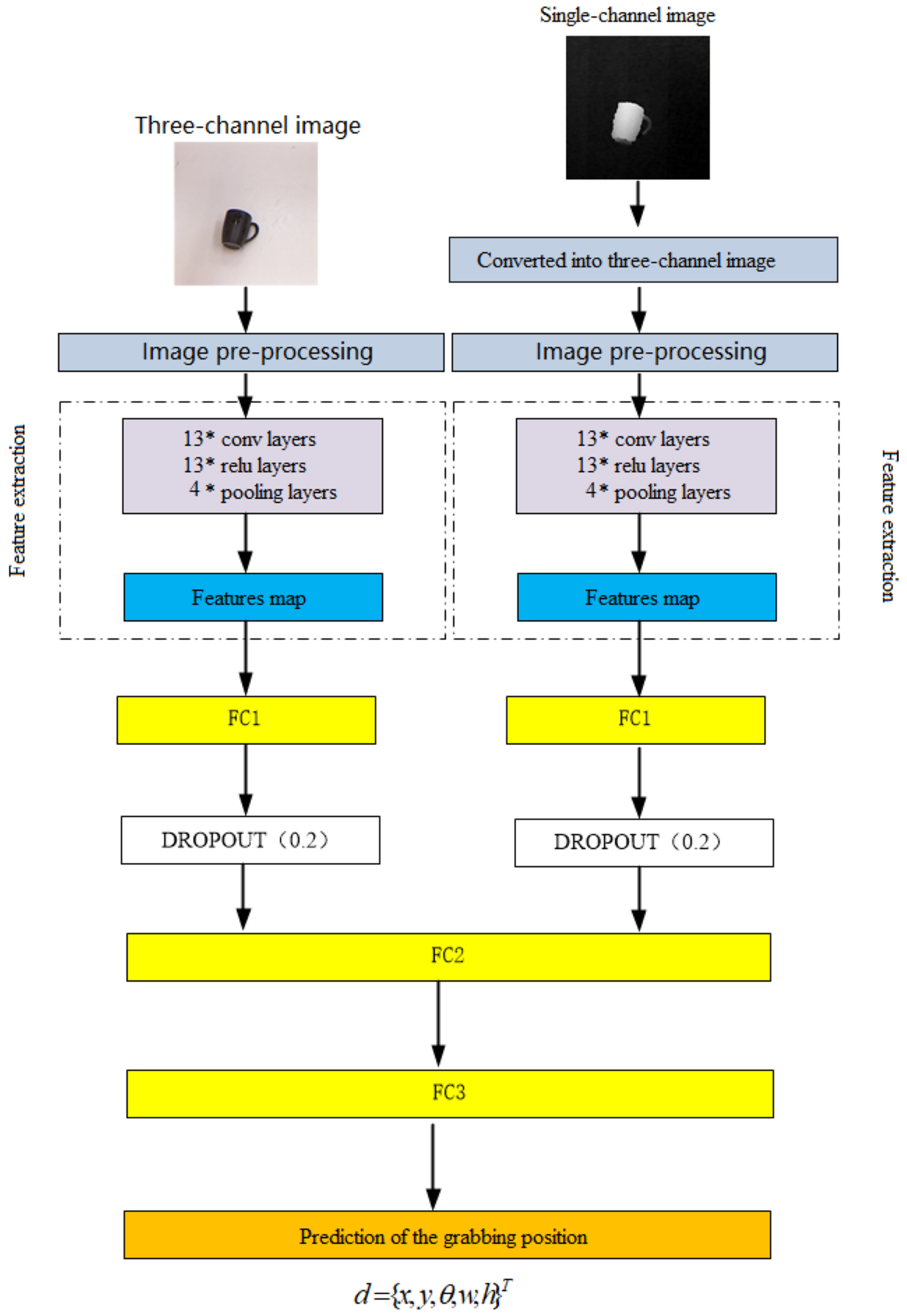


Fig. 10 Multimodal information fusion strategy

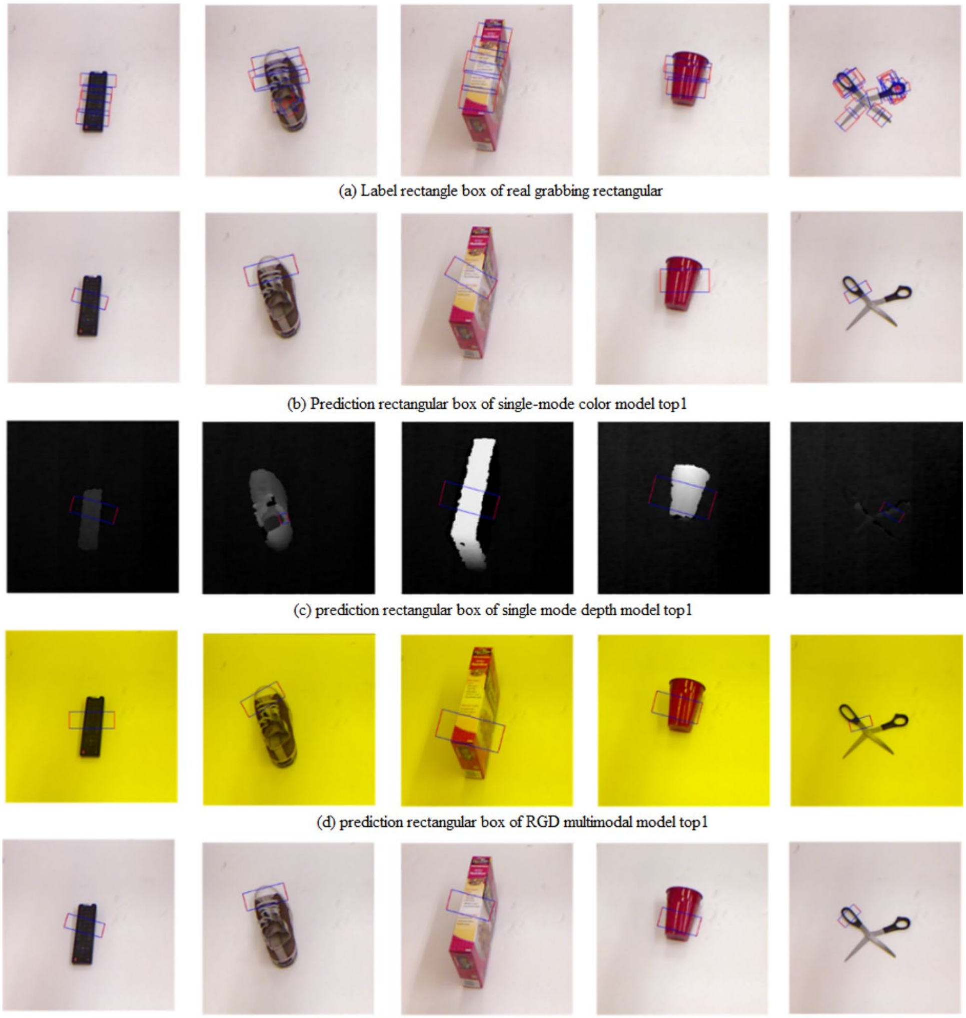
Table 1 Effect of different anchors on crawl detection results

Anchor scales	Top 1	Top 5
8,16,32	0.825	0.872
4,8,16	0.852	0.913
4,8,16,32	0.860	0.935

Table 2 Effect of different feature extraction networks on crawl detection results

Method	Split dataset based on imge	Split dataset based on object
	Success rate (%)	
VGG_16	0.860	0.823
MobileNet	0.842	0.831
ResNet-50	0.923	0.894

Fig. 11 Top1 prediction of five types of objects under different modals



(a) Label rectangle box of real grabbing rectangular

(b) Prediction rectangular box of single-mode color model top1

(c) prediction rectangular box of single mode depth model top1

(d) prediction rectangular box of RGD multimodal model top1

(e) Prediction rectangular box of color, deep multimodal model top1

Table 3 Effect of different input modals on crawl detection results

Method	Split dataset based on imge	Split dataset based on objec
	Success rate (%)	
Color	0.923	0.894
Depth	0.846	0.815
RGD	0.872	0.865
Multimodal	0.943	0.906

information to form RGD modal information for sample learning, the multimodal training method of two parallel neural networks is used to obtain higher detection accuracy.

