2020 International Conference on Computer Science and Management Technology (ICCSMT) | 978-1-7281-8668-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCSMT51754.2020.00034

Mask wearing detection method based on SSD-Mask algorithm

Mingyuan Xu
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China
e-mail: 824009787@qq.com

Heng Wang
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China
e-mail: 595390897@qq.com

*Shuqun Yang
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China
* Corresponding author e-mail: shuqyang@sues.edu.cn

Rui Li
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China
e-mail: 1024549774@qq.com

***Abstract*—With the spread of COVID-19 all over the world, under the background of normalized epidemic prevention and control, proper wearing and using of masks can effectively filter virus pathogen particles. In order to reduce the infection rate of ordinary people and improve the efficiency of monitoring people wearing masks in public places, this paper improves a method of face mask wearing detection in natural scenes. On the basis of SSD algorithm, SSD-Mask introduces a channel attention mechanism to improve the ability of the model to express salient features. At the same time, the information of different feature levels is fully utilized, and the loss function is optimized. The final experimental results show that the algorithm can effectively achieve the goal of face recognition and mask detection.**

Keywords- Face mask detection; SSD algorithm; channel attention mechanism; mask standard wearing detection; face recognition

I. INTRODUCTION

With the spread of Corona Virus Disease 2019 (COVID-19), the economic development of society and people's production and life are not affected.[1] COVID-19 is highly infectious, and it can be transmitted through contact or droplets and aerosols in the air, and can survive for 5 days in suitable conditions.[2-4] In order to effectively prevent the spread of the virus, it is an effective defense means to measure body temperature and wear masks correctly when entering and leaving public places. "Novel coronavirus pneumonia prevention guidelines" issued by the national health and health committee emphasizes that when wearing personal public transport or traveling by public transport, a person must wear a standard surgical mask or N95 mask.[5] Compared with pedestrian detection and face recognition, mask wearing detection not only needs to recognize the face, but also needs to accurately judge whether the face is wearing the mask correctly or wearing the mask incorrectly. Therefore, the ability of the detection algorithm to extract the detailed features has higher requirements.

At present, there is little research on the algorithm of face mask wearing detection, and there is still a lot of room for development in the field of intelligent detection of face mask wearing. SSD (Single Shot Multibox Detector) algorithm is a target detection algorithm with VGG as the front-end network.[6] It will evenly distribute on the image to produce candidate frames of different sizes and aspect ratios. Then, the convolution layer is used to extract image features, and finally regression and classification are carried out. SSD algorithm has the advantages of fast detection speed and strong comprehensive performance, but when it is directly applied to some specific detection objects, the detection effect can not meet the requirements due to the complexity of the scene and the diversity of features, so it is necessary to improve it. Ding Fenglong and others improved the SSD algorithm and realized the defect detection on solid wood board with high accuracy.[7] In order to solve the problem of missing detection in SSD, Yang Dawei et al. Improved the SSD algorithm and fused the feature layer to supplement the object contour, improve the expression ability of object feature information, enhance the detection ability and reduce the missed detection rate.[8] Jian Taoyu et al. Combined SSD target detection algorithm with lightweight neural network mobilenetv2, which greatly improved the detection speed of system model and improved the overall recognition efficiency of indoor service robot to a certain extent.[9]

At the same time, through the research of target detection algorithms in related fields, we find that there are many similarities between mask wearing detection and deep learning face recognition. A new algorithm pyramid is proposed in reference[10] The algorithm is a context assisted single detection algorithm. It learns the feature information of mask, head and body, and uses Low Feature Pyramid Networks(LFPN) to integrate high-level semantic information and low-level mask features to achieve the effect of predicting all scale faces. Finally, context sensitivity is introduced to increase the capacity of the prediction network to improve the output the accuracy of the final mask wearing detection. The reference[11] shows that a single shot face detection network algorithm based on feature fusion and segmentation supervised detection is designed. The algorithm introduces more effective feature fusion pyramid and segmentation branch to solve the problem that the face position

978-1-7281-8668-9/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCSMT51754.2020.00034

can not be accurately located. The semantic information from higher level feature map is applied as context prompt to enhance the low-level through the attention mechanism of space and channel Feature map can increase the recognition degree of mask and improve the accuracy of detection. In reference[12], the data set was trained by migration learning and improved RetinaNet network. Finally, good results were obtained in the test.

