A Robot Manipulator Grasping Method Based on Improved YOLOX

Yu Pan
College of Automation Engineering
Shanghai University of Electric Power
Shanghai, China
panyuyu@mail.shiep.edu.cn

Fei Xia
College of Automation Engineering
Shanghai University of Electric Power
Shanghai, China
xiafei@shiep.edu.cn

Jianliang Mao
College of Automation Engineering
Shanghai University of Electric Power
Shanghai, China
jl_mao@shiep.edu.cn

***Abstract*—The overlap and coverage of objects can affect the grasping success rate for robot manipulator grasping in multi-object scenarios. We propose an enhanced grasping algorithm based on YOLOX-that can predict the bounding box with a smaller aspect ratio, thereby more accurate spatial location. Due to the limits of sensor's environment and physical factors, the depth map will lose some depth values. We propose a depth value repair algorithm based on the FMM algorithm, through which the lost depth values in the grasping region can be repaired. In pose estimation, we use the aspect ratio of the bounding box to determine the rotation angle of the robot manipulator jaws. We use a six-axis robot manipulator combined with a depth camera to achieve object grasping in multi-object scenes. The experimental results show that the enhanced grasping algorithm makes the grasping area prediction more accurate, and the distance between the object and the camera is obtained more accurately.**

Keywords—yolox, object detection, pose estimation, robot manipulator grasping

I. INTRODUCTION

Robot manipulator grasping technology has a crucial role for robots, as one of the most common and fundamental skills, robot manipulator grasping has been extensively studied in recent decades [1]. In recent years, object detection accuracy has been dramatically improved with the continuous development of neural convolutional networks and sensors. For multi-object scenarios, the stacked coverage between objects and the influence of the environment can affect the robot manipulator grasping success rate.

As shown in Fig 1, we use the improved YOLOX[2] algorithm to predict the object grasp area, and the depth camera collects the depth map, the bounding box's center point depth value is calculated, the Zhang Zheng You calibration method is used to complete the robot manipulator hand-eye calibration[3]. In the end, the three-dimensional spatial information of the object is provided to the robot manipulator to achieve object grasping.

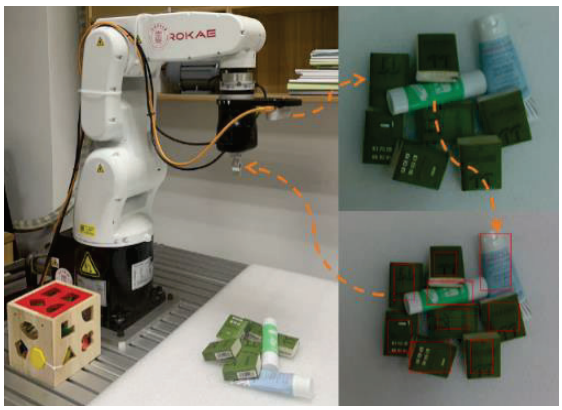


Fig. 1. Flowchart of robot manipulator grasping system

II. RELATED WORK

For traditional robot manipulator grasping, the robot manipulator operation is guided by manual demonstration [4], and this approach suffers from low grasping accuracy and low grasping efficiency in practical tasks. A two-step approach is taken for the input RGB-D images by roughly selecting the possible grasp pose and using the object's color and edge information as the grasp feature [5]. To reduce the operational efficiency of the algorithm, the two-step approach introduces a neural network that transforms the 3D problem into a 2D problem [6]. For machine vision robot manipulator grasp, the detection speed is slow and easily affected by the environment, making it difficult to accurately predict the object's contour, such as light, and the object's reflection, which cannot achieve real-time detection.

As deep convolutional networks evolve, using convolutional neural networks to extract grasping features is a general research topic, from the original AlexNet network transforming the grasping problem into an end-to-end issue [7], and subsequently using ResNet instead of AlexNet network for end-to-end prediction [8]. The advent of the YOLO object detection algorithm [9-12] and SSD object detection algorithm [13] has led to the rapid development of object detection, which uses the concept of the sliding window to chunk the image, predicts the presence of graspable objects in small chunks of images by a classifier, and outputs the optimal grasp pose [14]. However,

978-1-6654-9899-9/22/$31.00 ©2022 IEEE

convolutional neural networks are prone to false detection in predicting the grasp pose for multi-object scenes, such as object stacking and the influence of the environment and sensors, and cannot achieve real-time grasping. This paper proposes an improved YOLOX algorithm for robot grasping with the following main elements:

1) *A modified-YOLOX algorithm incorporating the bounding box prediction method is put forward, which has a high overlap rate between the bounding box and the object.*
2) *For the lost depth values in the bounding box region, we use the FMM[15] algorithm to restore regions with lost depth values adaptively.*
3) *By using the aspect ratio of the bounding box to select the optimal grasping pose, a better grasping success rate can be achieved.*

III. System framework and algorithmic process

Fig 2 depicts the robot manipulator grasping system's experimental framework and algorithm flow. The system adopts the "eye in hand" camera calibration method and attaches the depth camera to the two-finger jaw. The grasping process can be divided into object detection, pose estimation and robot manipulator grasping planning.

Initially, the depth camera captured the RGB-D images and input into the neural network, which outputs the class and spatial location of the object. Because of the effect of light and sensors, there is noise in the depth map, we use the FMM algorithm to restore regions with lost depth values. Finally, the object's spatial position is transmitted to the robot manipulator to implement grasping.

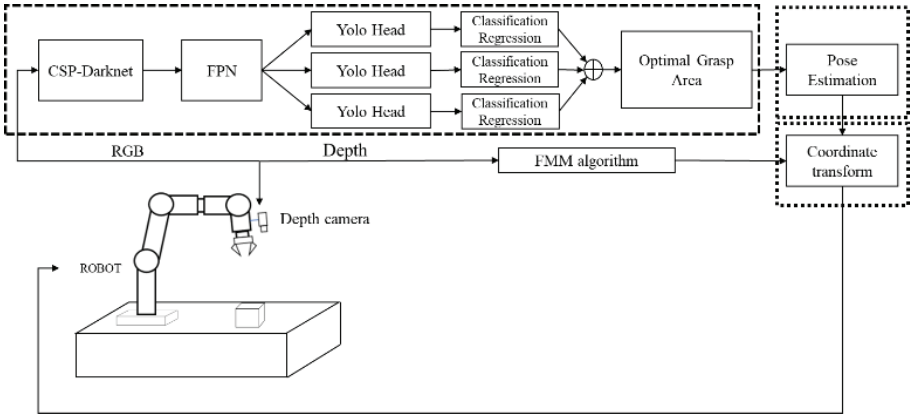


Fig. 2. Experimental framework and algorithm flow

IV. Object detection algorithms

A. Improved YOLOX Object Detection Model

In contrast to other YOLO algorithms, in the YOLO Head session in YOLOV2-V4, classification and regression are implemented in a 1×1 convolution layer. YOLOX differs in that classification and regression are performed independently, and the predictions are finally combined.

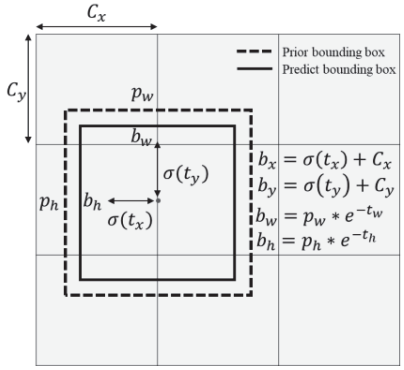


Fig. 3. Prediction bounding box

In order to change the size of the prediction bounding box, in the network architecture of YOLOX, three scaled feature layers are obtained by FPN (1024, 512, 256), using a 1x1 convolution layer to reduce the dimensionality, two 3x3 convolution layers are used for classification and regression. Each YOLO Head part will predict 4 coordinates for each bounding box t_x, t_y, t_w, t_h, where t_x, t_y is the center of the bounding box, t_w, t_h is the width and height. When indicating the bounding box size at the network's end, the prediction of the bounding box is shown in Fig 3, we use e^{-x} expected for the predicted bounding box aspect ratio, where b_x, b_y are the coordinates of the center point of the predicted bounding box, and b_w, b_h is the height and width. If the cell is offset from the top left corner of the image by (C_x, C_y) and the bounding box prior has width and height p_w, p_h, then the predictions correspond to:

$$b_x = \sigma(t_x) + C_x \quad (1)$$

$$b_y = \sigma(t_y) + C_y \quad (2)$$

$$b_w = p_w * e^{-t_w} \quad (3)$$

$$b_h = p_h * e^{-t_h} \quad (4)$$

One can refer to the improved YOLOX-YOLO Head network structure in Fig 4, we use the new bounding box prediction method to replace the original bounding box width and height. This prediction bounding box of the method will be smaller than the original YOLOX-YOLO Head, which can effectively reduce the predicted object bounding box size and improve the robot manipulator grasping success rate.