This paper compares the grasping success rate and detection speed of the previous research models (Table 4).

Based on the multi-target data set, a grabbing rectangle detection framework with arbitrary angle is proposed. For small area grabbing frame detection, smaller anchor is used to achieve higher detection success rate; four different modes

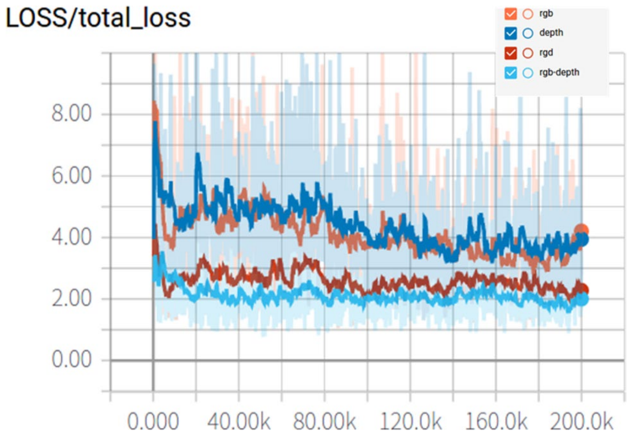


Fig. 12 The convergence of the loss value of the model under different modals

Table 4 Accuracy and speed of different detection models

Model	Split dataset based on imge	Split dataset based on objec	Speed
	Success rate (%)		fps
Lenz et al. (2013)	73.9	75.6	0.07
Wang et al. (2016)	81.8	–	7.10
Redmon et al. (2014)	88.0	87.1	3.31
Kumra et al. (2017)	89.2	88.9	16.03
Guo et al. (2017)	93.2	89.1	–
This paper	94.3	90.6	8.04

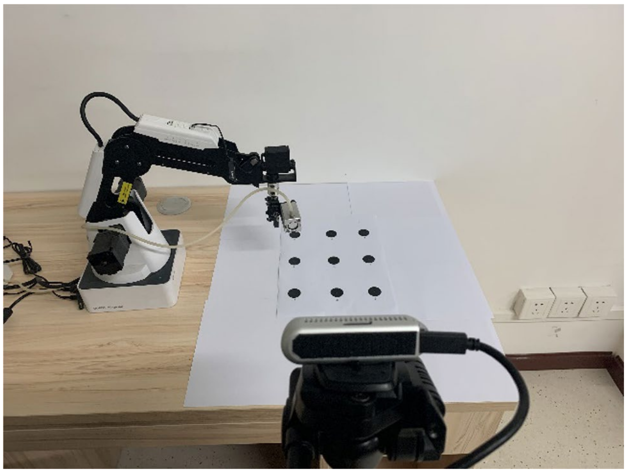


Fig. 13 Camera robot arm position diagram

of grabbing detection models are established, and the highest top-1 capture detection success rate of 94.3% is achieved in color and depth multimodal model which keeps good real-time performance.

5.2 The classification effect in the actual scene was verified and statistically analyzed

The grabbing experimental platform is composed of Dobot manipulator, ReaLsense d435 camera and computer. The position relationship between camera and manipulator is shown in Fig. 13. In the environment of ubuntu16.04 system, the color and depth images of the object to be grasped are obtained and preprocessed. The preprocessed images are input into the classification model and the grabbing rectangle detection model respectively, and the corresponding categories of objects and their optimal grabbing rectangle are output. The corresponding objects are captured through the coordinate position conversion between the manipulator and the camera.

Grabbing and classification experiments are based on RGB-D images, so the influence of illumination is great. In order to minimize the influence of light, the 20 groups of experiments are carried out in the same room at night, so that the illumination of each experiment is the same. For the object to be grasped, a single object is obtained by background segmentation of color information. As shown in Fig. 14, the RGB image of the target object is segmented, and then the image of the target object is input into the previously trained classification and grabbing detection model.