SSD network structure is a typical algorithm in the field of target detection, which has the advantages of fast processing speed and high detection accuracy. However, in the actual application of mask wearing target detection, due to the complexity of the scene, the difference between wearing mask correctly and wearing mask incorrectly is small, so the detection effect of SSD algorithm is not ideal. Especially in the detection of small targets, the detection accuracy is not high. SENet (Squeeze and Exception Networks)[13] is a kind of network structure which pays attention to the relationship between feature channels. The expression ability of network model can be improved by modeling the interaction between each channel of convolution feature. Applying it to the detection of mask wearing target, the channel weight of important features can be enhanced, so as to improve the detection effect. In this paper, SeNet is introduced into SSD algorithm, and a SSD-mask wearing detection algorithm based on VGG-16 network and adding deconvolution and feature fusion is proposed. The algorithm enhances the ability of context description and feature map expression, and combines mask wearing detection algorithm, mask standard wearing detection algorithm and face recognition algorithm. Experimental results show that the improved SSD algorithm has better accuracy and robustness. The logic structure of the intelligent detection system for mask wearing is shown in Figure 1.

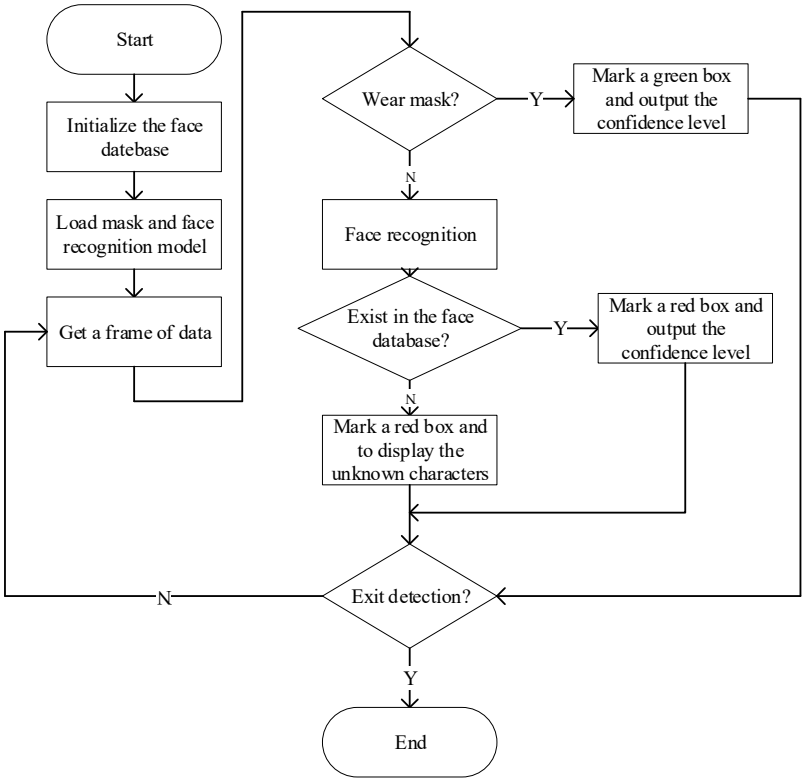

Figure 1. The system logic diagram

II. SSD-MASK ALGORITHM RELATED MODEL

A. SSD algorithm principle

SSD algorithm framework belongs to one stage method. It mainly relies on uniform and dense feature extraction on the detected target, and then classifies these features for kernel regression processing. It is a direct prediction target category and multi box prediction algorithm. The structure of SSD algorithm is shown in Figure 2. Firstly, image features are extracted through VGG-16[14] network. VGG-16 network is a kind of CNN network, which is based on convolution kernel and extracts image features through multiple convolution layers. SSD algorithm structure map predicts the target from different scales, and uses six different feature maps to detect different size targets. The six detection layers are convolution layers, and each convolution layer corresponds to a sliding window of different sizes. The higher layer predicts the larger target, and the lower layer predicts the smaller target. In this way, all targets of different scales can be detected completely.

