Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220527

Face detection and recognition of mask wearing in normal environment based on neural network

Yanyan Li[1], Shubing Xie[2,5], Yifei Ren[3] and Xinyu Li[4]

[1]School of Information Engineering Shanghai Maritime University, Shanghai, 201306, China
[2]School of Physics, Xi'an JiaoTong University, Xian, Shaanxi, 710049, China
[3]University of Liverpool, Liverpool L1 8JX, UK
[4]Department of International Baccalaureate Diploma Programme, Shanghai World Foreign Language Academy, Shanghai, 200233, China

[5]xsb19990228@stu.xjtu.edu.cn

Abstract. With the outbreak of COVID-19, wearing masks has become a hot topic again. In public places, Not wearing a mask correctly has almost the same harm as not wearing a mask. Mask detection is an extension of face detection. Therefore, it is of great practical significance to design a system that can correctly identify whether pedestrians wear masks correctly. The face recognition technology based on neural network has been relatively mature, but there is still a lack of work related to mask recognition, especially whether to wear masks correctly. This is not only the promotion of a two-classification problem to a three-classification problem, but also faces many practical problems, including data set acquisition, in this paper, mask recognition is divided into a multi-stage work. In the first link, Yolo network is used to recognize facial areas, and in the second link, RESNET is used to realize mask recognition. Finally, a layer of RESNET network is improved to achieve higher recognition accuracy.

Keywords: COVID-19, face detection, mask detection, Yolo network, RESNET

1. Introduction

With the outbreak of the new crown epidemic, health issues have attracted people's attention, and the smooth progress of epidemic prevention work is particularly important. The spread of the virus is very prone to infection in public places with dense human traffic, and the effective way to prevent the spread of the virus is to wear a mask. In public, if the mask is not worn correctly, the degree of harm is no less than that of not wearing a mask [1]. Therefore, the paper decided to adopt a machine learning method based on a neural network mechanism to make a mask recognizer to detect whether the subject is wearing a mask correctly, saving human resources for public health announcements, and providing technical guarantees. (Figure 1 shows the diagram of whether to wear a mask correctly).

DOI: 10.54254/2755-2721/2/20220527

Figure 1. Schematic diagram of whether to wear a mask correctly.

There are two points in the work. The first is to find a data set that can correctly distinguish whether the mask is worn correctly. This may be a little difficult at first, so we have manually processed some face data sets and artificially processed the facial area. Covers the mask; the second is our training on the network and the definition of whether to wear a mask correctly. Considering most of the public, a sanitary product is defined as a mask. Generally, it is worn on the nose and mouth to filter the entrance. The air in the nose is made of gauze or paper to block harmful gases, odors, droplets, viruses, and other substances. The improper wearing is limited to failure to cover specific organs such as the nose, mouth, or chin; for attempts to deceive the detector with a mask shaped like a mask, such as a human hand or a piece of cardboard, not considering for the time being. After the experiments and improvements, the work has basically realized the basic function of mask detection. In the test of samples. Finally, through the combination of dense net and yolo network, it can be achieved a recognition accuracy of up to 99.10 %

2. Related work

Mask detection is really a current hot topic during Covid outbreak, and many scholars have been working on their own approach. Generally, most works focus on adjusting existing face detection neural networks, amending them to detect masks. The group leading by G. Jignesh Chowdary applied transfer learning model [2], InterceptionV3, to judge the existence of a mask on a specific face.

Their result is quite satisfying, with 99.92% training accuracy and 100% testing accuracy while using SMFD dataset. Still, they commit that their data set is limited and simple, leading to the results of low universality of the model in complex environments. Furthermore, It is pointed out that only judging the existence of masks is not practical enough, since conditions of improper wearing may occur and the masks cannot function well resisting the virus.