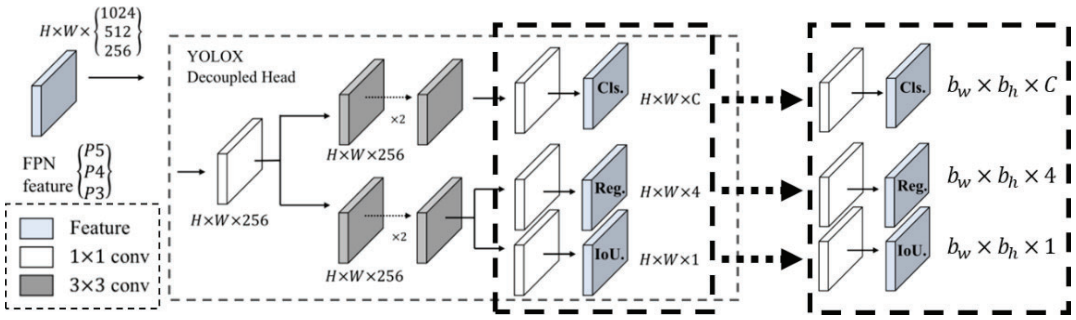


Fig. 4. Improved YOLOX YOLO-Head network structure

B. Depth Value Restoration

The depth map acquired by the depth camera will contain some noise, the traditional filtering noise algorithm is inadequate for noisy images. We use the FMM algorithm as the depth value repair algorithm for repairing the lost depth values in the centroid region Ω, use (5) to determine whether region Ω contains noise, let M be the mask matrix with the same size as the depth map D. Define (i,j) as the pixel coordinates, where α is the empirically obtained segmentation threshold, and M(i,j)=0 means that the depth value of the pixel point is not available.

$$M(i,j) = \begin{cases} 0, D(i,j) \le \alpha \\ 1, D(i,j) \ge \alpha \end{cases} \tag{5}$$

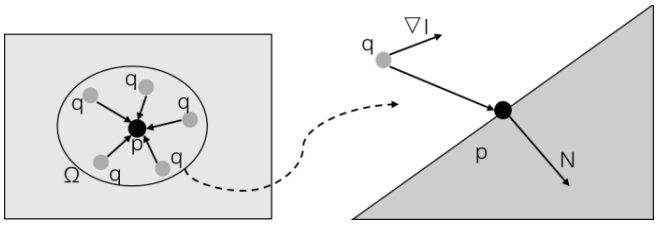


Fig. 5. FMM algorithm

Fig 5 shows that point p is the depth value to be repaired in the region Ω with p as the center, and the depth values of other pixel points in this region are known. The effect of improving the depth value is achieved by calculating a new depth value to replace the lost value, q is a point in Ω, and the grayscale value of p is calculated from point q according to the following equation:

$$I_q(p) = I(q) + \nabla I(q)(p-q) \tag{6}$$

Since each pixel point plays a different role, the weight function $w(p,q)$ is used to rank the influence size of the pixel points around the point to be repaired, using the following equation:

$$I(p) = \frac{\sum_{q\in\Omega} w(p,q)[I(q)+\nabla I(q)(p-q)]}{\sum_{q\in\Omega} w(p,q)} \tag{7}$$

$$w(p,q) = dir(p,q) * dst(p,q) * lev(p,q) \tag{8}$$

$$dir(p,q) = \frac{p-q}{||p-a||} * N(p) \tag{9}$$

$$dst(p,q) = \frac{d_0^2}{||p-q||^2} \tag{10}$$

$$lev(p,q) = \frac{T_0}{1+|T(p)-T(q)|} \tag{11}$$

Where (8) is the weight formula, (9) is the direction factor, which determines that the pixel point closer to the standard direction $N = \nabla T$ contributes the most to point p, (10) is the geometric distance factor, which determines that the closer to pixel point p the more significant the contribution, and (11) is the horizontal distance factor, which specifies that the pixel point closer to the contour line of the central region contributes more to point p. From Fig 6, the depth map repair structure and the process can be recognized.