The quality of Top1 rectangle determines the feasibility of grasping parameters, and the quality of top5 rectangle determines whether the predicted rectangle is stable, as shown in Fig. 15a, b. Due to the load limitation of Dobot manipulator, some light objects are selected as the

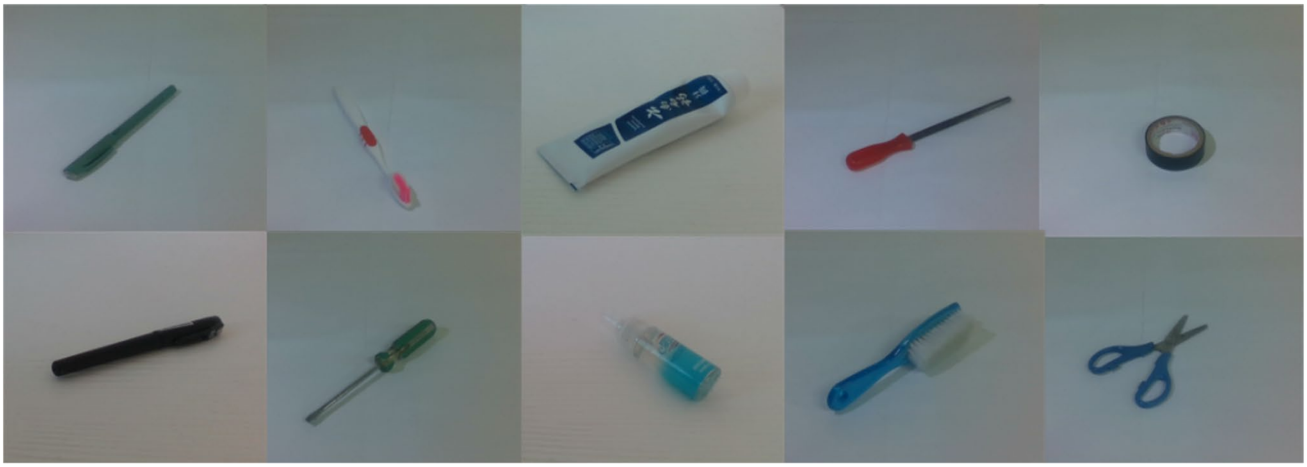


Fig. 14 Color-based background segmentation

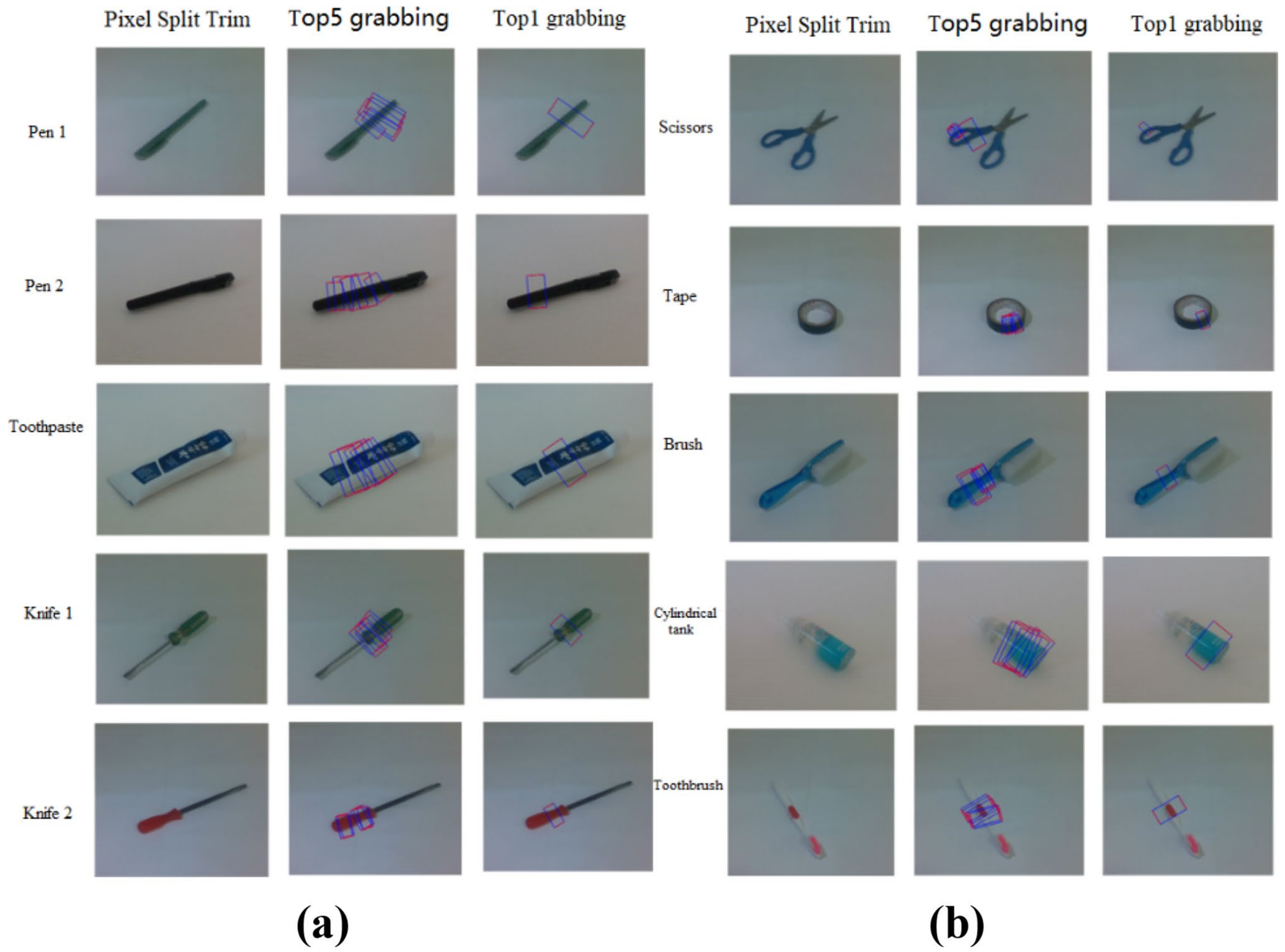


Fig. 15 Experimental Crawl results

Table 5 Crawl ingato-verification experiment results

Target object	Number of trials	The correct number of times	Number of crawl successes
Pen 1	20	18	18
Pen 2	20	19	16
Toothpaste	20	17	14
Knife 1	20	17	18
Knife 2	20	18	19
Scissors	20	20	14
Tape	20	19	14
Brush	20	18	17
Cylindrical tank	20	16	17
Toothbrush	20	19	17
Total	200	181	164

grasping objects in the experiment. From the distribution of the top 5 rectangular box, the learning rate of lighter, pen and other cylindrical or strip-shaped objects is better. And the result of the top 1 shows the top 5 prediction grabbing rectangle of the rectangle detection model is more inclined to grab the center of the object which is relatively stable on the strip and column objects. However, there will be some errors for some more complex objects such as scissors. In general, the prediction of the top 1 grabbing rectangle of the model is the most whose predicted phase success rate is higher.

According to the predicted optimal grabbing rectangle, the grab operation is performed on the grabbing experimental platform. In this paper, ten common objects are captured and classified for many times, in which, the success rate of grasping is 82.0%, as shown in Table 5. The experimental results show that the classification and grasping detection model proposed in this paper can complete the classification and grasping tasks well.

6 Conclusion

In order to overcome the mutual interference between objects and realize multi-target object detection in unstructured environment, a multi-target object grasping detection model is established in this paper. Based on the two-stage target detection network, the basic framework of grabbing detection model is constructed. In which, the angle variable is added to the anchor box generation of area suggestion network to realize the parameter expression of grabbing rectangle at any angle. Taking the rectangle as the learning feature, the four modes of grasping detection model are established. The capture detection model based on color and deep multimode fusion achieves 94.3%. Based on Dobot robot and ReaLsense camera, a grabbing verification platform is built to verify the feasibility of the model. In the actual environment, the capture success rate is 82.0%, which verifies the validity of the multi-target object capture detection model.

