Fast Recognition and Location Method of Parts for Assembly Robot Based on Deep Learning Network

Yongwei Yu, Jianheng Zhang, XinHan, Liuqing Du* and Zhenggen Li
College of Mechanical Engineering, Chongqing University of Technology, Chongqing, 400054, China
*Corresponding author’s e-mail: ftong@126.com

Abstract: Aiming at the problems of low components recognition rate and poor robustness under complex conditions such as adhesion, stacking and light source interference, a fast recognition and position method of parts based on deep convolution neural network was proposed. A multilayer convolution neural network for real-time detection of parts was constructed. Firstly, multi-scale extraction of candidate regions for target images was carried out. Then, the feature vectors of candidate candidate regions were automatically extracted by convolution neural network. The softmax classifier was used to recognize and locate. The experimental results show that the method realized the recognition and location for multiple parts in complex environment. The accuracy rate is over 95.2%. The detection took only 78ms. The accuracy rate and real-time performance of the method proved its feasibility.

1. Introduction
At present, combining robots with visual technology to achieve automatically parts recognition is the development and research trend in the field of intelligent assembly. Traditional image recognition methods mainly use artificial design features, such as SIFT features, LBP features, HOG features, etc. [1-6]. Due to the complicated conditions of the parts in the production line, such as messy placement, parts mixed, occlusion, overlap, etc., there are still many defects in the traditional method in the aspect of accuracy rate and real-time detection performance of the parts in the above production environment.

In recent years, deep learning methods have shown superior performance in terms of natural images, speech, and natural language processing [7]. Based on the detection framework of Faster R-CNN [8], this paper proposes a method for fast recognition of parts based on deep convolutional neural network. A multilayer convolutional neural network model was built. Firstly, the network does the multi-scale extracting of candidate regions for target images. Then feature vectors of candidate regions are extracted automatically based on depth learning method, and parts are identified and located automatically by combining with softmax classifier. Experimental results show that the method can realize accurate real-time recognition and positioning of robots on industrial production lines for complex parts in complex environment.

2. Part area extraction method based on deep convolutional neural network
In the traditional target detection task, the common methods for extracting the possible position of the target include the selective search algorithm and the Edge Boxes method. The above methods have defects in the field of extracting time and accuracy rate. However, although the target region generation method of the Faster R-CNN region selection network improves the speed, the image information is lost too much due to the multi-convolution calculation of the part image. Therefore, the
second and fifth layers of the feature extraction network are adopted. The convolution feature map is used to generate the candidate frame; and the area and the aspect ratio of the candidate frame are reset to achieve the purpose of improving the quality of the candidate region. The model realization flow of the multi-scale area extraction network proposed in this paper is shown in Figure 1. It is realized by steps such as image input, deep convolution feature extraction, part area position calculation and area classification discrimination.

![Image of multi-scale area extraction network flow chart]

**Figure 1. Multi-scale area extraction network flow chart**

The regional deep convolution feature extraction is implemented by a shared convolutional network, which is a five-layer deep convolutional neural network whose structural parameters as showing in Table 1.

| Number of layers | Types                  | Number of feature maps | Feature size | Convolution kernel |
|------------------|------------------------|------------------------|--------------|--------------------|
| 0                | Input Layer X0         | 3                      | 224×224      |                    |
|                  | The First Layer        |                        |              |                    |
| 1                | Convolution Layer C1   | 96                     | 110×110      | 7×7                |
|                  | The First layer        |                        |              |                    |
|                  | Pooling Layer S1       | 96                     | 56×56        | 3×3                |
|                  | The Second layer       |                        |              |                    |
| 2                | Convolution Layer C2   | 256                    | 28×28        | 5×5                |
|                  | The Second Layer       |                        |              |                    |
|                  | Pooling Layer S2       | 256                    | 14×14        | 3×3                |
| 3                | The Third Layer        |                        |              |                    |
| 4                | Convolution Layer C3   | 384                    | 14×14        | 3×3                |
|                  | The Fourth Layer       |                        |              |                    |
| 5                | Convolution Layer C4   | 384                    | 14×14        | 3×3                |
|                  | The Fifth Layer        |                        |              |                    |
|                  | Convolution Layer C5   | 256                    | 14×14        | 3×3                |

The coordinate calculation and area discrimination of the part area are realized by the frame correction network and the softmax discriminant network respectively. The softmax network calculates whether there is a target in the area, so that the detection area is divided into the foreground and the background, the foreground is used as the candidate frame to recognize the part type by the subsequent part classification network.

For the 6 candidate regions generated by each point, 6 kinds of parameters are output in the end. First of all, the target box score whether included in the candidate box, which is the scores of the foreground and the background; then, calculate the 4 parameters that need to be translated and scaled by the coordinate values of the generated candidate frames. The scores belonging to the foreground and background are calculated by the softmax function. The expression for the softmax classifier is as follows:
Where $x_i$ represents the feature vector of the feature map, foreground and background categories are denoted by $y_i$, $w$ and $b$ represent the weight and offset parameters of the model, respectively.

