An improved gesture recognition algorithm for human-robot collaboration

Yiqun Liu¹, Xiaogang Wang¹*, Yang He², Hua Mu³ and Yuewei Bai¹

¹School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai, 201029, P. R. of China
²China Tobacco Guizhou Industrial Co., Ltd, Guizhou, 550001, P. R. of China
³Wuhan KM Information Technology Co., Ltd., Hubei, 430076, P. R. of China
*Corresponding author’s e-mail: xgwang@sspu.edu.cn

Abstract. In this paper, an improved gesture recognition algorithm framework based on the improved SSD algorithm and OpenPose model for solving the robot’s mobile grasping problem in human-robot environment is proposed. Improved SSD algorithm by optimizing the front network of SSD algorithm and adjusting the prediction box is designed to identify the position of the hand rapidly, and to reduce the time for the training model. Testing experiments are performed for recognizing the hand’s key points information and eight designed gestures are compared with the traditional algorithm. The results show that the method can greatly improve the recognition distance of the original model under the dense environment, and also enhance the robustness and accuracy of the model, and have improved the ability of long-distance human-robot collaboration to a certain extent.

1. Introduction

With the rapid development of computer technology, especially virtual reality technology [4], the research on gesture recognition has gradually become a focus of human-robot collaboration. Although the communication channels between humans and robots are still limited, using gestures as the interface between humans and machines have been extensively applied for a long time [1]. Human beings have unique problem-solving abilities and flexible movement capabilities, but they have defects in strength and accuracy [2-3]. The robot can bear higher work intensity and has faster speed and higher efficiency, but the flexibility is quite insufficient. Human-robot collaboration allows people and robots to collaborate on a shared manufacturing environment. Gestures are used as human-robot interfaces, for example, using gesture to control the robot for mobile grabbing, allowing people to establish communication channels with robots, which can liberate people from tedious repetitive work, so as to get better benefits. At the same time, the recognition distance and accuracy of gestures in the intelligent manufacturing environment have higher requirements. In complex background environments, it is always a challenge to realize gesture recognition without contact.

Before the emergence of deep neural networks, traditional gesture recognition methods mainly include methods based on template matching [5], methods based on geometric features [6-7], and methods based on hid Markov models [8], etc. However, in practical applications, gestures will have the characteristics of diversity and polysemy because of different environmental backgrounds and different people, which will increase the difficulty of image processing, image segmentation, and skin color detection, and the complexity and variety of hand shape, which also increases the difficulty of gesture
recognition. There are also applications that rely on the depth information provided by Microsoft's Kinect body sensor for gesture recognition and control [9], but the cost is brought up and the use is not flexible enough.

Since the emergence of deep neural networks, more and more researches have been conducted on gesture recognition using convolutional neural networks [10-11] and recurrent neural networks. Aiming at the requirement of human-robot cooperation in intelligent manufacturing environment and the problem of robotic mobile grasping, a gesture recognition algorithm and a control robot mobile grasping system are designed, the gesture recognition is performed through the OpenPose model algorithm, combined with SSD algorithm. The improved SSD algorithm is used to preprocess the image and extract the image which contains only the hand, and then call the OpenPose model for gesture recognition.

2. Framework of Gesture recognition algorithm

2.1. Procedure of gesture recognition

In the intelligent manufacturing environment, there are some problems in the robot's mobile grasping, such as the complex background and need to keep a certain safe distance between human and robot, which requires the gesture recognition algorithm to be robust to the complex background and to have a relatively long detection distance. When the distance increases, the RGB image of gesture will contain a lot of background information that influences the recognition accuracy. This paper aims to adjust the network structure and prediction box. The improved SSD algorithm is used to extract hand parts to reduce the impact of background information. Report to direct gesture recognition, only hand recognition will greatly reduce the difficulty of recognition. The hand image was input into the OpenPose model, and the key points of the hand were direct output. The key points were classified for gesture recognition. The OpenPose model opens up source library developed by Carnegie Mellon university (CMU) based on convolutional neural network and supervised learning and based on Caffe. It can estimate the posture of human body movement, facial expression, finger movement and so on. The hand key points detection model [20] outputs a total of 22 key points, including 21 points of the hand. The 22nd point represents the background, as showing in figure 1. The structure of gesture recognition algorithm is given in figure 2.

