Human Detection Based on Improved Mask R-CNN

Yuejuan Wang, Ji Wu and Heting Li
State Grid Jiangsu Electric Power Limited Company, Suzhou Power Supply Branch
No. 555 Laodong Road, Suzhou Jiangsu 215004, China
Email: 215691852@qq.com
Email: 13862159678@163.com
Email: loading.527@163.com

Abstract. Human detection has been widely concerned by academic and industrial circles. The gradual maturity of deep learning framework further improves the accuracy and speed of detection. However, the relatively mature human detection methods cannot get accurate detection results because of the complex background, shooting angle and many limitations of human behaviour. Aiming at solving the problems of human detection in complex scenes, a novel human detection algorithm based on the improved mask R-CNN framework is proposed by implementing the leading research results of object detection through deep learning. The algorithm combines ResNet and FPN to extract the features of the image, and then takes advantage of RoIAlign and fine-grained Slic to proofread the pixels. In the experiments that compared with the original mask R-CNN algorithm on the same data set was carried out to verify the effectiveness of the proposed algorithm. The mAP and AR value of the improved mask R-CNN algorithm is greater than that of the mask R-CNN algorithm when the IoU is 0.5-0.95, showing that the improved mask R-CNN framework was able to detect human from video better.

1. Introduction
Human detection has always been an important issue in computer vision research, which is to exploit computer vision technology to determine whether there are human bodies in the input image or video sequences and then quickly and accurately determine the location of the traveller, and is widely applied in intelligent monitoring, security, auxiliary driving and other fields [1]. Because the efficiency of human detection is affected by the complexity of background, different lighting conditions, wearing, posture, visual angle etc., the high quality image feature information can rarely be produced and the recognition rate and detection speed need to be improved.

The traditional human detection methods mainly rely on Haar feature [2], directional gradient histogram HOG [3], local binary mode LBP [4] and classifier [5]. However, these methods not only suffer from high complexity and poor robustness but also produce a large number of redundant candidate areas that may even lead to the failure of target detection. With the great breakthrough of deep learning in various fields, the upsurge of deep-learning-based target detection and recognition has emerged. At present, methods of human detection based on deep learning are mainly divided into two categories: regression-based method and region candidate frame ROI based method. The regression-based method mainly depends on the selection of default box and the design of loss function. The main representative models are Yolo [6] and single shot multi-box detector (SSD) [7]. Then Yolo V2 [8] came out, which not only kept the real-time property of Yolo, but also made up for the defect of low detection accuracy of Yolo. The emergence of SSD further proves the advantages of regression method, but because of the problems of missing detection, there are FSSD [9] and DSSN [10] methods.
The representative work of region based candidate box method is the region-based convolutional neural network (R-CNN) [11], which uses candidate region and classification convolutional neural network training model for detection. Its R-CNN algorithm follows the traditional idea of target detection, and also uses four steps of extracting frame, extracting features from each frame, image classification and non-maximum suppression to detect the target.

In this paper, we introduce fast R-CNN [12], faster R-CNN [13], mask R-CNN [14] and other deep network frameworks to solve the problem of repeated calculation of R-CNN model. With the gradual maturity of deep network framework, the requirements for human detection technology are further improved, that is, higher detection accuracy and speed. However, for the images where human bodies are in the complex scene or far away from the target, current detection methods are far from satisfactory. Considering that fine-grained image recognition [15-19] changes between and within classes due to highly similar subordinate categories, it is helpful to distinguish objects of different subclasses, to learn key feature parts of the target, and to learn human bodies features more accurately. The algorithm proposed in this paper is based on the mask R-CNN structure, which is applied to human detection and the fine-grained is introduced to reduce the error caused by complex scenes, so as to achieve fast and accurate human detection.

2. Related Work

2.1. Human Detection

Human detection is a hot spot in computer vision technology, which plays an important role in vehicle assistant driving and video monitoring. The detection of human detection is an extension of human detection, which is more challenging, mainly because people have different postures and body appearances, and the background of the human’s environment is generally complex, and some other external factors such as light change and occlusion.

The traditional detection methods are mainly based on background-modelling and statistical learning whose performance is greatly affected by the feature description and classifier of human. The main features of the human bodies in the description image are Haar [2] and gradient direction histogram [3], and the main classifiers used are neural network, support vector machine and boosting.

2.2. R-CNN

With the development of deep learning method in the field of object detection, the research of object detection based on CNN framework has attracted the attention of scholars. It is the first time to use CNN to achieve the classification of image pixels. Researchers began to use CNN for image segmentation, object detection, and make it the mainstream algorithm.