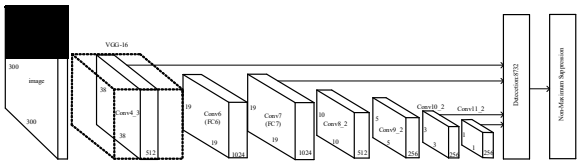

Figure 2. The structure of SSD algorithm

The traditional CNN network uses convolution kernel with fixed size and the same number of input layers to obtain a layer of characteristic data after convolution, and the calculation formula of each data in the characteristic layer is as follows:

$$C_{\mathrm{p}} = D \times F \times F \tag{1}$$

In the formula 1: D is the number of input data volume channels; F is the size of convolution kernel. In the case of padding operation, the calculation formula of each CNN layer is as follows:

$$C_{CNN} = K \times D \times F \times F \times N \times N \tag{2}$$

In the formula 2: K is the number of convolution kernels in the current layer; N is the width and height of the input data (the same).

After feature extraction, the feature is input into the subsequent network to calculate the target location and target type. For different sizes of targets, the previous solution is to transform the original image to different scales, and SSD discretizes the output space of bounding box into a series of default boxes, which is similar to anchor in Faster RCNN. Different layers will output different default boxes, and finally summarize them for non maximum suppression. In this way, the combination of different scales of the box can effectively deal with multi-scale problems.

B. SENet module

By modeling the dependency relationship between channels, the SENet module calibrates the weights of each channel, so as to recalibrate the characteristics of each channel, so as to selectively enhance the channel with stronger feature representation ability. The structure of SENet is shown in Figure 3.

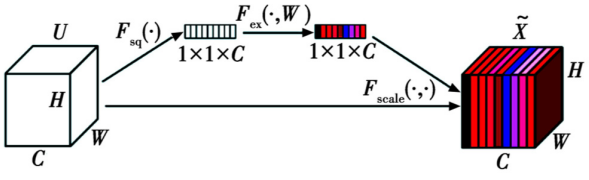

Figure 3. The structure of SENet

In Figure 3, U can be regarded as the output data volume of any convolution layer. By embedding the SENet module to calculate the weight value of each channel, it is obtained that the dimension of output $\tilde{X}$ remains unchanged.

The calculation process of SENet module is mainly divided into two steps.

1) The global information embedding (squeeze) compresses each feature map and compresses the input features from the spatial dimension, which is equivalent to pool operation with global receptive field. The feature data of W × h × C is compressed into 1 × 1 × C one-dimensional vector.

2) SENet captures the dependency relationship between channels by adaptive exception. It includes two full connection layers, relu and sigmoid. The calculation process is

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{3}$$

In the formula, Z is the output of squeeze process; σ is sigmoid function; δ is ReLU function; W_1 and W_2 are two full connection layers; output vector s is the importance weight of each channel learned by module through network training. Multiply S with the corresponding channel of input U to complete the calculation of SENet module.

The SENet module can effectively improve the performance of convolution network by embedding the existing network. In this paper, we build a new basic network by embedding the SENet module into the VGG-16 network structure.

III. DETECTION ALGORITHM OF SSD-MASK WEARING

A. SSD-Mask network structure

In this paper, VGG-16 is used as the basic network of SSD mask algorithm. Firstly, the two full connection layers of VGG-16 are replaced by convolution layers, and the depth of convolution layer is increased, and the four convolution layers Conv_8 _ 2, Conv9_ 2, Conv10_ 2, Conv11_ 2 are added at the end. Through the SENet module, the process of extrusion and excitation is realized. The formula of extrusion is as follows:

$$Y = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i, j) \tag{4}$$

H, W and C in the formula represent the length, width and channel number of input dimensions of low resolution and high semantic information graph X respectively. After squeezing the information graph X, a group of arrays with length C can be obtained. The (i, j) in formula (4) represents the (i, j) points on the characteristic graph of size H * W, and the output y is a one-dimensional array of length C. The following excitation process is actually to model the correlation between channels, and the formula is as follows:

$$S = Sigmoid(W_2 \bullet \mathrm{Re}LU(W_1 Y)) \tag{5}$$

In the formula, the dimension of W_1 is C * C, the dimension of W is C * C, and C is C * 1 / 4. After training and learning these two weights through relu activation function and sigmoid function, a one-dimensional incentive weight can be obtained, which is used to activate each layer of channels. Finally, the dimension of S is C * 1 * 1.