![Fig.1 Detection model of hand critical points](image1.jpg)

![Fig.2 Structure of gesture recognition algorithm](image2.jpg)
2.2. An improved SSD algorithm

2.2.1 Improvement of front network

SSD (Single Shot MultiBox Detector) algorithm is a multi-box prediction, end-to-end target detection method. Although SSD algorithm has achieved a relatively speedy detection speed, there is still room for optimization and improvement of the vgg-16 front network it uses. Depth of Separable Convolution (Depthwise Separable Convolution) to standard decomposition of Convolution kernels, reduce the amount of calculation and improve the running speed. DSC includes two parts: depthwise convolution (DWC) and pointwise convolution (PWC). DWC filters the input channels without increasing in the number of channels. PWC is used to connect different channels of PWC and can increase in the number of channels. By means of this decomposition, the amount of computation can be considerably reduced. DSC divides a standard convolution kernel into a deep convolution kernel and a 1×1 point convolution kernel. Assuming that the input is the feature graph of F channels, the size of the convolution kernel is H×H, and the output channel is E, the standard convolution kernel is E×H×H×F.

![Fig.3 Convolution structures](Traditional- Depthwise- Convolution)

The calculation amount of traditional convolution is:
\[ E \times F \times H \times H \times G \times G \]  
(1)

The total calculation amount of DSC is:
\[ F \times H \times H \times G \times G + E \times F \times G \times G \]  
(2)

For example, if the input characteristic graph is m×n×16, and 32 channels are output, the convolution kernel should be 16×3×3×32, then it can be decomposed into deep convolution: 16×3×3 to obtain the characteristic graph of 16 channels. The point convolution is 16×1×1×32. If the standard convolution is used, the calculation amount is: 
\[ m \times n \times 16 \times 3 \times 3 \times 32 = m \times n \times 4608. \]
The computation of the deep decomposable convolution is 
\[ m \times n \times 16 \times 3 \times 3 + m \times n \times 16 \times 1 \times 1 \times 32 = m \times n \times 656. \]
Therefore, compared with the standard convolution kernel, the calculation ratio is:
\[ \frac{m \times n \times 4608}{m \times n \times 656} = \frac{1}{E} + \frac{1}{H^2} \]  
(3)

2.2.2 Improvement of default box

The default box in SSD algorithm is determined according to the size of the feature diagram of each layer, and the initial size and length-width ratio of the default boxes have to be set manually. The basic size and shape of default boxes in a network can't be learned directly. They need to be set manually. However, the default box size and shape used by each layer feature in the network are exactly different, which makes the debugging process highly dependent on experience. In this paper, only the hand part has to be tested, and the length to width ratio of the hand default boxes is relatively fixed. Setting the default box parameters can improve the accuracy and speed of hand testing. Different from the manual setting method, this paper adopts k-means clustering analysis algorithm to carry out clustering analysis on the target box, get the optimal K value for the hand detection problem through iterative analysis, and set the number of default boxes to K. Traditional K-means algorithm using Euclidean distance as the judgment standard, this will lead to big goals box produce the bigger error values, inconsistent with our aim, so this article use the IOU (overlapping), namely two box intersection and the ratio of two box and set as the judging standard, the K-means cluster analysis, and make a balance between accuracy and algorithm complexity, finally selected K value for 4 candidate box for the default box of initial setting.

In addition, the default boxes are generally designed to surround the hand or cut off part of the finger, as shown in fig.4(a), the size of the default boxes is increased to make the OpenPose recognize more hands and increase the robustness of recognition, as shown in fig.4(b).
3. Comparison and analysis of experimental results

3.1 Experimental environment and data set

The experimental environment and configuration of this paper are shown in Table 1.

| Experimental environment | Windows10pro, TensorFlow1.13 |
|--------------------------|------------------------------|
| CPU                      | INTEL® CORE™ i7-8750H        |
| GPU                      | GeForce GTX 1050 Ti         |
| Video Memory             | 4G                           |
| Memory                   | 16G                          |
| The data set             | Egohands hand data sets and self-made data sets |

In this paper, the Egohands hand data set from Indiana University was used to train the improved SSD model. The Egohands dataset consists of 48 Google Glass videos showing interactions between two people in a complex setting, with 15,053 accurately tagged hand infographics. At the same time, a test set containing eight gestures was made, and 100 pieces of each gesture were collected. A total of two people participated in the gesture collection. Gestures were collected by the 720p camera at 1m, and a total of 1600 images of 640×480 pixels were obtained, and 1600 images of 4032×1960 pixels were collected at 3m, which were used as the test data set of this experiment. Part of the self-made gesture data set is shown in Fig 5.

