International Journal of Computational Intelligence Systems and Applications. Volume 10, Issue 1, (2023) 29-37
http://www.cayley-nielson.com/ijcisa/

Design of Gesture Recognition System based on YOLOV5

Dongyan Cui[1,2]

[1]School of Artificial Intelligence, North China University of Science and Technology ,Tangshan, Hebei, China

[2]Xingtian (Suzhou) Intelligent Control Technology Co., Ltd.Suzhou, Jiangsu, China

E-mail: *cdy_xxz@163.com*

Abstract

Gesture recognition, as a representative way of human-computer interaction, has a large demand in fields such as VR, sign language recognition, and safe driving. This article conducts gesture recognition experiments based on two deep learning network models of YOLOv5, and uses Faster R-CNN as a comparative algorithm to recognize gesture images in simple and complex environments. The average recognition accuracy, recall rate, and detection time are used as algorithm efficiency evaluation indicators to evaluate the accuracy of gesture recognition, and the recognition results of the two algorithms are compared and evaluated. The experimental results show that in a simple background, the detection accuracy of Faster R-CNN can reach 86.46%, but there are errors and omissions. The YOLOv5 algorithm has a recognition accuracy of 93.89% and can accurately recognize gestures, and the model size is much smaller than that of Faster R-CNN. In complex background environments, the recognition accuracy of both algorithms has decreased to varying degrees, but the model recognition performance based on YOLOv5 network is still better than Faster R-CNN. Therefore, gesture recognition based on YOLOv5 network has better recognition performance, and can achieve accurate recognition of gestures in complex backgrounds, with good adaptability and stability.

Keywords: YOLOV5 algorithm; gesture recognition; object detection; deep learning

1. Introduction

Gesture recognition has a very wide range of applications in production and daily life, with huge market demand and development prospects, and has become one of the hot research fields [1-2]. In 1997, Assam et al used an implicit Markov model for feature extraction through colored gloves, which achieved a gesture semantic recognition accuracy of 91.3%[3]. In 2017, Fan Wenbing proposed a pose recognition method based on skin color feature extraction, which used Haar features to remove facial regions and obtain gesture segmentation images, achieving a recognition accuracy of 95.8% [4]. Deep learning has greatly improved the recognition effect of gesture recognition and is the future development direction of gesture recognition technology. Oyebade et al successfully identified 24 American sign languages using CNN [5]. Wu Xiaofeng's team used Faster R-CNN for gesture recognition, using ZF and VGG networks to extract gesture features, and then debugged the network input parameters, target framework, and training hyperparameters to improve the efficiency of gesture detection and recognition [6]. In 2020, Bochkovskiy A et al proposed the YOLOv4 network, which integrates multiple techniques that can improve accuracy [7]. In June 2020, the YOLOv5 network was proposed, which has the advantages of small model size, low deployment cost, and high flexibility [8-10].

How to make computers more accurate and efficient in recognizing the meanings represented by

different gestures has become an urgent issue to be solved. Therefore, research on gesture recognition is of great significance for improving recognition efficiency, enhancing human-computer interaction experience, and promoting the development of artificial intelligence.

2. Gesture Recognition Algorithm based on YOLOv5 Network Model

2.1. YOLOv5 Network Model Architecture

The YOLOv5 network model architecture is shown in Figure 1.

