Garbage Sorting System Based on Composite Layer CNN and Multi-Robots

Liang Binyan1*, Wang Yanbo1, Wang Yao1, Chen Zhihong1, Lin Junqin1

China Academy of Launch Vehicle Technology (CALT), Beijing, China
liang_by@foxmail.com

Abstract. Object detection technology is the key problem in the garbage sorting system. There are many difficulties in identifying the target in the garbage, such as poor visibility and targets being coverage. The existing image processing methods are difficult to detect the target objects in the garbage sorting system. To solve this problem, this paper proposes a garbage sorting system based on convolution neural network and robots. The CNN network combines the features of three different network layers and predicts the location and angle information of the target objects at the same time, which we called composite layer. Coupled with the training solution of classification regression and angle prediction, the performance of object detection is further improved. The entire sorting system is composed of three subsystems: vision system, electrical system and mechanical system. The vision system is responsible for the image processing and send the position and attitude information to the electrical system. The electrical system controls the different robots to do sorting motion. We construct a Test Set which has 23983 objects to be detected. The experiment results show that our system’s Drop rate is 1.7% and False rate is 4.8%, which is good enough to meet the needs.

1. Introduction
Using machine vision to detect the position and posture of moving objects in complex environment is one of the most common methods. The purpose of developing machine vision is to make workers free from monotonous repetitive work and dangerous environment. With the maturity of industrial robot technology and the rapid development of machine vision, the research of garbage sorting system based on computational vision technology has been widely studied.

The welding seam tracking control system based on machine vision, which is studied by S.Murakami, uses visual sensor and image processing with neural network to obtain the shape data of weld seam, and can effectively track the weld seam. The Swiss SIG company has developed a combination of visual system, a parallel robot XR22 for taking parts on high speed conveyor[1]. Australia Telerobot, developed by Western University in Australia, is an industrial robot with six degrees of freedom with a camera. Telerobot process the image to generate a space position by continuously shooting and updating the image to get the space position. The ZenRobotics [2] company in Finland has designed a robot recognition and sorting system that using machine visual, 3D scanning, and spectral information to pick wood, stone, metal and other objects in the garbage. It can pick up 6000 recyclable garbage in an hour, and increase the efficiency by 429%. But the ZenRobotics system uses a variety of sensors, and the cost of equipment is very high.

Image processing is the key in the garbage sorting system. The traditional edge detection methods through the tuning of threshold for edge extraction will make more mistakes when the background is complex. Jilin Tu[3] propose an object tracking model based on Mean-Shift key points. But the
algorithm will be fail when the interference and objects’ colors are the same or close. There are many target detection techniques such as Viola and Jones's proposed face detection framework based on the Haar feature cascade classifier, a generic object detection framework based on the Deformable Partbased Model (DPM)[4]. However, those methods are designed using manual characteristics and are difficult to deal with the real complex environment.

The detection method like neural network[5-8] can process images fast and accurate. Ren Shaoqing[9-11] Proposed the RPN network to share the convolution layer with the classification network, and finally improved the detection accuracy. But it can’t get the position and posture information of the object and therefore it is difficult to be applied to the robot classification system.

In this paper we design a robot sorting system that solves the above problems in the complex industrial environment. Our contributions can be summarized as follows:

- We design the garbage sorting system and the vision system. The target image captured from the garbage disposal are analysed by a CNN based network. Then the robots do the continuous sorting job according to detection result.
- We design a composite layer CNN network by integrating the high level and low level feature maps. The accuracy of the object location can be improved, especially the small targets.
- We divide the system into three subsystems, each of which can be implemented separately and iteratively.

2. The CNN model design
In order to solve the problem of inaccurate detection of small targets detection network, we design composite layers CNN network by integrating the high level and low level feature maps. Because of the rich semantic information of high level feature maps, the precision of target detection based on neural network is greatly improved. However, the high level feature map has caused some problems, such as details lose because of multiple pooling layers. In order to solve this problem, we merge the different layers of feature maps. The composite layers CNN can solve the problem of the lack of details of small targets. Through composite layers CNN, the location of objects can be improved, especially the small targets.

