Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

# A Vision-based Robotic Grasping System Using Deep Learning for Garbage Sorting

Chen Zhihong[1], Zou Hebin[1], Wang Yanbo[1], Liang Binyan[1], Liao Yu[1]

1. Beijing institute of precision mechanical and electrical control equipment, Beijing, 100076
E-mail: zhihongch@outlook.com

**Abstract:** This paper proposes a robotic grasping system for automatically sorting garbage based on machine vision. This system achieves the identification and positioning of target objects in complex background before using manipulator to automatically grab the sorting objects. The object identification in complex background is the key problem that machine vision algorithm is trying to solve. This paper uses the deep learning method to achieve the authenticity identification of target object in complex background. In order to achieve the accurate grabbing of target object, we apply the Region Proposal Generation (RPN) and the VGG-16 model for object recognition and pose estimation. The machine vision system sends the information of the geometric centre coordinates and the angle of the long side of the target object to the manipulator which completes the classification and grabbing of the target object. The results of sorting experiment of the bottles in the garbage show that the vision algorithm and the manipulator control method of the proposed system can achieve the garbage sorting efficiently.

**Key Words:** Machine vision, Complex backgrounds, Deep Learning, Robotic grasping, Garbage sorting

## 1 Introduction

The classification means putting garbage into different cans so that they can become new resources through different ways of cleaning, shipping and recycling. However, the environment has continued to worsen with problems related to atmospheric pollution, ecological disasters, water pollution and cities besieged by garbage. In this situation, classification and collection of refuse have been an urgent affair and imperative trend. Nowadays, the drawback of low efficiency for the manual sorting of garbage needs to be improved, being replaced by robotic grasping system. Robotic grasping is extremely difficult because of unknown objects and poses. Robust object recognition and pose estimation are the key problems for the automatic-grasping task of robotics. Canny operator is a method that gets the edge of image and segments it by using image edge detection [1]. In the case of the object being in the complex background, recognition algorithms cannot be applied to garbage sorting. Alvaro Collet and Manuel Martinez achieved robust object recognition and pose estimation with MOPED, a framework for Multiple Object Pose Estimation and Detection [2]. But the object recognition method has difficulty in the model building stage. Nowadays, more and more approaches are presented for robotic grasping. Ellen Klingbeil proposed a novel algorithm for grasping unknown objects given raw depth data obtained from a single frame of a 3D sensor [3]. Kai Huebner and Danica Kragic proposed an algorithm that efficiently wraps given 3D data points of an object into primitive box shapes by a fit-and-split algorithm based on Minimum Volume Bounding Boxes [4]. But above methods cannot recognize the object which is covered by unknown substance.

In this paper, we propose a new system which can not only recognize objects but estimate their poses by using deep neural network model. Region Proposal Network (RPN) and the VGG-16 model [5] are optimized to the object detection. In the case of the object being covered by unknown substance over 30 percent coverage, experiments have been made to demonstrate the practicality, effectiveness and superiority of the proposed optimization algorithm. The information of the geometric centre coordinates and the angle of the long side of the target object is sent to the S7-300 PLC to distribute to the KUKA robotic arm by high-speed field-bus ( profibus ).

### 1.1 Instructions for Authors

Chen zhihong, Zou Hebin，Wang Yanbo, Liang Binyan and Liao Yu are with Beijing institute of precision mechanical and electrical control equipment. Beijing. China. Our professional work deals with the cutting edge of robot technology and machine vision technique.

## 2 System Description

As the Fig. 1 shown, the robotic grasping system has three components: the camera used for image collection, the conveyor belt used for objects transporting, and the manipulators for object grasping. The camera is installed in the confined space with artificial light to eliminate the environmental impact and obtain stable images. The number of the robotic arms are more than one to implement cooperating grasp tasks.

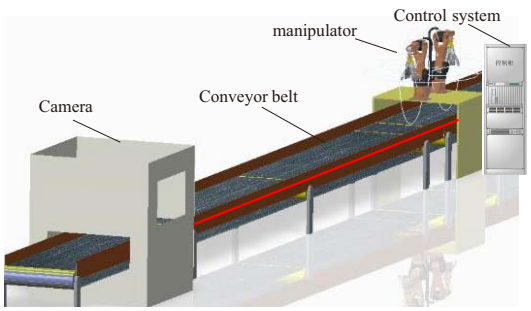


Fig. 1: The sketch of the grasping system

*This work is supported by the Industry Foundation of China Academy of Launch Vehicle Technology.