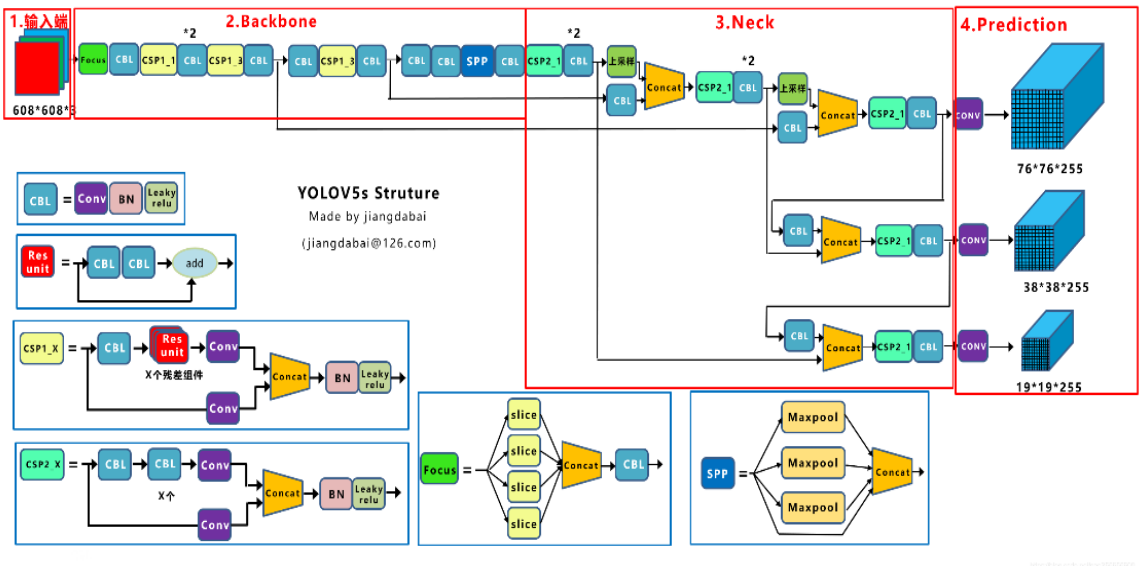


Fig.1 YOLOv5 Network Model Architecture

2.2. Steps for Gesture Recognition based on YOLOv5

The entire YOLOv5 is divided into three parts: the backbone extraction network Backbone (CSPToken), the enhanced feature extraction network FPN, and the classifier and regressor YOLO Head. The input image first enters the CPSDarknet for feature extraction to obtain a feature layer, which is the feature set of the input image. Obtain three feature layers in the backbone, namely the effective feature layer. Then input three effective feature layers into FPN for feature fusion, including upsampling the features for feature fusion and downsampling the features again for feature fusion, to obtain three strengthened effective feature layers, each with width, height, and channel number. The feature map can be viewed as a collection of feature points one after another, with each feature point having several channels and features. Finally, YOLO Head judges the feature points to determine whether there are objects corresponding to them, achieving object detection. The entire YOLOv5 network detection process includes three steps: feature extraction, feature enhancement, and prediction of the object situation corresponding to feature points.

3. Experimental Environment and Data Preparation

3.1. Experimental Environment and Development Platform

The gesture recognition system designed for this experiment was built on Pycharm 2021, using Python 3.8 as the development language and PyTorch 1.8.0 as the deep learning framework. The system was implemented based on the Opencv library and pyqt. The experimental environment is shown in Table 1.

Tab.1 Experimental Environment

Experimental Environment	
CPU	Intel i5-7200U
Memory	16GB
GPU	NVIDIA GeForce MX130
Video storage	8GB
Camera	Resolution 640x480, 30 frames/second
development language	Python

3.2. Gesture Dataset

The gesture dataset used in this experiment is the sign language number dataset and the gesture letter dataset.

The sign language digital dataset consists of 2180 gesture images captured by different experimenters under simple backgrounds, including ten international sign language digital gestures, with an image size of 100x100 pixels. In the later stage of this experiment, the original dataset was expanded through image flipping, adding noise, and color transformation. The definition of gestures in the dataset is shown in Figure 2.