2.1. CNN network
Compared with the one order information of the feature maps, the two order information contains more useful details. In this paper, we propose two order information CNN model, which starting from the idea of polynomial kernel function. By using the tensor learning method, the original external product is approximated by a few tensors. Then the formula is showed as below:

\[ f(x) = \langle w, x, b \rangle + \sum_{r=2}^{R} \sum_{d=1}^{D} \alpha_{r,d}^{s} \prod_{j=1}^{r} \langle c_{r,d}^{s}, x, b \rangle \]

\[ = \langle w, x, b \rangle + \sum_{r=2}^{R} \langle \alpha_{r}, \beta_{r} \rangle \]

(1)

Here, \( x \) is the input feature map. \( w \) is the weight variables. \( b \) is the bias. \( r \) is the \( r \)-th order. \( \alpha^{r} = [\alpha_{r,1}, ..., \alpha_{r,D}]^{T} \) is a set of weight vectors consisting of a tensor of rank 1. The formula can be decomposed into two groups of vector forms, namely weight vectors \( \{ w, \alpha^{2}, ..., \alpha^{R} \} \) and response vectors \( \{ c_{r,d}^{s} \}_{r=2, ..., R} \). The response vector can be decomposed into a group of responses \( x \) through a group of different size convolution. Through the above formula, we can insert the model into the whole convolution neural network. Then the backpropagation formula of the model is as follows:
The final CNN model is shown in figure 1

\[
\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y'} \sum_{d=1}^{D} \sum_{r=1}^{R} \left( \prod_{t=1}^{C_t} c_{t}^{r,d}, x \right) c_{t}^{r,d}
\]  

(2)

2.2. Position and posture prediction layer design

In our system, the images are taken as inputs to the VGG net. Then the data will be processed through convolution layers (including ROI pooling layers and ReLU layers[12]). The following full connect layer is to calculate the target class probability, the position and the posture. The full connect layer can be implemented by two forms: regression layers coupling and regression layers decoupling.

Regression layers coupling is implemented by adding fully connected layer Fc6 and Fc7 after the ROI pooling layer. Then Fc7 is divided into three outputs: classification, position and posture. Three outputs share a common Fc7 layer. Because Fc7 contains three kinds of information, they are softmax[13] probability, position regression information and posture regression information. Considering the less correlation between the network outputs, regression layers coupling may cause some problems e.g. the accuracy rate decrease, longer training time and hard to converge. We design the regression layers decoupling method.

The decoupling method adds two full connection layers after ROI pooling layer: Fc6 - Fc7 responsible for objects classification prediction and position regression prediction, and Fc6_1 and Fc7_1, responsible for the posture regression prediction. Regression layers decoupling separates the two different regression information for avoiding the intersection, improving the detection precision, shortening the training time and accelerating the model convergence. The decoupling method is shown as figure 2.
2.3. Training the model

Using the back propagation method to adjusting the parameters of the CNN. Here the loss function contains three parts[14], namely position loss, category loss, and angle loss:

$$L = \alpha \sum_i L_{\text{cls}}(p_i, p_i^*) + \beta \sum_i L_{\text{pos}}(t_i, t_i^*) + \lambda \sum_i L_{\text{ang}}(a_i, a_i^*)$$

Here, \(i\) is the index of anchor and \(p_i\) is the predicted probability of anchor \(i\) being an object. And \(p_i^*\) is the ground-truth label. \(t_i\) is the position vector of the \(i\)-th anchor, and \(t_i^*\) is the ground-truth position vector. \(a_i\) is the angle of the \(i\)-th anchor, and \(a_i^*\) is the ground-truth angle. \(\alpha, \beta, \lambda\) is weight of each item.

Then update the CNN model parameters according to the loss value:

$$h(j) = h(j - 1) - \delta \frac{L}{h^{(j)}(j)} , j = 1 \sim N$$

Here, \(N\) is the layer number. \(h(j)\) is \(j\)-th layer parameters. \(\delta\) is the learning rate.

3. Intelligent Sorting System

Considering the intelligent sorting system is used to sorting different kinds of objects, we divided the system into three subsystems, each of which can be implemented separately and iteratively. each subsystem has a certain degree of substitutability, which is not only convenient for system development, but also for changing using.