The data flow chart of system is illustrated in Fig. 2. The image of the moving object is captured by the CCD camera, and the data of the image is transferred by GigE to the computer to process. Serial communication bus is employed to transfer the position and orientation information. The real-time controller of S7-300 distributes the data to the robotic arms. The KUKA robotic arms are compatible for the Profibus.

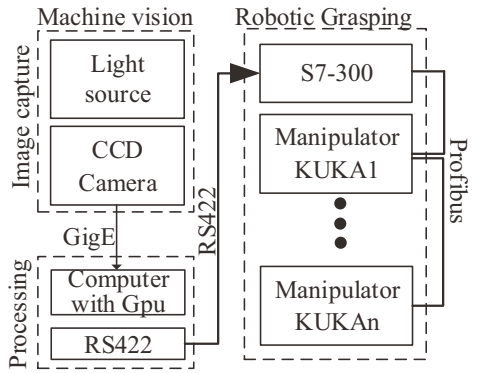


Fig. 2: Data flow chart of system

## 3 Object Recognition and Pose Estimation

### 3.1 Brief Introduction of RPN

Fast R-CNN achieves near real-time rates using very deep networks. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy[6]. Fast R-CNN is composed of two sub-nets: Region Proposal Generation (RPN) and VGG-16. RPN shares full-image convolutional features with the detection network, which is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position.

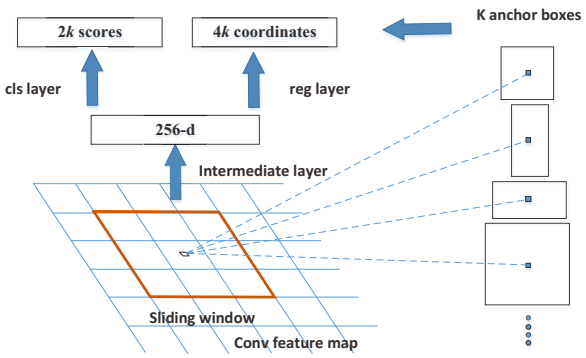


Fig. 3: The Region Proposal Network (RPN)

### 3.2 Optimization

The estimation of the angle of the long side of the target is connected to the VGG-16, after the FC. Here, we minimize an objective function following the multi-task loss in Fast-RCNN [6]. Our loss function for an image is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i, t_i^*\right) \quad (1)$$

Here, $i$ is the index of an anchor in a mini-batch and $p_i$ is the predicted probability of anchor $i$ being an object. The ground-truth label $p_i^*$ is 1 if the anchor is positive, and is 0 if the anchor is negative. The $t_i$ is a five-dimensional vector, which is the coordinates of the object's bounding box. $t_i^*$ is that of the ground-truth box associated with appositive anchor.

For the object's bounding box, the parameterization methods of R-CNN is employed. The normalization is showed as (2).

$$\begin{aligned} t_x &= (x - x_a)/w_a,\ t_y = (y - y_a)/h_a, \\ t_x^* &= (x^* - x_a)/w_a,\ t_y^* = (y^* - y_a)/h_a, \\ t_w &= \log(w/w_a),\ \ t_h = \log(h/h_a), \\ t_w^* &= \log(w^*/w_a),\ \ t_h = \log(h^*/h_a) \\ t_p &= t/2\pi,\ t_p^* = t^*/2\pi \end{aligned} \quad (2)$$

$x$ and $y$ is the coordinates of the bounding box, $w$ and $h$ is the width and length of the bounding box. The network is illustrated in Fig. 4. The modified VGG-16 net. 'Conv' and 'FC' represent the convolution and full connected layers respectively. We also omit the 'ReLU' and 'Max Pooling' layer after each 'Conv' operation for presentation clarity. The modified VGG16-net decouples the angle loss and original detection loss (including both classification and regression loss) for improving the performance.

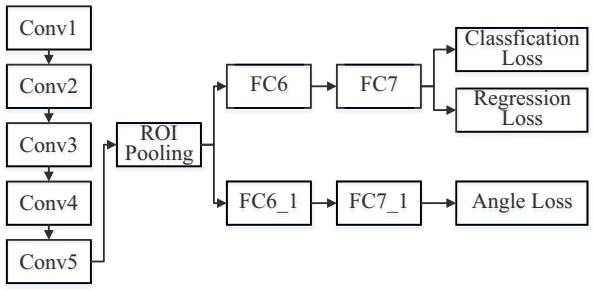


Fig.4: The modified VGG-16 net

### 3.3 Removing repeating identification

The target object is moving on the conveyor belt, and a target occurs multiple times in a series of continuous images. The method of removing the data of the repeating identification can be considered as the abstraction of the useful data.