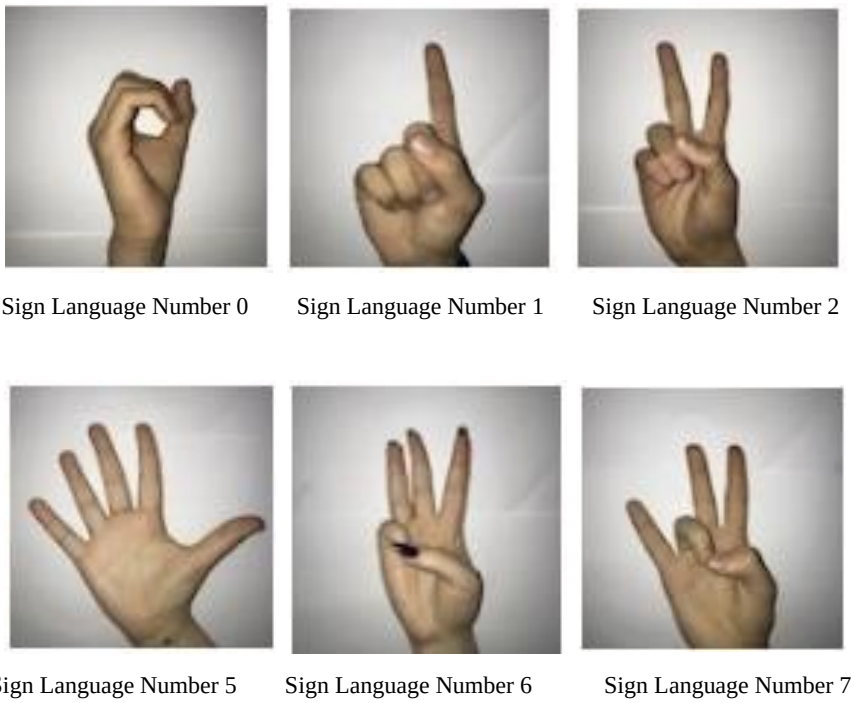


Fig. 2 Definition of some gesture images in the dataset

The gesture letter dataset includes 10 gestures, including 'A', 'number7', 'D', 'I', 'L', 'V', 'W', 'Y', 'I LOVE YOU', and 'number5'. Twelve individuals captured the above 10 gestures. The environment and background of the photographer in this dataset are relatively complex, with an image size of 1280x960 and a total of 1200 images. The gesture images and their meanings in the dataset are shown in Figure 3.

Letter 'A' Letter 'D' Letter 'V'

Letter 'W' Letter 'Y' 'i love you'

Fig.3 Definition of Hand Gesture Letters

3.3. Label of Datasets

The gesture dataset is produced using the VOC2007 dataset format. Firstly, use Labelimg software to frame the gestures in the dataset images, input categories, and save them. Generate an XML format label file and randomly divide the training and testing sets in a 9:1 ratio. The dataset annotation process is shown in Figure 4.

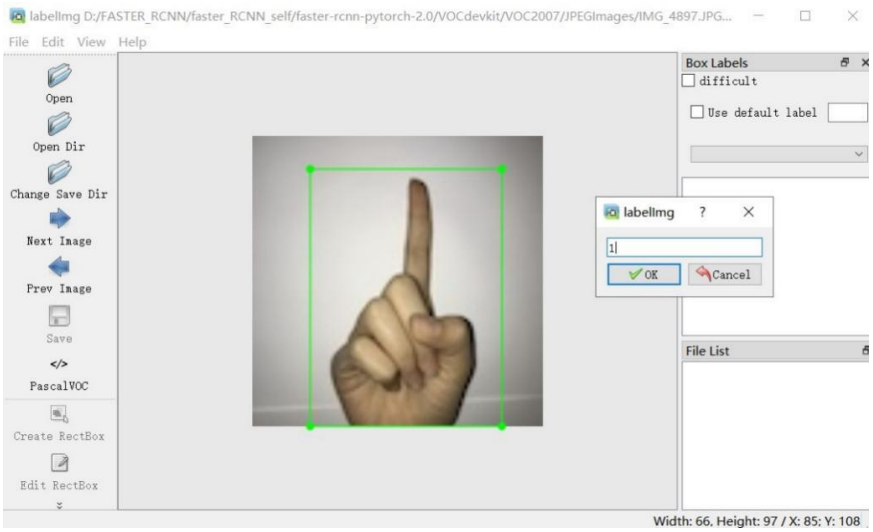


Fig. 4 LabelImg annotation interface

3.4. Data Enhancement

Due to the fact that there are only 2062 photos in the sign language digital dataset, and the experimental dataset images are all taken in the same single background environment, which is less affected by other environmental conditions. Therefore, in order to enrich the dataset and achieve better training results, this experiment uses data augmentation methods to scale, add noise, adjust brightness, cut, and change HSV on

the photos, expanding the dataset to 10310 photos. The partial sample images after data augmentation and expansion are shown in Figure 5.