Finally, the spatial relationship of feature space is used to generate attention feature map, which mainly focuses on the feature information of target location. In order to calculate the spatial attention feature map, average pooling and maximum pooling operations are usually applied along the channel, and

they are connected to generate effective feature descriptors. In the information region, the application of pooling operation can effectively improve the saliency of target features in the channel. The formula is as follows:

$$X' = X \bullet S \tag{6}$$

Replace the original input X· with the feature graph X obtained by the attention mechanism module and send it to the SSD mask network in this paper for detection. In other words, this process is actually a process of scaling down. The values of different channels are multiplied by different weights to enhance the attention to the key channels. The algorithm flow is shown in Figure 4.

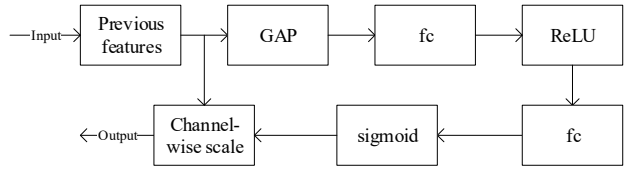

Figure 4. SSD-Mask algorithm flow chart

Firstly, the global average pooling(GAP) of image features is started, and then the feature values in each channel are added and averaged. Through the full connection layer, the pooled features are synthesized, and the two weights are trained by the ReLU function and the sigmoid function respectively. Finally, a one-dimensional excitation weight is obtained and each channel is activated. Finally, with the original input feature map, the scale function is used to enlarge and shrink the image features according to different weights, so as to pay attention to the mask wearing feature information that we need to pay attention to, so as to improve the mask wearing detection rate.

B. Loss function

The target loss function of SSD-Mask is $L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g))$, where n is the number of matched default boxes. If n is 0, set loss = 0. X for input image; C for multi category confidence level of target; l for prediction box; g for calibrated real data; L_{conf} for confidence loss and L_{loc} for location loss.

$$L_{loc}(x,l,g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^{k} smooth_{L1}(l_i^m - g_j^m) \tag{7}$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx}) / d_i^{w} \tag{8}$$

$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy}) / d_i^{h} \tag{9}$$

$$\hat{g}_j^{w} = \log(g_j^{w} / d_i^{w}) \tag{10}$$

$$\hat{g}_j^{h} = \log(g_j^{h} / d_i^{h}) \tag{11}$$

In the formula: L_{loc} is the $smooth_{L1}$ loss between prediction box L and calibrated real data g; g_j^{cx} and g_j^{cy} are the abscissa and ordinate of the center point of the jth calibrated real data; d_i^{cx} and d_i^{cy} are the horizontal and vertical coordinates of the center point of the i-th default box; D_i^{w} and D_i^{h} are the width and height of the i-th default box; G_j^{w} and G_j^{k} are the width and height of the j-th calibrated real data. It can be seen from the calculation expression of, that regression does not directly get the center point coordinates and width and height of the prediction box, but the offset values relative to the default box, but can be obtained by simple transformation,

$$L_{conf}(x,c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log(\hat{c}_i^{p}) - \sum_{i \in Neg} \log(\hat{c}_i^{0}) \tag{12}$$

$$\hat{c}_i^{p} = \frac{\exp)(c_i^{p})}{\sum_{p} \exp(c_i^{p})} \tag{13}$$

In the formula: L_{conf} is the softmax loss of multi category confidence C; x_{ij}^{p} is the identification of whether the i-th default box matches the j-th calibrated real data of category P, with the value of {0,1}; $\sum_{i} x_{ij}^{p} \geq 1$ the confidence level of the category of the i-th default box; c_i^{p} is the output of softmax of the class confidence of the i-th default box; $\hat{c}_i^{p}$ is the confidence of the background class of the ith default box; Pos and Neg represent positive and negative sample sets respectively.

IV. Experiment and Result Analysis

A. Experimental environment

The experimental environment configuration is as follows: Intel i7-9750H processor, NVIDIA geforce RTX 2060 video card, 16GB memory, and the operating system is Ubuntu 16.04. Python programming language is used to train and test the network, and the framework of deep learning is pytorch.

B. Data set construction and model training

Due to the fact that the data set of face mask wearing pictures is still relatively small and the quality of photos is not high, this paper uses two open-source face image data sets, including WIDER Face and MAFA, and collects them through personal shooting and Internet collection. The data set contains 3000 pictures of face mask wearing in various scenes. The data set images include three cases of people wearing mask correctly, wearing mask with occlusion and not wearing mask in different scenes, including partial occlusion and shadow, as shown in Figure 5.