Acknowledgements This work was supported by grants of National Natural Science Foundation of China (Grant Nos.51575407, 51505349, 61733011, 41906177); the Grants of Hubei Provincial Department of Education (D20191105); the Grants of National Defense Pre-Research Foundation of Wuhan University of Science and Technology (GF201705) and Open Fund of the Key Laboratory for Metallurgical Equipment and Control of Ministry of Education in Wuhan University of Science and Technology (2018B07, MECOF2019B13).

References

Agrawal P, Nair A, Abbeel P (2016) Learning to poke by poking: experiential learning of intuitive physics. Adv Neural Inf Process Syst, p. 5074–5082

Billard A, Kragic D (2019) Trends and challenges in robot manipulation. Science 364(6446):eaat8414

Bohg J, Morales A et al (2014) Data-driven grasp synthesis—a survey. Robotics IEEE Trans Robot 30(2):289–309

Cai X, Niu Y, Geng S et al (2020a) An under-sampled software defect prediction method based on hybrid multi-objective cuckoo search. Concurr Comput Pract Exp 32(5):e5478

Cai X, Hu Z, Zhao P et al (2020b) A hybrid recommendation system with many-objective evolutionary algorithm. Expert Syst Appl 159:113648

Caldera S, Rassau A, Chai D (2018) Review of deep learning methods in robotic grasp detection. Multimodal Technol Interact 2(3):57–81

Chen T, Li Q, Yang J et al (2019) Modeling of the public opinion polarization process with the considerations of individual heterogeneity and dynamic conformity. Mathematics 7(10):917

Cheng W, Sun Y, Li G et al (2018) Jointly network: a network based on CNN and RBM for gesture recognition. Neural Comput Appl 31(Supplement 1):309–323. https://doi.org/10.1007/s00779-019-01268-3

Cheng Y, Li G, Yu M et al (2020a) Gesture recognition based on sEMG-feature image. Concurr Comput Pract Exp. https://doi.org/10.1002/CPE.6051

Cheng Y, Li G, Li J et al (2020b) Visualization of activated muscle area based on sEMG. J Intell Fuzzy Syst 38:2623–2634

Chu FJ, Xu R, Patricio V (2018) Real-world multi-object multi-grasp detection. IEEE Robot Autom Lett 3(4):3355–3362

Cui Z, Xue F, Zhang S et al (2020a) A hybrid BlockChain-based identity authentication scheme for multi-WSN. IEEE Trans Serv Comput 13:241–251

Cui Z, Zhang J, Wu D et al (2020b) Hybrid many-objective particle swarm optimization algorithm for green coal production problem. Inf Sci 518:256–271

Duan H, Sun Y, Cheng W et al (2020) Gesture recognition based on multi-modal feature weight. Concurr Computat Pract Exp. https://doi.org/10.1002/cpe.5991

Fontanelli D, Moro F, Rizano T et al (2014) Vision-based robust path reconstruction for robot control. IEEE Trans Instrum Meas 63(4):826–837

Guo D, Sun F, Liu H et al (2017) A hybrid deep architecture for robotic grasp detection. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), p. 1609–1614. https://doi.org/10.1109/ICRA.2017.7989191

Han J, Zhang D, Cheng G et al (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100

Hassan MU, Rehmani MH, Chen J et al (2020) Differential privacy techniques for cyber physical systems: a survey. IEEE Commun Surv Tutor 22(1):746–789

He Y, Li G, Liao Y et al (2019) Gesture recognition based on an improved local sparse representation classification algorithm. Clust Comput 22(Supplement 5):10935–10946. https://doi.org/10.1007/s10586-017-1237-1

Hsiao K, Chitta S, Ciocarlie M, Jones EG (2010) Contact-reactive grasping of objects with partial shape information. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, p. 1228–123

Huang L, Fu Q, Li G et al (2019) Improvement of maximum variance weight partitioning particle filter in urban computing and intelligence. IEEE Access 7:106527–106535. https://doi.org/10.1109/ACCESS.2019.2932144