Fig.5 Data samples after data augmentation

Similarly, the gesture letter dataset was expanded using the above method, from the original 1200 sheets to 5000 sheets. The partially enhanced images are shown in Figure 6.

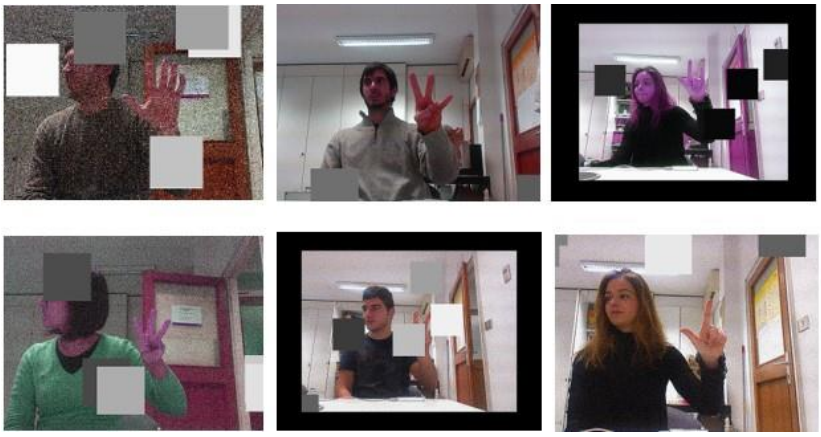


Fig. 6 Image of Enhanced Part of Gesture Letter Dataset Data

4. Experimental Results and Analysis

4.1. Model Training

The entire training process is divided into 100 epochs for training, which is divided into two stages: the freezing stage and the thawing stage. During the freezing training phase, the model backbone is frozen, and the feature extraction network remains unchanged. Freeze_ Batch_ Set size to 8 and learning rate to 1e-3. During the thawing phase, the feature extraction network changes and unfreezes_ Batch_ Set the size to 4 and the learning rate to 1e-4. Loss function curve are shown in Figure 7.