The abstraction of the useful data is illustrated in Fig. 5. The target object can be classified into two groups: one is that just getting into the vision of the camera or just getting out of the vision, like ① and ③; the other is that all in the vision of the camera, like ②. We assume that the times of the same target a occurs in the image is $N_c$.

$$N_C = \frac{S_Y - 2L_{\max}}{V_C} \times F_{\text{ps}} \tag{3}$$

$V_c$ is the speed of the conveyor belt, $F_{\text{PS}}$ is the frame rates of the camera, $L_{\max}$ is the max length of the target along the axis-Y direction. The abstraction of the useful data is that occurs $N$c times will be recorded.

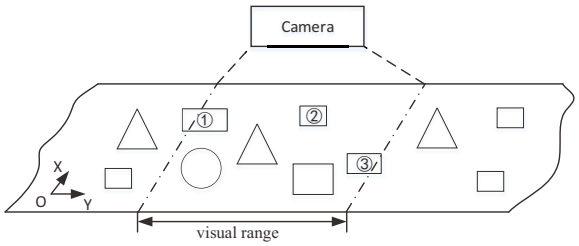


Fig. 5: The abstraction of the useful data

## 4 Robotic grasping control

The key problem of the grasping control is to distribute the target information to the robotic arms. The scheduling module for multiple robotic arms cooperative control is S7-300 which is built by Siemens. KUKA arms are employed to be the executives, which implement the expectant motion locus. Profibus is a kind of serial real-time communication. Originally developed to meet the needs of the discrete manufacturing industry, PROFIBUS has evolved into a fieldbus standard for both discrete manufacturing and process automation. The protocol that transmits the coordinates, the attitude angle and the real time clock. The information of the target object is transmitted to the robotic arm controller, then the motion locus. The control process causes a delay, $t_2 + \Delta T$, $\Delta T$ is the time consuming of motion locus. The tracking coordinate of feature point is $\left(x_{C\_target}, y_{C\_target}\right)$, which is the targeting coordinate.

$$x_{c\_target} = x_{c\_0} + v \times (t_2 + \Delta T - t_1) \tag{4}$$

$$\Delta T = \frac{\sqrt{(x_{R\_target} - x_{R\_int})^2 + (y_{R\_target} - y_{R\_int})^2 + (z_{R\_target} - Z_{R\_int})^2}}{V_{R\_hand}} \tag{5}$$

$$\begin{bmatrix} x_{R\_target} \\ y_{R\_target} \\ 1 \end{bmatrix} = \mathbf{M}_{C\_R\_i} \bullet \begin{bmatrix} x_{C\_target} \\ y_{C\_target} \\ 1 \end{bmatrix} \tag{6}$$

$$Z_{R\_target} = 0 \tag{7}$$

$$y_{C\_target} = y_{C\_0} \tag{8}$$

Where $\left(x_{R\_int}, y_{R\_int}, z_{R\_int}\right)$ is the starting position coordinate of the $i$th robotic arm. $V_{R\_hand}$ is the Point-to-Point moving speed of the end- effector. $v$ is the speed of the conveyor belt. $\mathbf{M}_{C\_R\_i}$ is the transformation matrix equation for transforming the belt- frame to the tool-frame. $t_2$ is the current time.

The scheduling module can judge whether the target is in the workspace of the $i$th robot arm according to the following two stepps:

(1) Point $\left(x_{C\_i}, y_{C\_i}\right)$ can be obtain by transforming the origin of the $i$th robot arm from frame $X_{Ri}O_{Ri}Y_{Ri}$ to the frame of the conveyor belt.

(2) Judge whether the following condition is satisfied the workspace of the robot arm is a sector whose radius is R, if Eq. (9) is satisfied, point $\left(x_{C\_i}, y_{C\_i}\right)$ is in the workspace.

$$\begin{cases} x_{C\_target} < x_{C\_i} + \sqrt{R^2 - (y_{C\_target} - y_{C\_i})^2} \\ x_{C\_target} > x_{C\_i} - \sqrt{R^2 - (y_{C\_target} - y_{C\_i})^2} \end{cases} \tag{9}$$

## 5 Experiments

The 1999 images are captured in laboratory simulation. The target object is bottle. The training set is 9 times than test set. The image is 600×1200. The parameters of the model are illustrated in Table 1.