Figure 5. Sample dataset image

In order to solve the problem of sample imbalance caused by the relatively small number of wrong wearing mask pictures in the data set, this paper uses random cutting, rotation, color transformation and other methods to enhance part of the data, and expand the data set to 4500, which makes the diversity of samples increase, improve the quality of samples, and effectively alleviate the over fitting phenomenon in the training process.

In the data set, 4000 images were randomly selected as training samples and 500 images as test samples. In this paper, the Adam optimizer is used to optimize the network. The initial learn rate is set to 0.001, and the adaptive adjustment strategy is used to dynamically adjust the learning rate. The adjustment factor is 0.5, and the patience is 2. When the two Epoch indexes do not change, the learning rate is adjusted to factor times of the current learning rate. In order to prevent over fitting, the training process was terminated when six consecutive epoch evaluation indexes did not change. In this paper, a batch contains 4 images and 120 epoch are trained. Finally, the loss converges to about 3. The parameter settings are shown in Table 1.

TABLE I. SELECTION OF KEY PARAMETERS

Batch size	Factor	Patience	Learn rate
4	0.5	2	0.001

C. *Evaluating indicator*

In order to verify the effectiveness of the training model for wearing masks, the detection effect of SSD algorithm before and after the improvement is compared, and the detection performance of the two algorithms in complex situations is compared, and the experimental conclusion is drawn.

Because of the need to intuitively evaluate the recognition results, the precision(P) and recall(R) rate should be considered in the evaluation index. Finally, the average precision (AP) and mean average precision (map) are selected as the evaluation indexes of the target detection algorithm. Both precision and recall are taken into account.

$$P(\text{classes}) = \frac{T_P}{T_P + F_P} \times 100\% \tag{14}$$

$$P(\text{classes}) = \frac{T_P}{T_P + F_P} \times 100\% \tag{15}$$

Set J as the evaluation index of mask wearing recognition results, J is the function of recall rate and precision rate, which is defined as formula (16):

$$J = 2PR/(P+R) \tag{16}$$

Since the target detected in this paper is mask recognition, T_P in the above formula represents the correct number of mask wearing targets in the detection model, F_P represents the number of targets with or without masks as errors, and F_N represents the number of objects correctly wearing masks as errors.

D. *Experimental analysis*

500 images randomly selected from the data set were used as the test set. The experimental results are compared with the original SSD target detection algorithm. The test results are shown in figure 6.

Figure 6. Algorithm comparison detection results. (a)(b) SSD detection results; (c)(d) SSD-Mask detection results

From the detection results of the two algorithms, SSD-Mask algorithm is more effective than SSD algorithm in the detection of small targets wearing masks. It not only ensures the detection accuracy, but also improves the detection rate of small targets and occluded targets, and has strong robustness.

The test results of the improved SSD-Mask network are compared with the original SSD algorithm in Table 2.

TABLE II. COMPARISON OF DETECTION RESULTS

Model	P/%	R/%	F_1/%	Detection time/ms
SSD	88.6	84.2	86.3	26.8
SSD-Mask	90.2	86.5	88.2	16.3

Table 2 shows that compared with the original SSD algorithm, the SSD-Mask model proposed in this paper has greatly improved the detection speed, while the performance of both precision and recall has been improved.

The SENet module has a simple structure and less parameters, so it can improve the speed of operation obviously. Meanwhile, by calculating the importance weight of different feature channels and redistributing them to each channel, the weight of feature layer which can play a more important role in small target wearing mask is enhanced. At the same time, through the end-to-end training of the whole detection

framework, we can learn which features are more useful for pedestrian classification and spatial information regression, and improve the performance of the whole detection framework.

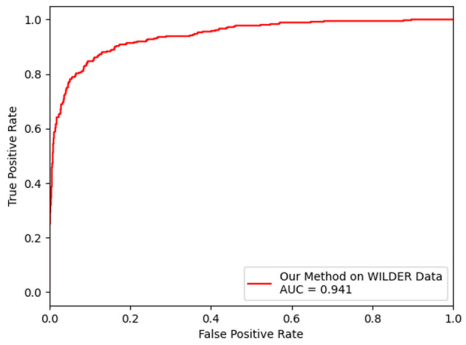

Figure 7. ROC curve characteristic diagram

The SSD-Mask model was tested and trained on WIDER Face dataset, and the performance of the model was evaluated by ROC curve. The figure7 shows that the ROC curve rises rapidly from 0.4, and the AUC area reaches 0.941 after the curve is stable, indicating that the classification of the model is obvious and the effect is remarkable.