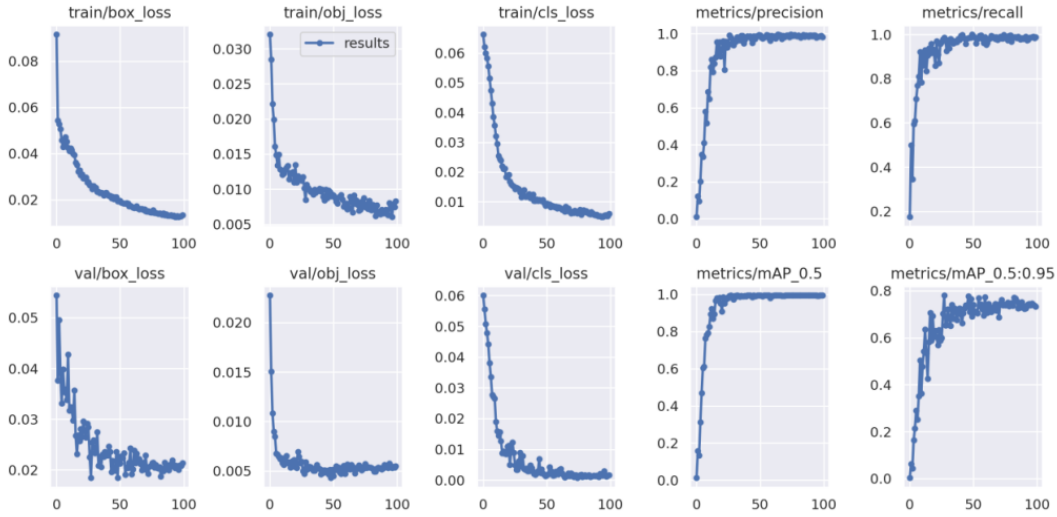


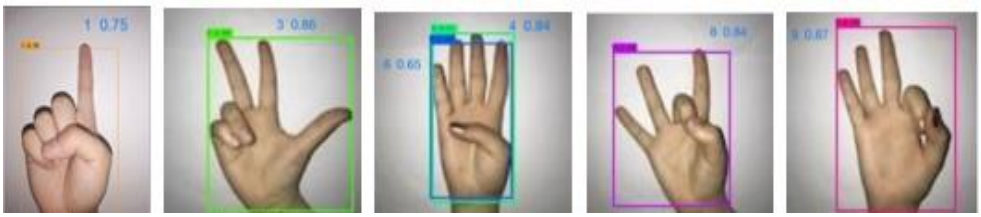
Fig. 7 Loss function curve

4.2. Evaluation Indicators

This experiment uses average accuracy (AP), recall (Recall), model size, and detection time as evaluation indicators for gesture recognition results. The calculation formula is shown in Equation 1.

$$\begin{cases} AP=\dfrac{TP}{TP+FP}\times 100\% \\ R=\dfrac{TP}{TP+FN}\times 100\% \end{cases} \tag{1}$$

Where, TP represents the number of correctly recognized gesture images, and FP represents the number of gesture images with incorrect recognition; FN represents the number of unrecognized gesture images.



Number 1 Number 3 Number 4 Number 8 Number 9

Fig. 8 Recognition Results of Faster R-CNN Network on Sign Language Digital Datasets

Number 1 Number 3 Number 4 Number 8 Number 9

Fig. 9 Recognition Results of YOLOv5 Network on Sign Language Digital Dataset

4.3. Analysis of Comparative Experimental Results

(1) Analysis of Recognition Results for Sign Language Digit Datasets

The comparison algorithm adopts the Faster R-CNN algorithm and uses two types of networks to detect sign language digital datasets. The partial detection results of the Faster R-CNN network in the sign language digital dataset are shown in Figure 8. The recognition results of YOLOv5 network on this dataset are shown in Figure 9.

The comparison of recognition results between the two algorithms on the sign language digital dataset is shown in Table 2.

Tab.2 Recognition Results of Two Algorithms on Sign Language Digit Datasets

Gesture	Recognition effect	Faster R-CNN	YOLOv5
Number 1	categories	1	1
	accuracy	0.75	0.86
	Detection box positioning	accurate	accurate
Number 3	categories	3	3
	accuracy	0.86	0.95
	Detection box positioning	accurate	accurate
Number 4	categories	Recognition error	4
	accuracy	×	0.97
	Detection box positioning	inaccurate	accurate
Number 8	categories	8	8
	accuracy	0.84	0.89
	Detection box positioning	accurate	accurate
Number 9	categories	9	9
	accuracy	0.67	0.96
	Detection box positioning	accurate	accurate

Tab. 3 Analysis of evaluation indicators for recognition rate of two algorithms

Recognition algorithm	AP	Recall	Model size	Detection time
Faster R-CNN	86.46%	89.52%	108MB	0.186s
YOLOv5	93.89%	96.64%	54.3MB	0.103s

According to Table 2, the detection performance of the Faster R-CNN algorithm is not ideal in the same gesture image, resulting in incorrect recognition, recognition of multiple gestures, and inaccurate positioning of the detection frame, with low recognition accuracy. The YOLOv5 algorithm has better recognition performance and can accurately recognize gesture categories with high accuracy. The detection box positioning is accurate, and there are no recognition errors. Multiple gestures are recognized. The analysis of the recognition accuracy of the two algorithms in the sign language digital dataset is shown in Table 3.

From Table 3, it can be seen that the YOLOv5 algorithm has an average accuracy of 94.89% and a recall rate of 96.64% on this dataset, which is 8.43% higher than Faster R-CNN and 7.12% higher than

Faster R-CNN. The model size is only 54.3MB, which is 53.7MB smaller than Faster R-CNN and has a shorter detection time, better meeting the requirements of real-time detection.

(2) Analysis of Recognition Results for Gesture Letter Datasets

Two algorithms are used for recognition on the gesture letter dataset, with the detection box selecting the gesture position and the upper left corner indicating the recognition category and confidence level. The recognition results of the Faster R-CNN algorithm are shown in Figure 10, and the YOLOv5 algorithm is shown in Figure 11.