There are also scholars who are keen on analyzing the effect of the mask on people's voice to solve the mask detection problem. Nicolae-Catalin Ristea and Radu Tudor Ionescu constructed a GANs model to improve mask detection [3]. Their main contribution is to advance a novel data augmentation to process the data. Plus, the creative way to solve the question in a voice detection way to judge whether people are wearing masks eliminates the misjudgment of wearing the mask properly. However, the weakness is also evident. The voice approach requires very critical conditions of a non-noise environment, and the accuracy turns out to be low in the view of an advanced network, specifically, around 70%. Thus, this creative approach is instructive, but still not fully solving the mask detection problem. Some latest research is much more revealing in realizing real-time mask detection. Scholars led by Preeti Nagrath applied single shot multibox detector and MobileNetV2 [4] (as shown in Figure 2) to establish a real-time DNN-based mask detection system.

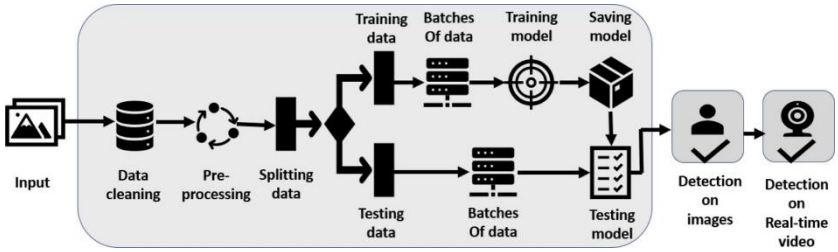

Figure 2. Flow Diagram of the SSDMNV2 model.

DOI: 10.54254/2755-2721/2/20220527

This model achieved high accuracy and high detection speed enabling the mask detection to remain real time. Still, the improper mask wearing is not separated from the mask detection, making the model less practical in real world application. Additionally, the scholars manually selected and filtered only the simplest image for training to achieve expected accuracy, which on the other hand further undermined the practical value of the model.

In conclusion, all the models above successfully judge the existence of a mask on the face and their results turned out to be satisfying, though sometimes low accuracy occurs. However, the paper points out that all these works leave the question of whether the people being detected wear the masks properly. Therefore, It aims to figure out the conditions where the masks are not properly used and maintain high accuracy simultaneously, which is definitely innovative among the related works.

3. Methodology

To make the model more interpretable, the mask discrimination task is divided into two independent and dependent sub-modules. In the first module, a face detection model is designed to recognize faces in the screen. After that, feature screening, feature extraction and feature fusion were carried out on the facial area of the face, in which the features of the mask were also integrated into the facial features, which could accelerate the training efficiency and frame rate detection of the model. Finally, it uses a simple classification module to identify and classify the features of face and mask and combine it with the feature extraction module to form the second module of the whole system.

Firstly, the input image is divided into grid cells, and the grid at the center of the target object is responsible for predicting the position of bBoxes (candidate boxes), corresponding confidence and class of the object. Bbox confidence includes two parts: the accuracy of candidate box and the confidence of existing target.

$$fi = Bool[object] \times iou_{pred}^{truth}) \tag{1}$$

It can found that if there is a target in the candidate box, confidence fi = IOU truth. If not, fi = 0. At this point, the model will return five values of Bbox: x,t, W, H and fi, where (x,y) represents^{n}the ratio of the center coordinate of the candidate frame to the length and width of the candidate frame, (w,h) represents the ratio of the length and width of the candidate frame to the whole image, and fi represents bbox confidence.

If there is a total of class C targets, each grid needs to predict which category the target with this grid belongs to. The conditional probability of prediction is indicating that the probability of each category should be predicted if the grid has targets. In the test, the confidence of the candidate box to a certain category is the product of the probability of the category and the confidence of the candidate box. In this paper, YOLOv5 is used for face detection. At the same time, the detected face is clipped, saved and sent into the next module of the system.

At the same time, the paper not only uses YOLO [5] to detect faces, it also has selected a pretrained Mobile Net to reduce computational cost [6].