Table 1: The parameters of the model

| | |
|---|---|
| Training set | resolution 600×1200 (1999) |
| Test set | resolution 600×1200 (199) |
| Anchor *n* | 901 |
| Anchor | [600,1000] |
| Image *n* | RPN: 1<br>VGG-16: 2 |
| RPN Mini-batch | 32 |
| RPN | 1:1 |
| RPN IOU-1 | 0.5 |
| RPN IOU-2 | [0,0.3] |
| Momentum | 0.9 |
| weight_decay | 0.0005 |
| Fine-tune | 40000 |
| NMS | 0.7 |
| Detection threshold | 0.99 |
| Weight function | P:1 A:5 |

The detection result is illustrated in Fig. 6. The bottle's position can be shown as the red Rectangular box, the line is the direction vector of the bottle.

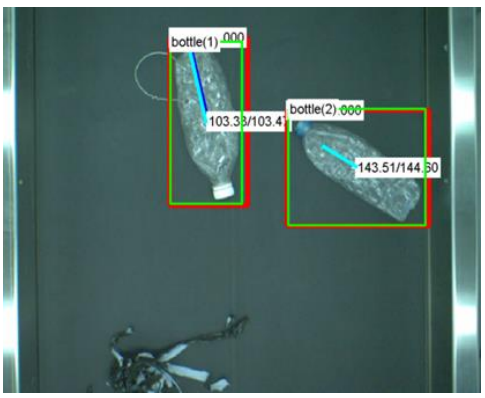


Fig. 6. The detection result

The test set is 199, rates of miss and false detection are defined as:

$$\eta_{missrate} = \frac{n_{missnum}}{n_{testset}} \times 100\% \quad (10)$$

$$\eta_{falserate} = \frac{n_{falsenum}}{n_{testset}} \times 100\% \quad (11)$$

Where $\eta_{missrate}$ is the target object is not be recognized, $\eta_{falserate}$ is not the target but be recognized. The $\eta_{missrate}$ is 3%, and the $\eta_{falserate}$ is 9%. The computing time is about 220ms.

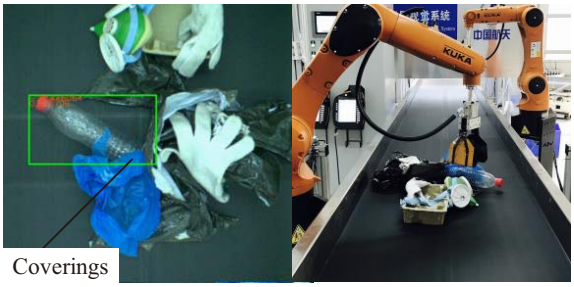


Fig. 7. The detection result with coverings

The detection result with covering and the robotic grasping are illustrated in Fig. 7. The Vision-based detection system is stable and efficient.

## 6 Conclusion

In this paper, we present a vision-based robotic grasping system that is capable of object detection, recognition and grasping with different poses by using a deep learning model. The Region Proposal Generation (RPN) and the VGG-16 model for object recognition and pose estimation are employed to identify the target object in complex background. Finally, the results of sorting experiment of the bottles in the garbage show that the vision algorithm and the manipulator control method of the proposed system can achieve the garbage sorting efficiently.

## References

[1] Tran S, Davis L S. 3D Surface Reconstruction Using Graph Cuts with Surface Constraints[C]. Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006: 219-231.

[2] A. Collet, M. Martinez, and S. S. Srinivasa, The MOPED framework: Object Recognition and Pose Estimation for Manipulation[J], International Journal of Robotics Research, 2011, 30(10): 1284-1306.

[3] E. Klingbeil, D. Drao, B. Carpenter, V. Ganapathi, O. Khatib, and A. Y. Ng, Grasping with application to an autonomous checkout robot[C]. Proc. IEEE Int. Conf. Robotics and Automation (ICRA), 2011, 19(6): 2837-2844.

[4] K. Hbner, D. Kragic, “Selection of robot pre-grasps using box-based shape approximation”,in IEEE Int. Conference on Intelligent Robots and Systems , 2008, pp. 1765-1770.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2015.

[6] R. Girshick. Fast R-CNN. arXiv: 1504.08083, 2015.