The Region-based CNN (R-CNN) [11] approach employed convolutional neural network to extract feature of extraction frame, and evaluate convolutional networks [20]. After that, on the basis of R-CNN, RoIPool [12] was introduced leading to higher speed and accuracy. Faster R-CNN [13] advanced further by learning the attention mechanism with a Region Proposal Network (RPN). However, due to the usage of candidate ROI and RPN, fast R-CNN is difficult to achieve real-time detection. In 2015, the appearance of ResNet [21] residual structure developed by Microsoft shortened the training time. Subsequently, RFCN [22] combines RPN and ResNet [23] to achieve high-precision detection.

3. Method

3.1. Mask R-CNN

The mask R-CNN algorithm extends the original fast R-CNN with a branch to use the existing detection to predict the target in parallel. At the same time, the network is often applied to other areas, such as target detection, segmentation, and key point detection. In this paper, the mask R-CNN algorithm is used to detect the human bodies in the image. The network structure is shown in Figure 1. First, the convolution feature of input human image of any size is extracted to form feature map. Then, in the region proposal network (RPN), the convolution layer is shared by the region, category and
expression to accelerate the calculation speed. The parallel feature pyramid network (FPN) distinguishes each ROI and divides the pixels of the human body target, and gives the coordinates of the traveller in the Figure 1.

![Figure 1. The flow chart of Mask R-CNN for instance segmentation [14].](image)

The basic structure of mask R-CNN consists of two steps. The first step is to generate target candidate regions and propose candidate boundary boxes; the second step is that mask R-CNN outputs binary masks for each region of interest, which are parallel to prediction classes and boundary box offsets, where classification depends on mask prediction. During training, the mask R-CNN algorithm defines the multi-task loss function on each ROI sampled as:

\[ L = L_{CLS} + L_{BOX} + L_{MASK} \]  \hspace{1cm} (1)

Where \( L_{CLS} \) represents classification loss, \( L_{BOX} \) represents boundary box loss and \( L_{MASK} \) represents segmentation loss.

The algorithm adopts the RoIAlign layer instead of the full connection layer, and it achieves one-to-one correspondence between output and input pixels. It uses bilinear interpolation method to get the image value on the pixel point whose coordinates are floating-point numbers, avoiding any quantization on the ROI boundary or interval, thus transforming the whole feature aggregation process into a continuous operation. In the specific algorithm operation, RoIAlign does not simply supplement the coordinate points on the boundary of candidate area for pooling, but through traversing each candidate area, keeps the floating-point boundary not to be quantified; secondly, the candidate area is divided into \( k \times k \) cells, and the boundary of each cell is not to be quantified; finally, four fixed coordinate positions are calculated in each cell, and bilinear interpolation is used which calculates the values of these four locations, and then performs the maximum pooling operation.

### 3.2. Improved Mask R-CNN

In this paper, based on the original mask R-CNN detection algorithm, some improvements are made, such as adding fine-grained to improve the prediction effect of the network. The network structure is shown in Figure 2, which is mainly divided into three parts: the first part is to generate candidate areas; the second part is to learn the features of extraction box; the third part is to use fine-grained segmentation again to further effectively separate the target and background information, see section 3.2.3 for details.

#### 3.2.1. Area generation network

RPN is used to generate all possible target candidate regions, which solves the problem that it takes long time to generate detection frame. RPN generates a feature map based on convolutional neural network, which generates multiple anchors on the scale of the original image, and then classifies and regresses the generated anchors.
RPN uses a sliding window to scan the whole image to find the existing target area. For each position on the image, K anchors are predicted at the same time, and each anchor generates two outputs, that is, anchor category and border adjustment. For overlapping multiple anchors, non-maximum suppression is used to give rough results of the target, and the anchor with the highest target score is retained. Therefore, the best anchor with the target can be selected by RPN modular.

3.2.2. FPN
Since objects in images are of various sizes, using image pyramid network to detect can adjust the image scale and make the convolution neural network easier to process. The FPN of mask R-CNN first obtains four sets of feature maps through convolution neural network from bottom to top. In order to solve the problem that the features of different levels of convolution neural network differ greatly, the top-down and horizontal connection methods are used to integrate these four sets of different feature maps in Figure 3, so that each layer of the fused network has deep and shallow features. In this module, we assign ROI with width \( w \) and height \( h \) to the \( P_k \) level of the feature pyramid:

\[
k = \left\lfloor k_0 + \log_2(\sqrt{wh} / 224) \right\rfloor
\]  

(2)

Where \( k_0 \) is the reference value.