Huang L, He M, Tan C et al (2020) Jointly network image processing: multi-task image semantic segmentation of indoor scene based on CNN. IET Image Process. https://doi.org/10.1049/iet-ipr.2020.0088

Jiang D, Li G, Sun Y et al (2019a) Grip strength forecast and rehabilitative guidance based on adaptive neural fuzzy inference system using sEMG. Pers Ubiquitous Comput. https://doi.org/10.1007/s00779-019-01268-3

Jiang D, Zheng Z, Li G et al (2019b) Gesture recognition based on binocular vision. Clust Comput 22(Supplement 6):2719–2729. https://doi.org/10.1007/s10586-018-1844-5

Jiang D, Li G, Sun Y et al (2019c) Gesture recognition based on skeletonization algorithm and CNN with ASL database. Multimed Tools Appl 78(21):29953–29970

Kumra S, Kanan C (2017) Robotic grasp detection using deep convolutional neural networks. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). https://doi.org/10.1109/IROS.2017.8202237

Lei Q, Chen G, Wisse M (2017) Fast grasping of unknown objects using principal component analysis. AIP Adv. https://doi.org/10.1063/1.4991996

Lenz I, Lee H, Saxena A (2013) Deep learning for detecting robotic grasps. Int J Robot Res 34:705–724

Li B, Sun Y, Li G et al (2017) Gesture recognition based on modified adaptive orthogonal matching pursuit algorithm. Clust Comput 22(Supplement 1):503–512. https://doi.org/10.1007/s10586-017-1231-7

Li G, Gan Y, Wu H et al (2018a) Cross-modal attentional context learning for RGB-D object detection. IEEE Trans Image Process 28(4):1591–1601

Li C, Li G, Jiang G et al (2018b) Surface EMG data aggregation processing for intelligent prosthetic action recognition. Neural Comput Appl 32(22):16795–16806. https://doi.org/10.1007/s00521-018-3909-z

Li G, Jiang D, Zhou Y et al (2019a) Human lesion detection method based on image information and brain signal. IEEE Access 7:11533–21154

Li J, Mi Y, Li G, Ju Z (2019b) CNN-based facial expression recognition from annotated RGB-D images for human-robot interaction. Int J Humanoid Robot 16(04):1941002

Li G, Li J, Ju Z et al (2019c) A novel feature extraction method for machine learning based on surface electromyography from healthy brain. Neural Comput Appl 31(12):9013–9022

Li G, Tang H, Sun Y et al (2019d) Hand gesture recognition based on convolution neural network. Clust Comput 22(Supplement 2):2719–2729. https://doi.org/10.1007/s10586-018-1844-5

Li G, Wu H, Jiang G et al (2019e) Dynamic gesture recognition in the internet of things. IEEE Access 7:23713–23724

Liao Y, Yu N, Tian D et al (2019) A quantized CNN-Based microfluidic lensless-sensing mobile blood-acquisition and analysis system. Sensors 19(23):5103

Liao S, Li G, Li J et al (2020a) Multi-object intergroup gesture recognition combined with fusion feature and KNN algorithm. J Intell Fuzzy Syst 38(3):2725–2735

Liao S, Li G, Wu H et al (2020b) Occlusion gesture recognition based on improved SSD. Concurr Comput Prac Exp. https://doi.org/10.1002/CPE.6063

Lin Y, Wei S, Fu L (2014) Grasping unknown objects using depth gradient feature with eye-in-hand RGB-D sensor. In: 2014 IEEE International Conference on Automation Science and Engineering (CASE), p. 1258–1263. https://doi.org/https://doi.org/10.1109/CoASE.2014.6899488

Lin D, Zhang R, Ji Y et al (2018) SCN: switchable context network for semantic segmentation of RGB-D images. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2018.2885062

Lin Y, Tang C, Chu F et al (2019) Using synthetic data and deep networks to recognize primitive shapes for object grasping. arXiv preprint arXiv:1909.08508

Lu L, Xie Z, Ye H (2016) Object recognition algorithm based on RGB feature and depth feature fusing. Comput Eng 42(5):186–193

Luo B, Sun Y, Li G et al (2020) Decomposition algorithm for depth image of human health posture based on brain health. Neural Comput Appl 32(10):6327–6342