Transfer learning was one of the widespread techniques used in computer vision missions such as classification and segmentation. It shares weights or and information learned while solving one problem to solve other similar problems. In general, transfer learning reduces training time if the areas of the tasks are closely related. We used one of the pre-trained models to complete the task (MobileNet). A pre-trained model is the model that has been trained on large-scale datasets. There are a number of pre-trained models such as ResNet, MobileNet, GoogleNet. These models accept RGB images as input and is capable of achieve classification for more than 1000 classes [7].

Depthwise separable convolution is the fundamental unit of the MobileNet model. It uses the factorization of filters to reduce computational cost and size. It consists of two layers: depthwise convolution and pointwise convolution. The depthwise convolution applies a single filter to each input channel and realizes layer filtering; the pointwise convolution creates a liner combination of the output with the 1 Œ 1 convolution.

DOI: 10.54254/2755-2721/2/20220527

In the next module, the extended network based on Resnet (Deep Residual Learning for Image Recognition) [8] is used to extract facial appearance features.

Because simply increasing the depth of neural network will cause problems such as gradient explosion and gradient disappearance, as well as degradation of network model performance, Resnet proposed in 2016CVPR can effectively solve the above problems. The main structure of Resnet is mainly based on the block unit, and the input and output of each block are connected through a shortcut, which is equivalent to a simple equivalent mapping. In this way, no additional parameters are added, that is, no extra calculation is required, and the performance of the network after deepening is not degraded.

In the actual training process, the output of Resnet average Polling Layer will be puted into the full Connected layer and ground truth to calculate CrossEntropyLoss. Loss is backpropagated to the whole network to update the weight. After the training, the network will drop the last classified layer and divide the output of the average polling layer into two norms to output it as the 512-dimensional feature vector of the face.

In order to make the network more robust, the work use the same face of the traditional random data enhancement, including the color, saturation, brightness of the random transformation, as well as random scaling, horizontal translation and rotation, resulting in optical distortion and geometric distortion of the image. However, different from other situations, in the face data set, especially in the case of wearing masks, in order to enable the model to pay more attention to and learn the external characteristics of the wearing area of masks, Therefore, during data enhancement, we deliberately discarded the background Region of the image through region-mask (random mask on upper Region). A comparison of the results of specific experiments will be shown in the next section.

Because Resnet is a very popular and widely used backbone network, Densenet [9] proposed in 2017CVPR also draws on part of Resnet's ideas. However, compared with Resnet, Densenet proposed a more radical dense connection mechanism: To be specific, each block will accept the output of all previous blocks as its additional input, so as to better realize feature reuse and improve efficiency. This is also the main difference between the two networks. So Densenet is also used to improve the previous model, so that it can obtain high accuracy.Later, Densenet is trained in different depths, and parameters are selected and adjusted among different optimizers, so that the model could be more suitable for face feature extraction and mask recognition. The specific results will be shown i n the next section. Although Densenet and respectively use log _softmax and log_softmax at the end of the network, both networks use full Connected layer at the end of the feature fusion and reduce the dimension of features, so the two models can only accept fixed image sizes. This is exactly what the paper did in the last step of the data enhancement module. (figure 3 shows a diagram of data argumentation methods used in the work)

Figure 3. Region mask and other data argumentation methods Resnet.

In order to solve the problem of loss of detailed information when a person's face is scaled or scaled, when one dimension is scaled to a fixed size, padding is used to make the image reach the input size expected by the neural network. In the actual test results, this method can indeed improve part of the accuracy. Then we optimize the model again and add SPPnet (Spatial Pyramid Pool) [10] before the fully connected layer. Regardless of the size of the feature map, SPPnet has a three-layer pyramid pooling layer, that

DOI: 10.54254/2755-2721/2/20220527

is, three different scales (4x4, 2X2, 1X1) are divided into the same large grid (16, 4, and 1). Each network is a feature point, and (16, 4, 1) features are obtained through the Max polling layer. Finally, the output dimension of the feature map is (16+4+1) Œthe number of channels, so that regardless of the size of the input image, a fixed size output can be obtained. (The number of channels in this example is 256). After modifying the Dataloader, the entire model can adapt to images of different sizes without losing face information due to image scaling. The detailed before and after comparison results will be shown later. Finally, some simple scenarios are made and masked data sets are designed to train shallow networks and test models trained on other public datasets.