3.2.3. Fine-grained Slic
For fine-grained Slic [24] super-pixel segmentation, the image should be transformed into five-dimensional feature vectors \((L, A, B, X, Y)\)^1 normalized by CIELAB chromaticity space and space coordinate XY, and the image pixels should be locally clustered to generate similar and compact super-pixel blocks, so as to complete image super-pixel segmentation. The super pixel block generated by this algorithm has uniform shape and size, good compactness and boundary fit. The mean value of all the pixels in the Mask R-CNN segmentation results is obtained. According to the mean value, the adjacent and similar super-pixel blocks in the fine-grained super-pixel segmentation image \( Q_k \) are combined to reconstruct a more accurate target contour.

\[
Q_k = S_{th} (M_k)
\]  

(3)
\[ P_N = S_{con}(Q_K, A_L) \]  \hspace{1cm} (4)

where \( S_{div} \) is the fine-grained super-pixel segmentation function; \( M_n \) is the coarse-grained segmentation image divided into \( N \) equal parts; \( Q_K \) is the fine-grained super-pixel segmentation image with \( K \) tags; \( S_{con} \) is the Slic merge function; \( Q_K \) is the fine-grained super-pixel segmentation image with \( K \) tags; \( P_N \) is the reconstructed target object segmentation result; \( \text{AL} \) is the mean value of the center of mass in all the pixel blocks of the Mask R-CNN color segmentation result.

The algorithm of Human Detection based on Improved Mask R-CNN is as follows.

**Algorithm: Human Detection based on Improved Mask R-CNN**

**Input:** image \( M \), slic merge function \( S_{con} \), fine-grained super-pixel segmentation function \( S_{div} \), threshold \( \epsilon_0, \epsilon_1 \)

**Return:** reconstructed image

1: divide the image \( M \) into \( n \) equal parts \( M_n \)
2: \( P_0 \leftarrow 0 \)
3: Repeat
4: \( \text{iter} + 1 \)
5: for \( i = 1 \) to \( n \) do:
6: calculate the super pixel block information of \( M_i \) by \( Q_K = S_{div}(M_i) \)
7: calculate \( \text{loss} = \text{difference}(Q_K, \text{norm}_{X1}(M_i)) \)
8: if \( \text{loss} < \epsilon_0 \):
9: merge \( M_i \) super pixel block with adjacent super pixel block by \( P_n = S_{con}(Q_K, A_L) \)
10: \( \nabla P \leftarrow \| P_n - P_0 \| \)
11: until \( \nabla P < \epsilon_1 \)

4. Result

4.1. Data Set
We use Penn-Fudan Database, which were specified for human detection in the experiments reported in [25]. There are 170 images with 345-labeled human bodies, among which 96 images are taken from around University of Pennsylvania, and other 74 are taken from around Fudan University. The images are taken from scenes around campus and urban street, and each image will have at least one human in it.

4.2. Parameters Settings and Metrics
In this paper, the improved mask R-CNN structure is used as the model to train the human detector. In order to speed up the training and prevent over fitting, the parameters settings during the training are shown in Table 1.

| Parameter       | Description                  | Value   |
|-----------------|------------------------------|---------|
| lr              | Learning-rate                | 0.005   |
| momentum        | the momentum coefficient     | 0.9     |
| decay           | weight attenuation coefficient| 0.0005  |

The evaluation indexes used are the accuracy IOU [26] of predicted class pixels and correct class pixels which is defined as follows:
\[ IoU(i) = \frac{\sum_{j} n_{ij}}{(n_{ii} + n_{ji}) - n_{ij}} \]  

(5)

Where \( n_{ii} \) represents the number of correctly segmented pixels; \( n_{ij} \) represents the number of pixels whose category \( i \) is predicted as category \( j \); \( n_{ji} \) represents the number of pixels whose category \( j \) is predicted as category \( i \); \( N \) represents the number of categories.

### 4.3. Analysis of Experimental Results

We trained and tested on the Penn-Fudan data set, and results of mask R-CNN and the improved mask R-CNN algorithm are shown in Table 2 and Table 3. Bbox and segm represent the boundary boxes of human detection and segmented pixels of human detection respectively. The mAP value of the improved mask R-CNN algorithm is greater than that of the mask R-CNN algorithm when the IoU is 0.5-0.95. The detection accuracy of bbox under different thresholds is increased by 2.45%, 0.4%, 2.23%, and the detection accuracy of segm is increased by 4.12%, 0.41%, 4.83%, respectively. For AR, the improved mask R-CNN algorithm improves bbox and segm by 1.64% and 3.7% respectively. The results show that the improved mask R-CNN algorithm effectively improves the accuracy of human detection.