By extracting the candidate frames of the feature maps of different layers, the information of different feature maps can be utilized effectively, and the loss of image information after convolution calculating can be avoided; using a convolution operation with stride of 2 for the second-level convolutional feature map can reduce the coincidence with the centre point of the candidate frame of the fifth-level feature map, and reduce the redundant candidate area at the same time. The above working part has finished the area network extracting, it would save the relevant information of the extracted candidate frame for further using in the next detection network.

3. The recognition of part based on deep convolution feature

After obtaining the region candidate frame, the type of parts is recognized by the feature map which based on the deep convolution network. The recognition principle of part is shown in Figure 2. Firstly, the recognition network maps the candidate regions in the original part image back to the convolution feature map, and then adopts. According to the feature vector obtained by the mapping realize the classification of convolution feature map with the softmax layer.

![Figure 2. The recognition network of part](image)

The candidate area is the maybe-area of the part in the original picture. In order to realize the convolution feature sharing of the regional extraction network and the recognition convolution network. And remap the candidate regions in the original image back to the convolution feature map.

After extracting the feature vector of candidate region, input it into the softmax layer for classification. The category classification is not only the binary classification for predicting the existence of target when the candidate frame is extracted, but the multi-classification task classifies different targets accurately.

Set loss function of the classification network as follows:

$$L = -\sum_{i=1}^{N_{reg}} \log \left( \frac{e^{x_i^T y_i + b_j}}{\sum_{j=1}^{N_{cls}} e^{x_i^T y_j + b_j}} \right)$$

(1)

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

(2)

Where $p_i$ represents the confidence of the classification network for the category to which the target belongs, $p_i^*$ represents the true value label of the target, if the sample is positive, $p_i^*$ is 1, or 0 on the contrary. $t_i$ represents the four coordinate of parameter values of the image region position, $t_i^*$ represents the coordinate value of the real frame, $N_{cls}$ indicates the size of each small batch training, and $N_{reg}$ represents the number of candidate frames. $L_{cls}(p_i, p_i^*)$ represents the error function of target classification:
\[ L_{ch}(p_i, p_i^*) = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \]  

\[ L_{reg} \] represents the position error of coordinate, which indicates the translation scaling error between the real candidate region and the prediction candidate region.

By performing candidate region extraction and recognition on the image target, comparing with the existing method, the time-spending of the entire detection process is greatly shortened.

4. Experiment and Analysis

We chose 4 types of part as the object in this experiment, namely nut, gear1, gear2, and injector. This experiment simulates any production environment of different types of parts in the industry, such as arranging, occlusion, illumination changing, etc. A total of 6,000 images of 4 types of parts are collected, of which 5000 are used as training samples and the rest are used as test samples. After finishing the training process, the misclassification rate of the training sample parts is 4.9%, and the average misclassification rate of the test sample parts is 6.7%, and the total accuracy rate perform good.

In this paper, the category and position of the parts in the image are recognized at the same time. The categories are output in the form of probabilities in the figure, and the position of the parts is represented by a rectangular frame in the figure. Figure 3 shows a plurality of parts in a non-uniform illumination, overlapping occlusion effect identified under complicated conditions.

![Figure 3. Recognition effects in other complex environment](image)

Normally, the recognition rate of each part with this algorithm can reach above 91%, and it is also robust to illumination, multi-angle and other circumstances changing. It is also performing better in complex environments such as overlap, occlusion, and hybrid. The recognition time-spending is less than 78ms, which satisfies the real-time detection requirements.

5. Conclusion

this paper builds a deep learning recognition and positioning network based on the framework of Faster R-CNN, and realizes the recognition of multi-class parts in complex environment. The recognition accuracy can meet the needs of assembly robots.

Acknowledgments

Thanks to the funding of Chongqing Basic and Frontier Research Project (CSTC2017JCYJAX0344).

References

[1] Yuan J S, Cui K B, Li B S. (2015) Identification and location of insulator video images based on ASIFT algorithm [J]. Electrical Measure and Instrumentation, 7(7):106-112.

[2] Wan Y, Li H H, Wu K F, Tong H Q, (2015) Fusion with Layered Features of LBP and HOG for Face Recognition [J]. Journal of Computer-Aided Design & Computer Graphics,04:640-650

[3] Nie H T, Long K H, Ma J, et al. (2016) Face recognition based on fast scale invariant feature transform algorithm and fuzzy control [J]. Journal of Jilin University: Engineering and Technology Edition, 46(2): 549-555.
[4] Wang D, Zhang X H. (2013) Biomimetic recognition method of human behaviour based on HOG and SVM [J]. Journal of Jilin University: Engineering and Technology Edition, 43(S1): 489-492.

[5] Wang H T, Fu W P, Kang Y N. (2008) Study of edge matching approach to workpiece image recognition [J]. Chinese Journal of Scientific Instrument, 29(5): 986-991.

[6] Wang Y, Fu W P, Zhu H, Wang X H. (2009) Image Match for Workpiece Based on Edge Characteristics. 2009.10:1297-1302.

[7] Hinton G E, Osindero S, Teh Y W. (2006) A fast learning algorithm for deep belief nets [J]. Neural Computation, 18(7): 1527-1554.

[8] REN S Q, HE K M, GIRSHICK R, et al. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39 (6), 1137-1149.