4. Experimental results

In this paper, the medical mask data set is used to train the face detection model. The training set in the medical mask data set contains 3694 images, the test set contains 299 images, a total of 3993 images, and a total of 20 categories. Since this article requires the face detection model to detect the face regardless of the mask wearing condition, the data set is only used including: face-with-mask, face-other-covering, face-no-mask, face-with-incorrect mask.

Figure 4 shows Four category as the label of the face.

Figure 4. Medical mask dataset.

The machine configuration used in this article is shown in the following table 1:

Table 1. Experimental environment configuration.

Operating system	**CPU**	**RAM**
Windows 10	I7-7700HQ @ 2.80GHz*2	24.0GB
GPU	**Language**	**Algorithm framework**
1060 6G * 1	Python 3.7	Pytorch 1.8.1

In the detection stage, YOLO is used to perform face detection respectively to adapt to the requirements of different detection rates and different detection accuracy (as shown in Figure 5).

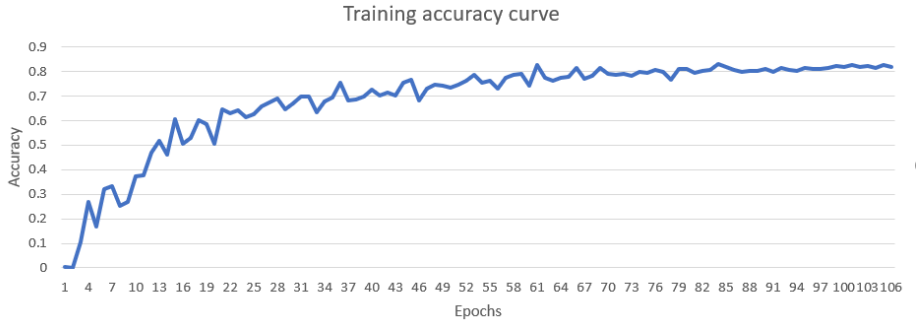

Figure 5. YOLO Training accuracy curve.

Table 2 compares the YOLO networks and Mobilenets:

DOI: 10.54254/2755-2721/2/20220527

Table 2. Comparison between YOLO and mobilenet.

	YOLO	Mobilenet
Number of parameters	6,861,110	2,257,984
mAp@.5 on the medical mask data set	0.811	0.794
Recall on the medical mask data set	0.775	0.833
Accuracy on self-made datasets in real scenes	93.01%	67.41%
Detection rate	53FPS (Test on RTX1060)	120FPS (Test on RTX3070)

A mask dataset was made, and the images were divided into three categories. Among them, there are 2403 images without masks, 1785 images with appropriately mask wearing, and 3231 images with appropriately mask wearing. Then the trained face detection model is tested on a self-made dataset. The figure 6 is the detection result of the face detection model:

Figure 6. Detection results.

After that, the paper uses Res34 to classify the detected faces based on the features of both the face and the mask. The training set used is the Masked Face Dataset (MFD) training set. The MFD includes a total of 6210 images, and the test set includes a total of 587 images, including two categories of inappropriately wear mask' and appropriately' wear mask'.

Figure 7 shows diagrams of Masked Face Dataset

Figure 7. Masked face dataset.