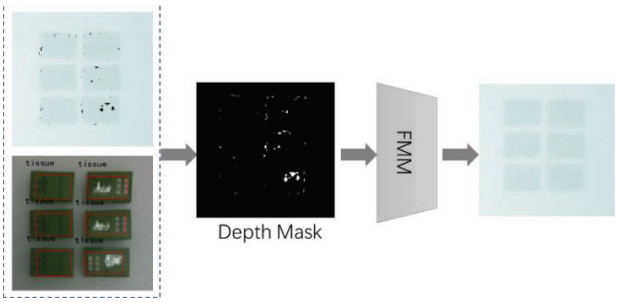


Fig. 6. Framework of depth image restoration

V. Grasp Pose Estimation

This paper classifies the object's pose into three types, horizontal, inclined, and vertical. We use the aspect ratio of the prediction bounding box for pose estimation. In a vertical posture, the length and width of the prediction bounding box are W and H. Define R as the aspect ratio of the prediction bounding box, and use the following equation:

$$R = \frac{H}{W}, Angle \in [0,90] \tag{12}$$

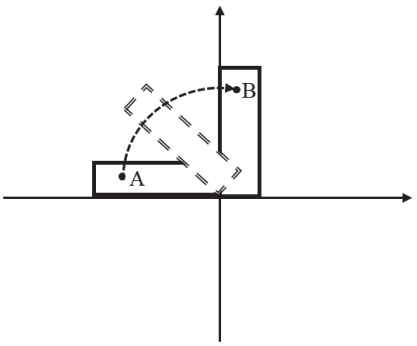


Fig. 7. Object pose change

As Fig 7 shows, when the object is transformed from point A to B, the width and height of the prediction bounding box will change, then R will be changed. From Table I, we can select the grasping pose by determining the R-value.

TABLE I. POSTURE CORRESPONDENCE

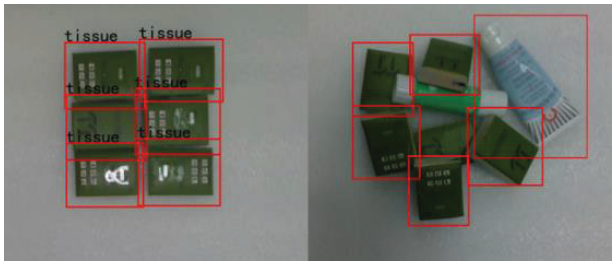
R	0.5	0.6	0.8	……	1.6	1.8
Angle	0°	10°	20°	……	80°	90°

VI. EXPERIMENT AND ANALYSIS

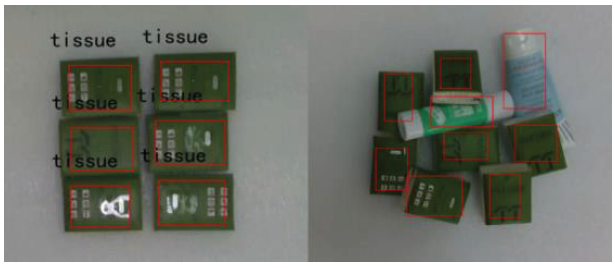
The robot manipulator grasp experiment was conducted in Ubuntu 18.04, using a Realsense-D435 depth camera and a six-axis robot manipulator. To build our dataset, the investigation is based on the PASCAL VOC dataset format, and six types of images (Tissue, Cup, Stick, Glue, Cream, Cola) are collected. 2000 images of each type were gathered, 1500 of which were used for training and 500 for testing. The GPU-GTX-2080Ti was used to accelerate the network during training, freezing the training first and thawing it later could speed up the training. The required parameters are as follows:

1) *Learning rate:1e-3.*
2) *Label smoothing:0.01.*
3) *Freeze training iterations:50.*
4) *Batch size:12.*

With the trained model deployed on the server and tested, the improved YOLOX model is compared with the original YOLOX model. As shown in Figure 8, the experimental results show that the improved YOLOX model can effectively reduce the overlap rate between the bounding box and the object. The improved model is more suitable for object detection under complex multi-object and can obtain the object position more accurately.



(a)



(b)

Fig. 8. Comparison of object detection model predictions: (a) Original YOLOX object detection. (b) Improved YOLOX object detection.

We simultaneously trained RCNN, SSD, and YOLO series networks and compared the model performance, the performance comparison of different networks is shown in Table II. The improved YOLOX is about 3% better than the YOLOX, in the real-time test, the improved YOLOX reached 28 FPS, an improvement of 8 FPS compare to YOLOX in terms of accuracy and can satisfy the requirements for real-time performance in the grasping task.