Ma C, Chen L, Yong J (2019) AU R-CNN: encoding expert prior knowledge into R-CNN for action unit detection. Neurocomputing 335:35–47

Ma R, Zhang L, Li G et al (2020) Grasping force prediction based on sEMG signals. Alex Eng J 59(3):1135–1147

Mahler J, Liang J, Niyaz S et al (2017) Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312

Nie S, Meng Z, Qiang J (2018) The deep regression Bayesian network and its applications: probabilistic deep learning for computer vision. IEEE Signal Process Mag 35(1):101–111

Paolini R, Rodriguez A, Srinivasa SS et al (2014) A data-driven statistical framework for post-grasp manipulation. Int J Robot Res 33(4):600–615

Qi J, Jiang G, Li G et al (2019) Intelligent human-computer interaction based on surface EMG gesture recognition. IEEE Access 7:61378–61387

Qi J, Jiang G, Li G et al (2020) Surface EMG hand gesture recognition system based on PCA and GRNN. Neural Comput Appl 32(10):6343–6351

Redmon J, Angelova A (2014) Real-time grasp detection using convolutional neural networks. Proc IEEE Int Conf Robot Autom. https://doi.org/10.1109/ICRA.2015.7139361

Sangwan D, Jain DK (2019) An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. Pattern Recognit Lett 120:112–119

Schwarz M, Milan A, Periyasamy AS, Behnke S (2018) RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. Int J Robot Res 37(4–5):437–451

Sombolestan SM, Rasooli A, Khodaygan S (2019) Optimal path-planning for mobile robots to find a hidden target in an unknown environment based on machine learning. J Ambient Intell Humaniz Comput 10(3):1841–1850

Sun Y, Weng Y, Luo B et al (2020a) Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. IET Image Process. https://doi.org/10.1049/iet-ipr.2020.0148

Sun Y, Xu C, Li G et al (2020b) Intelligent human computer interaction based on non redundant EMG signal. Alex Eng J 59(3):1149–1157

Tan C, Sun Y, Li G et al (2019) Research on gesture recognition of smart data fusion features in the IoT. Neural Comput Appl 32(22):16917–16929. https://doi.org/10.1007/s00521-019-04023-0

Tian H, Wang C, Manocha D et al (2019) Transferring grasp configurations using active learning and local replanning. In: 2019 International Conference on Robotics and Automation (ICRA) 2290–2295

Tian J, Cheng W, Sun Y et al (2020) Gesture recognition based on multilevel multimodal feature fusion. J Intell Fuzzy Syst 38(3):2539–2550

Wang A, Lu J, Cai J et al (2015) Large-margin multi-modal deep learning for RGB-D object recognition. IEEE Trans Multimed 17(11):1887–1898

Wang Z, Li Z, Wang B, Liu H (2016) Robot grasp detection using multimodal deep convolutional neural networks. Adv Mech Eng 8(9):1–12

Wang P, Zhang X, Hao Y (2019) Journal of Sensors 2019:1–8

Wang P, Huang J, Cui Z et al (2020) A Gaussian error correction multi-objective positioning model with NSGA-II. Concurr Comput Pract Exp 32(5):e5464

Wen Z, Liu D, Liu X et al (2019) Deep learning based smart radar vision system for object recognition. J Ambient Intell Humaniz Comput 10(5):829–839

Weng Y, Sun Y, Jiang D et al (2020) Enhancement of grasp detection by cascaded deep convolutional neural networks. Concurr Comput Pract Exp. https://doi.org/10.1002/cpe.5976

Yu M, Li G, Jiang D et al (2019) Hand medical monitoring system based on machine learning and optimal EMG feature set. Pers Ubiquitous Comput. https://doi.org/10.1007/s00779-019-01285-2

Yu M, Li G, Jiang D et al (2020) Application of PSO-RBF neural network in gesture recognition of continuous surface EMG signals. J Intell Fuzzy Syst 38(3):2460–2480

Zhang Q, Song X, Yang Y et al (2018) Visual graph mining for graph matching. Comput Vis Image Underst 178:16–29

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.