Resnet 34 and Densenet are used to extract and classify face and mask features. The training results are as follows(Figure 8):

DOI: 10.54254/2755-2721/2/20220527

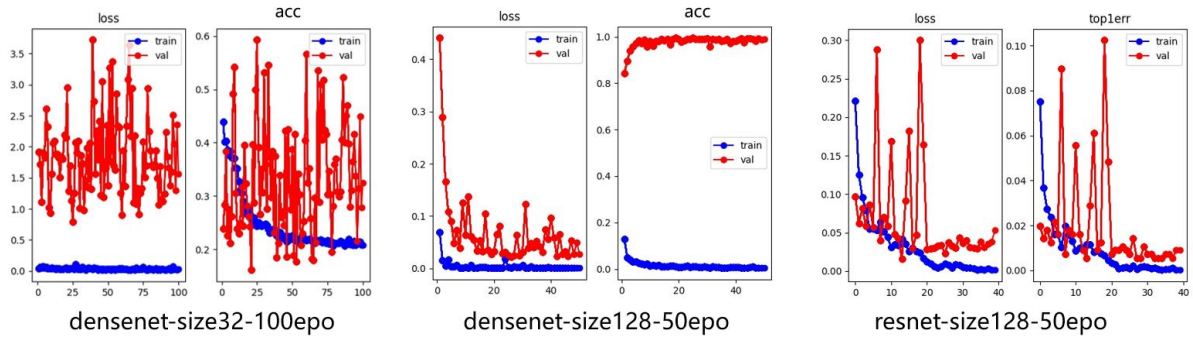

Figure 8. Training curves of Resnet and Densenet.

It can be found that on the same input image size, Densenet's top1error is smaller than Resnet of the same input image size. At the same time, if Densenet is also used, a larger input image will make the model learn faster and effectively prevent the model from shaking during training.

Table 3 compares the accuracy between the Resnet and the Densenet

Table 3. Accurancy between the Resnet and the Densenet.

Net	input 32*32	input 128*128	50 epochs	100 epochs	Final accuracy
ResNet	√			√	66.26%
ResNet		√	√		97.36%
DenseNet	√			√	74.39%
DenseNet		√	√		99.10%

To enable the network to accept images of different scales, the spatial pyramid pooling layer SPPnet is added to the model, so that the model can be multi-scale training or multi-scale pooling, thereby improving the overall accuracy.

The figure 9 shows the accuracy changes of densenet training for 50 and 15 epochs before and after SPPnet.:

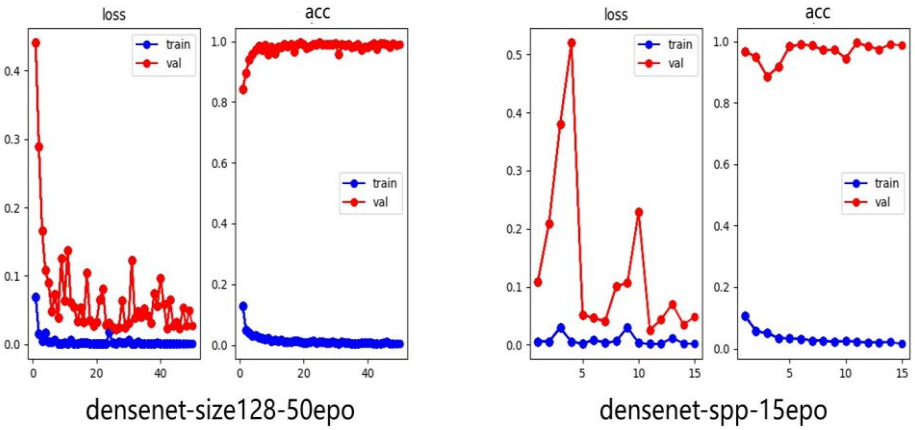

Figure 9. Densenet with and without SPPnet.