TABLE II. OBJECT DETECTION MODEL PERFORMANCE COMPARISON

Object Detection Network	**mAP**	**FPS**	**Boxes**
Improved YOLOX	*92.3*	*28*	*1405(False:24)*
YOLOX	*89.5*	*20*	*1405(False:32)*
YOLOV4	*86.4*	*21*	*1405(False:38)*
YOLOV3	*82.6*	*18*	*1405(False:52)*
SSD-512	*79.5*	*19*	*1405(False:48)*
Faster-RCNN(VGG16)	*73.2*	*14*	*1405(False:86)*

Finally, to illustrate the algorithm's effectiveness in this paper, we simulated the recognition grasping in a multi-object scene with different objects with random poses and positions. With stacking and covering between objects, as shown in Fig 9, the improved YOLOX can accurately classify and locate various objects. From Table III, we see that the grasp success rate of tissue is 0.86, and cream is 0.92.

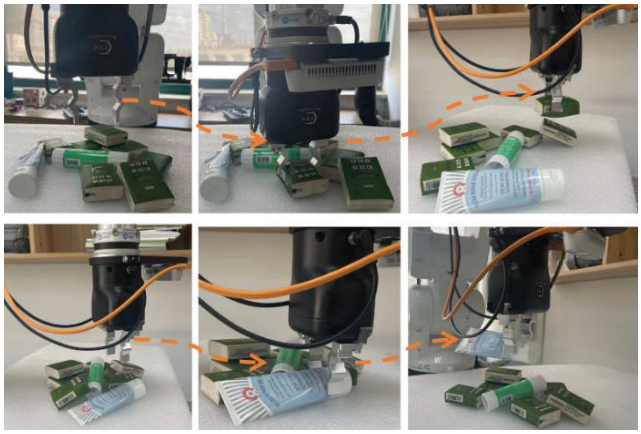


Fig. 9. Robot manipulator real grasp

TABLE III. GRASPING THE RESULT OF DIFFERENT OBJECTS

Object	**Average number**	**Successful grasp**	**Success rate**
Tissue	14	12	0.86
Cream	12	11	0.92
Stick	9	8	0.88
Cola	10	8	0.8

VII. CONCLUSION

We propose an improved YOLOX algorithm for grasping in multi-object scenes. It can predict the bounding box with a smaller aspect ratio and indicate the location of the grasping area more accurately. The FMM-based depth value repair algorithm can effectively repair the lost depth value, and accurately obtain the distance between the object and the camera. The object's pose is divided into three poses using the bounding box's aspect ratio to adjust the rotation angle of the two-finger jaw.

In the experimental part, the improved YOLOX can reach 28 FPS in NVIDIA-2080Ti with an average accuracy of 92.3%. The experimental results demonstrate that our proposed method improves the accuracy of object recognition and grasping. The disadvantage is that there is a specific error in the predicted angle for object pose estimation, 6D pose estimation is now widely used in robot manipulator grasping. Our future research is based on this method to improve objects' pose estimation and image processing speed.

REFERENCES

[1] B. Jeannette, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis-A survey," IEEE Transactions on Robotics, vol. 30, no. 2, pp. 289-309, 2014.

[2] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, and M. Technology, "YOLOX: exceeding YOLO series in 2021," arXiv: 2107.08430v2 [cs.CV], 2021.

[3] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, 2000.

[4] R. Balasubramanian, L. Xu, P. D. Brook, J. R. Smith, and Y. Matsuoka, "Physical human interactive guidance: Identifying grasping principles from human-planned grasps," IEEE Transactions on Robotics, vol. 28, no. 4, pp. 899-910, 2012.

[5] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 3304-3311, 2011.

[6] H. Lin and H. Chu, "Robotic grasp detection by rotation region CNN," 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), pp. 1-5, 2021.

[7] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 1316-1322, 2015.

[8] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 769-776, 2017.

[9] 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, 2016.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525, 2017.

[11] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[12] A. Bochkovskiy, C. Wang, H. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single shot multibox detector," European Conference on Computer Vision (ECCV), pp. 21-37, 2016.

[14] S. Kumra, C. Kanan, "Robotic grasp detection using deep convolutional neural networks," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 769-776, 2017.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.