The intelligent sorting system consists of three parts, namely: vision system, electrical control system and mechanical system. The vision system can implement different software and interfaces. The electrical control system can use kinds of hardware, such as PLC, Embedded hardware, motion controller, etc. Figure 3 shows the structure of the intelligent sorting system.
The vision system is composed of industrial camera and vision processor. The vision processor is an ARM card embedded with GPU. The CNN network is deployed in the GPU card. The CNN network processes the image and sends the recognition result to the electrical system through the RS422 bus. The motion controller of the electrical system calculates the recognition result and sends the motion instructions to the mechanical system through the bus controller. The mechanical system consists of three parts: industrial robot, linear cylinder and mobile robot. The motion instructions is sent to the mechanical system through three different buses: WiFi, RS485 and ethernet. The Control flow of the entire system is shown in Figure 4.

The visual system is arranged at the front of head. It takes a long time for the target to move from the visual system to the capture range of the mechanical system. So we use the open-loop control. The motion controller calculates the sorting motion according to the following control algorithm. Here we use s7-400 as the motion controller. The algorithm implementation show as follows:

1) According to the recognition results of the target, convert the recognition result from the image coordinate to the world coordinate. Then save the recognition results to the result database.

\[
\begin{bmatrix}
  x_c \\
  y_c \\
  z_c \\
  1
\end{bmatrix}
= K \cdot R \cdot T \cdot 
\begin{bmatrix}
  u \\
  v \\
  1
\end{bmatrix}
\]

(5)
Here $u, v$ is object position in image coordinate, $x_c, y_c, z_c$ is the object position in world coordinate. $K$ is the internal parameter matrix of the camera. $R$ is the rotation matrix from the image coordinate to the world coordinate. $T$ is the transition matrix from the image coordinate to the world coordinate.

2) The motion controller calculates the difference between the taking time and the receiving time and converts it to the position result.

$$x_p = x_c + v \times (t_2 + \Delta T_i - t_1), i \in 1 \sim I$$
$$y_p = y_c$$

Here $v$ is conveyor running speed. $\Delta T_i$ is the $i$-th mechanical system running time, where $i$ means the different robots.

3) Convert the position from the world coordinate to the robot coordinate.

$$\begin{bmatrix}
x_r \\
y_r \\
a_r \\
1
\end{bmatrix} = M_{RC} \begin{bmatrix}
x_p \\
y_p \\
a_c \\
1
\end{bmatrix}$$

Here $x_r, y_r$ is the object position in robot coordinate. $a_r$ is the sorting angle in the robot coordinate. $a_c$ is the object angle in world coordinate. $M_{RC}$ is the transfer matrix between the two coordinates.

4) If the target position meets the following criteria, do the sorting operation.

$$\sqrt{x_r^2 + y_r^2} < L_i, i \in 1 \sim I$$
$$y_r < y_{max}$$

Here $L_i$ is the $i$-th robot motion range.

5) Check the objects result database and repeat the above loop.

4. Experiment

In this paper, we first train the network on the PASCAL VOC 2007 and 2012 DataSet. After obtaining the trained model, we use the garbage picture to fine tune the model. The following Table 1 shows the first trained model result:

| Method      | PASCAL VOC 2007 mAP(%) | PASCAL VOC 2012 mAP(%) |
|-------------|------------------------|------------------------|
| Faster R-CNN| 76.4                   | 73.8                   |
| SSD321      | 77.1                   | 75.4                   |
| DSSD321     | 78.6                   | 76.3                   |
| R-FCN       | 80.5                   | 77.2                   |
| Ours        | 80.6                   | 77.5                   |

Then we use our own Testset(Figure 5) to train the model. The TestSet has 23983 pictures, including 48032 annotated objects in total. According to the actual needs of industrial sorting, we define two sensitive indicators:

$$FalseRate = \frac{False Positives(FP)}{True Positives(TP) + False Positives(FP)} \times 100\%$$
$$DropRate = \frac{True Negatives(TN)}{True Positives(TP) + True Negatives(TN)} \times 100\%$$
The experimental results are shown in table 2.