Although the Top1error of both can reach 0.02, the training curve of the model before joining SPPnet is more unstable than that after joining SPPnet. In other words, joining SPPnet can make the model converge faster. Although the addition of SPPnet does allow the network to adapt to images of different sizes, and at the same time to obtain and learn more information from larger-size images, the addition of SPPnet makes Densenet's GPU memory usage increase exponentially, which leads us to use a smaller batchsize, which also greatly increases the time to train the model in disguise. In the previous article, The paper talked about

DOI: 10.54254/2755-2721/2/20220527

the specific implementation of data enhancement including regionmask. Figure 10&11will compare the impact of adding region mask and related data enhancement methods on model accuracy:

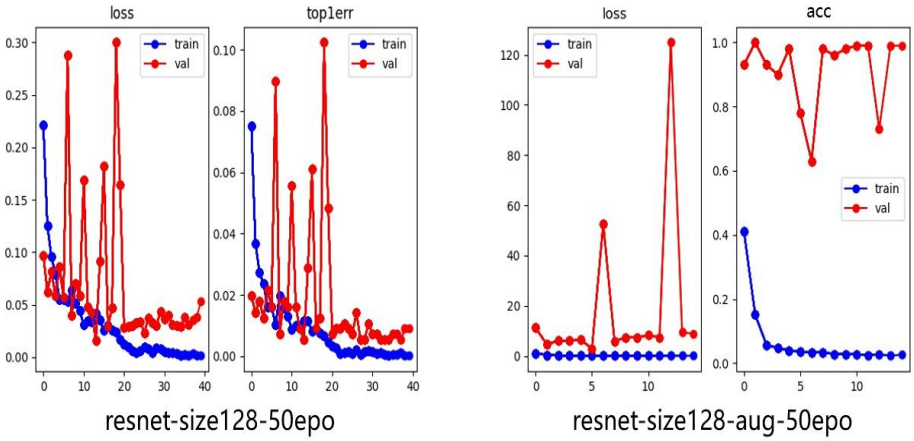

Figure 10. Resnet with and without argumentation.

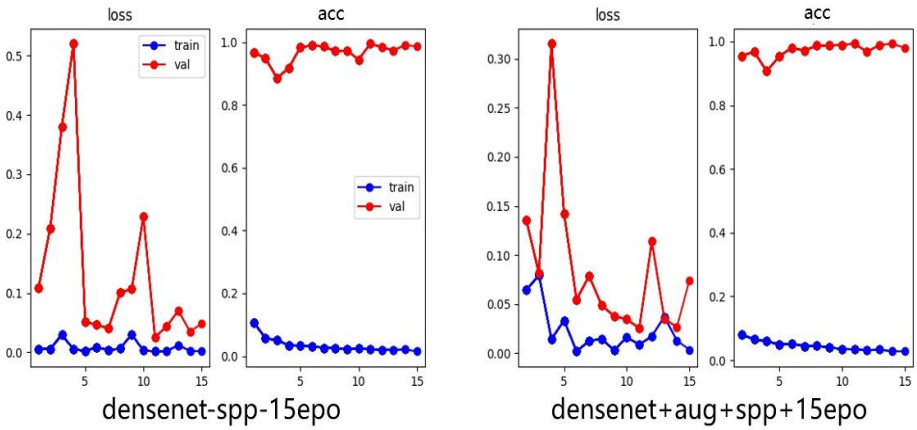

Figure 11. Densenet with and without argumentation.

It can be seen from the training results that the use of region mask and other argumentation methods makes the training of the model smoother, and it can also reduce the time required for the model to fully converge. (As shown in Table 4)

Table 4. Result of the models.

Net	input 128*128	SPPnet	15 epochs	50 epochs	argumentation	Top1 error	epochs at convergence
ResNet	√			√		2.64%	23
ResNet	√			√	√	1.1%	3
DenseNet		√	√			0.68%	5
DenseNet		√	√		√	0.90%	4

Finally, we need to talk about a method that have been tried but did not achieve a very good result in the end. Haar cascade is an Object Detection Algorithm to detect possible faces in an image. The algorithm uses edge or line detection features proposed by Viola and Jones in 2001 [11]. The algorithm is given a lot of positive images consisting of faces, and a lot of negative images not consisting of any face to train on them.