4.2 Experimental methods

After testing, the limit recognition distance of the OpenPose hand key point model of the 720p resolution is about 0.8-1m, beyond which the key point cannot be recognized or only a few key points can be recognized, as shown in Fig 6. The increase of distance makes the field of vision larger and generates a lot of interference information at the same time. Moreover, the pixel number at the 720p resolution is just 640×480, so the further increase of distance will also cause the decrease of recognition accuracy.
Since OpenPose model is done through high resolution data set training, thus improving resolution to improve the model identification of distance, but the increase of the distance and affect the identification accuracy, thus improved the SSD model were used to detect shots, then the hand image input OpenPose model, not only can improve the accuracy of recognition, also can increase the recognition distance. OpenPose model, SSD algorithm and OpenPose model, and improved SSD algorithm and OpenPose model were utilized to verify the test set. In this experiment, two kinds of distance test data sets were made, among which the 1m distance test data set was used to test whether adding the improved SSD algorithm would affect the original recognition accuracy, and the 3m distance test data set was used to explore whether the OpenPose model could increase the recognition distance and maintain the original accuracy after adding the improved SSD algorithm.

![Fig.6 Detection failure cases](image)

4.3 Comparison and analysis of experimental results
Firstly, the data set at 1m distance was used to test the three models. The missing test was that the model failed to detect the key points or only detected a few key points and could not carry out gesture recognition. The results are shown in table.2.

| Model                  | Total | Null | FP  | TP   | Accuracy |
|------------------------|-------|------|-----|------|----------|
| OpenPose               | 1600  | 329  | 85  | 1186 | 74.125%  |
| SSD algorithm and OpenPose | 1600  | 589  | 81  | 930  | 58.125%  |
| Improved SSD algorithm and OpenPose | 1600  | 350  | 86  | 1164 | 72.75%   |

According to the above experimental results, within the recognition distance of the key point model of OpenPose hand, the addition of the improved SSD algorithm did not affect the original accuracy, while the direct addition of SSD algorithm would significantly reduce the recognition accuracy.

The three models were then tested using the 3m distance data set, among which the OpenPose model was completely unable to detect any key points, and the results of the other two models are shown in table.3.

| Model                  | Total | Null | FP  | TP   | Accuracy |
|------------------------|-------|------|-----|------|----------|
| SSD algorithm and OpenPose | 1600  | 515  | 63  | 1022 | 63.875%  |
| Improved SSD algorithm and OpenPose | 1600  | 363  | 73  | 1164 | 72.75%   |

From the above data, we can be added to improve SSD OpenPose hand point model of algorithm greatly improve the detection distance, raised by 1 m to 3 m, at the same time improve the SSD algorithm comparing the original SSD for OpenPose model has the obvious improvement of gesture recognition accuracy, visible SSD algorithm for gesture recognition OpenPose model of ascension is bigger.

4. Conclusion
In this paper, an improved gesture recognition algorithm based on OpenPose model and an improved SSD algorithm were proposed. The experimental results were showed that this method can significantly improve the recognition distance and accuracy, increase the recognition distance of the OpenPose model, improve the accuracy of SSD algorithm alone, reduce the training amount of the model, and have good
accuracy and robustness in the complex background. It was also practical to control the robot mobile grasping system. The next work direction is to further optimize and improve the OpenPose hand critical point model, and to use other more advanced target detection methods to replace the improved SSD algorithm, and to test some more accurate gesture classification methods, so as to further improve the speed and recognition accuracy of the model.

Acknowledgments
The research is supported by the National Key R&D Program of China under Grant No. 2017YFE0118700, the National Natural Science Foundation of China under Grant No. 51605273, the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734599.

References
[1] Liu H, Wang L. Gesture recognition for human-robot collaboration: A review[J]. International Journal of Industrial Ergonomics, 2018, 68: 355-367.
[2] Krüger J, Lien T K, Verl A. Cooperation of human and machines in assembly lines[J]. CIRP annals, 2009, 58(2): 628-646.
[3] Green S A, Billinghurst M, Chen X Q, et al. Human-robot collaboration: A literature review and augmented reality approach in design[J]. International journal of advanced robotic systems, 2008, 5(1): 1.
[4] Rückert P, Wohlfromm L, Tracht K. Implementation of virtual reality systems for simulation of human-robot collaboration[J]. Procedia Manufacturing, 2018, 19: 164-170.
[5] Ciliberto M, Cuspinera L P, Roggen D. WLCSSLearn: learning algorithm for template matching-based gesture recognition systems[C]. 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR). IEEE, 2019: 91-96.
[6] Modanwal G, Sarawadekar K. A robust wrist point detection algorithm using geometric features[J]. Pattern Recognition Letters, 2018, 110: 72-78.
[7] Kumar S, Kaurav A. Hand gesture through geometric moments (HCI based)[C]/2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018: 561-565.
[8] Babu D, Konyo M, Nagano H, et al. Stable haptic feedback generation for mid-air gesture interactions: a hidden Markov model-based motion synthesis approach[J]. ROBOMECH Journal, 2019, 6(1): 2.
[9] Mueller F, Bernard F, Sotnychenko O, et al. Generated hands for real-time 3d hand tracking from monocular rgb[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 49-59.
[10] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[11] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.