| Table 2. Average Precision on bbox and segm |
|-------------------------------------------|
| Threshold | Mask R-CNN | Improved Mask R-CNN |
|-----------|------------|---------------------|
| bbox      |            |                     |
| 0.50:0.95 | 0.815      | 0.835               |
| 0.5       | 0.991      | 0.995               |
| 0.75      | 0.943      | 0.964               |
| segm      |            |                     |
| 0.50:0.95 | 0.752      | 0.783               |
| 0.5       | 0.981      | 0.985               |
| 0.75      | 0.911      | 0.955               |

| Table 3. Average Recall on bbox and segm |
|-----------------------------------------|
| Mask R-CNN | Improved Mask R-CNN |
|------------|---------------------|
| bbox       | 0.852               | 0.866               |
| segm       | 0.784               | 0.813               |

The visualization results of this experiment are shown in Figure 4. A typical image is selected from the dataset, and the human bodies in the image are marked correctly with a rectangular frame.

5. Conclusion

We presents an improved mask R-CNN algorithm for human detection in this paper. Based on the
initial mask R-CNN framework, we combine ResNet, FPN and other frameworks to improve the
detection ability of human bodies, and introduce fine-grained layer to improve local characteristics,
effectively removing redundant information and error fusion target. The experimental results in Penn-
Fudan data set show that the proposed method effectively improves the accuracy of target detection.

6. References

[1] Miao H, Haibo L, Zheng C, et al. Human detection with Semantic Regions of Interest[J]. Sensors, 2017, 17(11):2699-2708.
[2] Panning A, Alhamadi A, Niese R, et al. Facial expression recognition based on Haar-like feature detection[J]. Pattern Recognition and Image Analysis, 2008, 18(3): 447-452.
[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Computer Vision and Pattern Recognition, 2005: 886-893.
[4] Wang X, Han T X, Yan S, et al. An HOG-LBP human detector with partial occlusion handling[C]. International Conference on Computer Vision, 2009: 32-39.
[5] Girshick R, Donahue J, Darrell T, et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.
[6] Redmon J, Divvala S K, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. Computer Vision and Pattern Recognition, 2016: 779-788.
[7] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. European Conference on Computer Vision, 2016: 21-37.
[8] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. Computer Vision and Pattern Recognition, 2017: 6517-6525.
[9] Zhuang X, Zhang T. Detection of sick broilers by digital image processing and deep learning[J]. Biosystems Engineering, 2019: 106-116.
[10] Fu C, Liu W, Ranga A, et al. DSSD: Deconvolutional Single Shot Detector[J]. arXiv: Computer Vision and Pattern Recognition, 2017: 1-11.
[11] Girshick R, Donahue J, Darrell T, et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.
[12] Girshick R. Fast R-CNN[C]. International Conference on Computer Vision, 2015: 1440-1448.
[13] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. Neural Information Processing Systems, 2015: 91-99.
[14] He K, Gkioxari G, Dollar P, et al. 14R-CNN[C]. International Conference on Computer Vision, 2017: 2980-2988.
[15] Huang S, Xu Z, Tao D, et al. Part-Stacked CNN for Fine-Grained Visual Categorization[C]. Computer Vision and Pattern Recognition, 2016: 1173-1182.
[16] Lin D, Shen X, Lu C, et al. Deep LAC: Deep localization, alignment and classification for fine-grained recognition[C]. Computer Vision and Pattern Recognition, 2015: 1666-1674.
[17] Zhang Y, Wei X, Wu J, et al. Weakly Supervised Fine-Grained Categorization With Part-Based Image Representation[J]. IEEE Transactions on Image Processing, 2016, 25(4): 1713-1725.
[18] Xie G, Zhang X, Yang W, et al. LG-CNN: From local parts to global discrimination for fine-grained recognition[J]. Pattern Recognition, 2017, 71(71): 118-131.
[19] Lee S H, Chan C S, Mayo S J, et al. How deep learning extracts and learns leaf features for plant classification[J]. Pattern Recognition, 2017: 1-13.
[20] Lecun Y, Boser B E, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
[21] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. Computer Vision and Pattern Recognition, 2016: 770-778.
[22] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. arXiv: Computer Vision and Pattern Recognition, 2016.
[23] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. Computer Vision and Pattern Recognition, 2016: 770-778.
[24] Achanta R, Shaji A, Smith K, et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.

[25] Wang L, Shi J, Song G, et al. Object detection combining recognition and segmentation[C]. Asian Conference on Computer Vision, 2007: 189-199.

[26] Rezatofighi H, Tsoi N, Gwak I, et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]. Computer Vision and Pattern Recognition, 2019: 658-666.