DOI: 10.54254/2755-2721/2/20220527

In the process of classifying whether the person is properly wearing a mask, the paper intended to use the Haar cascade algorithm to detect the position of the nose and mouth of the person. If either the mouth or the nose is detected, then the person is not properly wearing a mask or not wearing a mask at all, it can further be distinguish the two cases by training a haar cascade model focusing on detecting the mask.

After training Haar cascade models, the drawback of Haar cascades is that it tends to be amend to false-positive detections, and the tuning on parameters is needed when being applied for detection. In general, the Haar cascade algorithm is not as accurate as the detector this work has developed based on YOLO when there are big variations in the resolution of the image set.

5. Conclusion

The main task is to perform detection for general scenarios of mask wearing. This paper use several different methods for face detection and It also compares these different face detection methods, which allows the user to make a trade-off between accuracy and detection speed by choosing different networks. After achieving good accuracy in the face detection phase, face images obtained from the complex scene is used to train the feature extraction and classification network.In this step ,a simple shallow network, Res34 has been trained, and the end of the network is modified so that it could extract and output the features of faces and masks. Then it can be found that Densenet can improve the accuracy of mask wear recognition without adding too many network parameters, so we replaced Resnet with Densenet network and compared the results of both networks in detail. This work also used region mask and numerous other random data enhancement methods to make the model more robust. It can then be found that roughly scaling the image size would lose much of the hidden relative position information that was already present, and the data size of the input network would affect the accuracy of the network in mask wearing situation recognition, but due to the existence of the fully connected layer, the image of the input network should be fixed, so SPPnet is added to the network, which allows the network to be trained at multiple scales and allows the network to accept. In this step, a simple shallow network res34 is first trained, and the end of the network is modified so that it can extract and output the features of face and mask. Then we found that densinet can improve the accuracy of mask wear recognition without adding too many network parameters, so densinet is used to replace RESNET. Moreover, the results of the two networks are compared in detail. This paper also uses region mask and many other random data enhancement methods to make the model more robust. Then it can be found that roughly scaling the image size will lose a lot of existing hidden relative position information. The data size of the input network will affect the accuracy of the network in mask wearing recognition. However, due to the fully connected layer, the image input to the network must be fixed, so sppnet will be added to the network, The network is allowed to train on multiple scales and accept the input of images of any size. Finally, this article creates its own mask wear recognition data set for the project, which does not contain a large number of images. However, it can be used to fine tune the network after learning to make it perform better in our expected operating environment. It can also be used to measure the accuracy of face detection network in real scenes.

References

[1] Alex P，M Yan，Y T Huang，C Gao，W Z Li，（2021）What shapes people’s willingness to wear a face mask at the beginning of a public health disaster? A qualitative study based on COVID-19 in China. [J]International Journal of Disaster Risk ReductionVolume 65, 2021. PP 102577-102577

[2] Chowdary G J, Punn N S, Sonbhadra S K, et al. （2020）Face mask detection using transfer learning of inceptionv3[C]//International Conference on Big Data Analytics. Springer, Cham, 2020: 81-90.

[3] Ristea N C, Ionescu R T. （2006）Are you wearing a mask? Improving mask detection from speech using augmentation by cycle consistent GANs[J]. arXiv preprint arXiv:2006.10147, 2020

[4] Nagrath P, Jain R, Madan A, et al. (2021) SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2[J]. Sustainable cities

DOI: 10.54254/2755-2721/2/20220527

and society, 2021, 66: 102692.

[5] Redmon J, Divvala S, Girshick R, et al. (2017) You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[J]. 2017:6517- 6525.

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[7] Redmon J, Farhadi A. (2018) YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.

[8] He K , Zhang X , Ren S , et al.(2016) Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[9] Huang G, Liu Z, Laurens V, et al. (2016) Densely Connected Convolutional Networks[J]. IEEE Computer Society, 2016.

[10] He K, Zhang X, Ren S, et al. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2014, 37(9):1904-16.

[11] Viola, P., Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.