V. Conclusion

In this paper, a mask wearing detection method based on SSD-Mask algorithm is proposed to solve the problem of small target and occluded target detection. Firstly, the SENet module is introduced to improve the VGG-16 network through the channel attention mechanism to improve the channel weight of useful features, and strengthen the use of low-level features for small target detection, so as to improve the detection effect of mask wearing. Finally, the data set of mask wearing in natural scene is established, and the algorithm is verified on the data set. The experimental results show that SSD-Mask algorithm has good accuracy and robustness. Compared with the original SSD algorithm, this algorithm has achieved good detection results in the detection of mask wearing, and has broad application prospects in epidemic prevention and control.

References

[1] Ahmed N J , Alrawili A S , Alkhawaja F Z . The Anxiety and Stress of the Public during the Spread of Novel Coronavirus (COVID-19)[J]. 2020.

[2] Jasper Fuk-Woo Chan,Shuofeng Yuan,Kin-Hang Kok,Kelvin Kai-Wang To,Hin Chu,Jin Yang,Fanfan Xing,Jieling Liu,Cyril Chik-Yan Yip,Rosana Wing-Shan Poon,Hoi-Wah Tsoi,Simon Kam-Fai Lo,Kwok-Hung Chan,Vincent Kwok-Man Poon,Wan-Mui Chan,Jonathan Daniel Ip,Jian-Piao Cai,Vincent Chi-Chung Cheng,Honglin Chen,Christopher Kim-Ming Hui,Kwok-Yung Yuen. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster[J]. Elsevier,2020(prepublish).

[3] Bai Lang, Wang Ming, Tang Xiaoqiong, et al. Thinking on hot issues in the diagnosis and treatment of new coronavirus pneumonia[J]. West China Medicine, 2020, 35(2): 125-131.

[4] Chan J F，Yuan Shuofeng，Kok K H，et al.A familial clus-ter of pneumonia associated with the 2019 novel coro-navirus indicating person-to-person transmission：a study.of a family cluster[J].The Lancet，2020，395（10223）：514-523.

[5] Ling L , Taisheng L . Interpretation of "Guidelines for the Diagnosis and Treatment of Novel Coronavirus (2019-nCoV) Infection by the National Health Commission (Trial Version 5)"[J]. National Medical Journal of China, 2020, 100(00):E001-E001.

[6] Jeong J , Park H , Kwak N . Enhancement of SSD by concatenating feature maps for object detection[J]. 2017.

[7] [1] Ding Fenglong College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. Zhuang Zilong College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. Liu Ying College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. Jiang Dong College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. Yan Xiaoan College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. Wang Zhengguang College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China. . Detecting Defects on Solid Wood Panels Based on an Improved SSD Algorithm.[J]. Sensors (Basel, Switzerland),2020,20(18).

[8] [1]Dawei Yang,Cheng Bi,Lin Mao,Rubo Zhang. Contour feature fusion SSD Algorithm[A]. Technical Committee on Control Theory, Chinese Association of Automation, Chinese Association of Automation, Systems Engineering Society of China.(3)[C]. Technical Committee on Control Theory, Chinese Association of Automation, Chinese Association of Automation, Systems Engineering Society of China, 2019:4.

[9] [1]Jiantao Yu,Weiping Hu. Static Target Recognition of Indoor Mobile Robot Based on Improved SSD Algorithm[P]. Information Technologies and Electrical Engineering,2019.

[10] Tang X , Du D K , He Z , et al. PyramidBox: A Context-assisted Single Shot Face Detector[J]. 2018.

[11] TIAN W,WANG Z,SHEN H,et,al. Learning Better Features for Face Detection with Feature Fusion and Segmentation Supervision[J/OL]. arXiv preprint,2018,http://arxiv.org/abs/1811.08557.

[12] Fu C Y , Shvets M , Berg A C . RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free[J]. 2019.

[13] Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[14] Polyakov, A. Nonlinear feedback design for fixed-time stabilization of Linear Control Systems[J]. IEEE Transactions on Automatic Control, 2012, 57(8):2106-2110.