Table 2. Trained result on the TestSet.

| Total objects | 23983 (TP+FN) | Dropout objects | 425 (FN) |
|---------------|---------------|-----------------|---------|
| Wrong detection | 1184 (FP) | Right detection | 23558 (TP) |
| Wrong position | 5 (ignore) | Wrong posture | 7 (ignore) |
| Drop rate | 1.7% | False rate | 4.8% |

As the table 2 shows, our Drop rate is 1.7% and False rate is 4.8%. Both the two sensitive indicators meet the actual industrial requirements. We show parts of the results in TestSet experiment as Figure 5.

5. Conclusion

In this paper, we propose a garbage sorting system based on composite layer CNN and multi-robots. We construct a TestSet which has 23983 objects to be detected. The experiment results show that our system’s Drop rate is 1.7% and False rate is 4.8%, which is good enough to meet the needs. This system achieves sorting automation in extremely complex environment, where the factory need some dedicated people to do the sorting jobs before.

The vision-based robot sorting system captures images from the complex industrial environment and calculate objects information by a CNN based network. Then the robots sort the objects according to detection result. We design composite layer CNN network by integrating the high level and low level feature maps. The system has a high precision output by adding the prediction layer of coupled regression.
We divide the system into three subsystems, each of which can be implemented separately and iteratively. Each subsystem has a certain degree of substitutability, which is not only convenient for system development, but also for replacement.

For future work, we hope to use more effective models to achieve higher precision. We also plan to improve our detection results by using more powerful CNN model [15-16] to detect more different kinds of objects. And we will move forward to search for methods to speed up image processing time.

Acknowledgments
Thanks for the support of the research work by China Academy of Launch Vehicle Technology (CALT). The authors would like to acknowledge the support of the team members for their outstanding contributions to the paper.

References
[1] Viola, P., and Jones, M. J. (2004). Robust Real-Time Face Detection. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (Vol.57, pp.747).
[2] Lukka, Tuomas J et al. “ZenRobotics Recycler--Robotic Sorting Using Machine Learning.” *Proceedings of the International Conference on Sensor-Based Sorting (SBS).* N.p., 2014. Print.
[3] Tu, J., Tao, H., and Huang, T. (2009). Online updating appearance generative mixture model for meanshift tracking. *Machine Vision and Applications, 20*(3), 163-173.
[4] Zhang, L., Yang, M., Jiang, Y., Regentova, E., and Lu, E. (2006). Generalized wavelength routed optical micronetwork in network-on-chip. *Proc.asted Int’l Conf.parallel and Distributed*, 698-703.
[5] Rawley Z. Heimer, Kristian Ove R. Myrseth, & Raphael Schoenle. (2019). Yolo: mortality beliefs and household finance puzzles. *The Journal of Finance, 74*(6).
[6] Huaizu Jiang, Ming-Ming Cheng, Shi-Jie Li, Ali Borji, & Jingdong Wang. (2019). Joint salient object detection and existence prediction. *Frontiers of Computer Science, 13*(4), 778-788.
[7] Wang, J., Chen, K., Yang, S., Loy, C. C., & Lin, D. (2019). Region proposal by guided anchoring.
[8] Moran Ju, Haibo Luo, Zhongbo Wang, Miao He. (2019). Improved yolo v3 algorithm and its application in small target detection. *Acta Optica Sinica, 39*(7), 0715004.
[9] Wang, X., Shrivastava, A., and Gupta, A. (2017). A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition* (pp.3039-3048).
[10] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *International Conference on Neural Information Processing Systems* (Vol.39, pp.91-99).
[11] Jiang, H., and Learnedmiller, E. (2016). Face detection with the faster r-cnn. 650-657.
[12] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science."
[13] Hinton, G. E., and Salakhutdinov, R. (2009). Replicated softmax: An undirected topic model. *International Conference on Neural Information Processing Systems* (pp.1607-1614). Curran Associates Inc.
[14] Aman, T. E., Ghosh, A., Sur, S. K., and Bhattacharya, D. K. (2010). Optimal control problem with comparatively large delay. 22(1), Pages. 23-30.
[15] Kong, T., Yao, A., Chen, Y., and Sun, F. (2016). Hypernet: towards accurate region proposal generation and joint object detection. 845-853.
[16] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science.*