Letter'A' Number 7 Letter'V' Letter'L'

Fig. 10 Recognition results of Faster R-CNN algorithm on gesture letter dataset

Letter'A' Number 7 Letter'V' Letter'L'

Fig. 11 Recognition Results of YOLOv5 Algorithm on Gesture Letter Dataset

The recognition results of the two algorithms on the gesture letter dataset are shown in Table 4.

Tab. 4 Recognition Results of Two Algorithms on Gesture Letter Datasets

Gesture	Recognition effect	Faster R-CNN	YOLOv5
Number '5'	categories	Number 5	Number 5
	accuracy	0.72	0.87
	Detection box positioning	inaccurate	accurate
Number '7'	categories	Number 7	Number 7
	accuracy	0.74	0.84
	Detection box positioning	accurate	accurate
Letter 'V'	categories	V	V
	accuracy	0.85	0.91
	Detection box positioning	accurate	accurate
Letter 'L'	categories	Unrecognized	L
	accuracy	×	0.90
	Detection box positioning	inaccurate	accurate

The recognition accuracy analysis of the two algorithms in the gesture letter dataset is shown in Table

5.

Tab.5 Analysis of Identification Rate Evaluation Indicators for Two Algorithms

Recognition algorithm	AP	Recall	Model size	Detection time
Faster R-CNN	77.64%	83.65%	108MB	0.194s
YOLOv5	88.76%	92.77%	13.7MB	0.113s

From Table 5, it can be seen that in complex scenarios, although the average recognition accuracy of the two algorithms has decreased, compared to the Faster R-CNN algorithm, the YOLOv5 algorithm also has higher recognition accuracy, more accurate target box positioning, and higher confidence. The YOLOv5 algorithm has better practicality and robustness in practical applications.

5. Conclusion

Gesture recognition has developed rapidly in recent years and has been increasingly applied in human-computer interaction. Compared with the Faster R-CNN network, the gesture recognition algorithm based on YOLOv5 proposed in this paper outperforms the Faster R-CNN network in gesture category recognition, confidence, detection box localization, and can accurately recognize gestures in complex background environments, which has better practicality and robustness.

References

[1] Liu XH, Hong Q, He YL, et al. Design of Gesture Recognition System Based on Data Gloves[J]. *Microcontrollers &Embedded Systems*, 2020, 20(06): 16-19.

[2] Liu K, Zhang JS. Research on the Application of Gesture Control System in Ground Penetrating Radar[J]. *Equipment for Geotechnical Engineering*, 2023, 24(03): 12-15.

[3] Grobel K, Assan M. Isolated sign language recognition using hidden makov model [C]// *IEEE International Conference on Computational Cybernetics and Simulation*, 1997, 1(1): 162-167.

[4] Fan WB, Zhu LJ. Research on hand gesture detection and recognition method based on skin color feature extraction[J]. *Modern Electronics Technique*, 2017, 40(18): 85-88.

[5] Oyedotun O K,Khashman A. Deep learning in vision-based static hand gesture recognition[J]. *Neural Computing and Applications*, 2017, 28(12): 3941-3951.

[6] Wu XF, Zhang JX, Xu XC. Hand Gesture Recognition Algorithm Based on Faster R-CNN[J].*Journal of Computer-Aided Design & Computer Graphics*, 2018, 30(03): 468-476.

[7] Wang C Y, BochkovskiyA, Liao H. Scaled-YOLOv4: Scaling Cross Stage Partial Network[C]// *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2021.

[8] Song SJ, Xia HJ, Li G.Research on Improved YOLOv5 Algorithm and Its Application in Multi-Object Detection for Automatic Driving[J]. *Computer Engineering and Applications*, 2023,59(15):68-75.

[9] Mao Z, Ren YM, Chen XY, et al. An Improved Multi-Scale Object Detection Algorithm for YOLOv5s[J].*Chinese Journal of Sensors and Actuators*, 2023,36(02):267-274.

[10] Wang X, Wang S. Traffic police gesture recognition based on improved YOLOv5 Algorithm[J]. *Electronic Measurement Technology*, 2022,45(02): 129-134.