Multi-Crowd Mask Detection Method Based On SSD

Written by XMU Students[1]
Qian Hui Yang, Jing Chen, Jie Lian, Yu Xing Dai
[1]Electronic Information Major, Department of Software Engineering, School of Information, Xiamen University

Abstract

Coronavirus disease 2019 has affected the world seriously. One major protection method for people is to wear masks in public areas. Wearing face masks and following safe social distancing are two of the enhanced safety protocols need to be followed in public places in order to prevent the spread of the virus. Furthermore, many public service providers require customers to use the service only if they wear masks correctly.

Therefore, we proposes a multi crowd mask wearing detection method based on SSD (single shot multibox detector), which can automatically detect masks in real-time video frames or static images. After 200 rounds of training, the accuracy rate of the model in the test set reaches 0.919, which can accurately detect whether people in the crowd do not wear masks. As one of the auxiliary means of epidemic prevention and control, the multi-crowd mask detection method has certain application value. It can reduce the workload of relevant staff, supervise the public wearing masks and ensure that people do a good job of health protection when they go out.

Keywords: Coronavirus, Safe Social Distancing, Multi Crowd Mask Detection, SSD

Introduction

The situation report of world health organization (WHO) presented that coronavirus disease 2019 (COVID-19) has globally infected a large number of people and caused a lot of deaths. In addition, there are several similar large scale serious respiratory diseases, such as severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS), which occurred in the past few years (Rota et al. 2003; Memish et al. 2013). Liu et al. (Liu et al. 2020b) reported that the reproductive number of COVID-19 is higher compared to the SARS. Therefore, more and more people are concerned about their health, and public health is considered as the top priority for governments (Fang, Nie, and Penny 2020). Fortunately, Leung et al. (Leung et al. 2020) showed that the surgical face masks could cut the spread of coronavirus. At the moment, WHO recommends that people should wear face masks if they have respiratory symptoms, or they are taking care of the people with symptoms (Feng et al. 2020). Furthermore, many public service providers require customers to use the service only if they wear masks. Therefore, face mask detection has become a crucial computer vision task to help the global society, but research related to face mask detection is limited.

Figure 1: Examples of Images in the Face Mask Dataset

There are examples of images in the face mask dataset which is shown in Fig. 1. Face mask detection refers to detect whether a person wearing a mask or not and what is the location of the face (Wang et al. 2020). The problem is closely related to general object detection to detect the classes of objects and face detection is to detect a particular class of objects, i.e. face (Zhao et al. 2019). Applications of object and face detection can be found in many areas, such as autonomous driving (Lee et al. 2020), education (Savita et al. 2018), surveillance and so on. Traditional object detectors are usually based on handcrafted feature extractors. Viola Jones detector uses Haar feature with integral image method (Viola and Jones 2001), while other works adopt different feature extractors, such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT) and so on (Felzenszwalb, McAllester, and Ramanan 2008).

Recently, deep learning based object detectors demonstrated excellent performance and dominate the development of modern object detectors. Without using prior knowledge for forming feature extractors, deep learning allows neural networks to learn features with an end-to-end manner (Liu et al. 2020a). There are one-stage and two-stage deep learning based object detectors. One-stage detectors use a single neural network to detect objects, such as single shot detector (SSD) (Liu et al. 2016) and you only look once (YOLO) (Redmon et al. 2016). In contrast, two-stage detec-

tors utilize two networks to perform a coarse-to-fine detection, such as region-based convolutional neural network (R-CNN) (Girshick et al. 2014) and faster R-CNN (Ren et al. 2015). Similarly, face detection adopts similar architecture as general object detector, but adds more face related features, such as facial landmarks in RetinaFace (Deng et al. 2019), to improve face detection accuracy. However, there is rare research focusing on face mask detection.

Related Work

In recent years, object detection techniques using deep models are potentially more capable than shallow models in handling complex tasks and they have achieved spectacular progress in computer vision. Deep models for person detection focus on feature learning contextual information learning, and occlusion handling. Deep learning object detection models (Farfade, Saberian, and Li 2015) can now mainly be divided into two families: (i) two-stage detectors such as R-CNN, Fast R-CNN and Faster R-CNN and their variants and (ii) one-stage detectors such as YOLO and SSD. In two-stage detectors detection is performed in stages, in the first stage, computed proposals and classified in the second stage into object categories. However, some methods, such as YOLO, SSD MultiBox, consider detection as a regression issue and look at the image once for detection.

The Fig. 2 below shows the SSD system and it is Single Shot Detector MultiBox(SSD) which seems to be a good choice for real-time object detection and the accuracy trade-off is also very little. SSD uses the VGG-16 model pre-trained on ImageNet as its basic model to extract useful image feature. At the top of VGG16, SSD adds several convolutional feature layers of decreasing sizes.

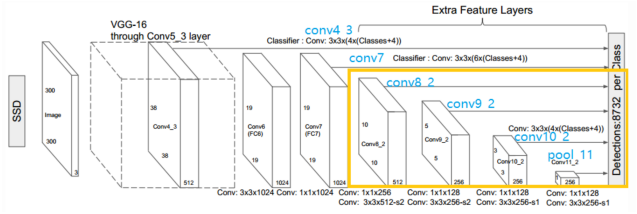

Figure 2: Model architecture of SSD

The Viola Jones object detection system can be trained to detect any object, but is especially common for facial detection and is more accurate and faster. The Viola and Jones process is an example of supervised learning. Zhu(Zhu and Ramanan 2012) also shared another very widespread facial detection algorithm is a neural network-based detector.

It only works well with the front, upright face. Ren et al.(Ren and Liu 2020) propose an improved Face-mask Net detection method for convolutional neural network based on YOLOv3. It can effectively detect the target of wearing mask and not wearing mask, and its performance can be close to real-time and the accuracy is higher than that of the network before the improvement. Nicolae-Ctlin Ristea's approach(Ristea and Ionescu 2020) is based on (i) training Generative Adversarial Networks (GANs) with cycle-consistency loss to translate unpaired utterances between two classes (with mask and without mask).Park et al. (Park et al. 2020) combine the object detecting network based on MobileNetV2 plus SSD and the quantization scheme for integer-only hardware, which can also be deployed on Coral Dev Board, a commercially available development board containing Google Edge TPU.Li et al. (Howard et al. 2017), suggested another model for facial detection which was a MultiView Face Detector with surf capabilities. Oro et al. (Oro et al. 2011) also proposed a haar-like feature based face detection algorithm for HD video on the GTX470 and obtained an improved speed of 2.5 times. However, they only used CUDA which is a GPU programming tool for NVIDIA GPUs.

Compared to OpenCL which is used in a number of computed components, it is unable to resolve the imbalanced workload issue experienced during the implementation of the viola-jones face detection algorithm in GPUs. Glass et al. (2006)(Glass et al. 2006) addressed the importance of social differencing and how the risk of pandemic growth can be slowly decreased by successfully preserving social distance without the use of vaccines or antiviral drugs.

The authors have carried out an exhaustive study on this in both rural and urban communities in order to demonstrate a reduction in the growth rate. Z., Luo(Ge et al. 2017) studies the identification of people with full-face or partial occlusion. This approach categorizes into way, people with hand over their faces or occluded with objects.

These related work can bring us enlightenment and ideas for our future work to complete our work.

Methodology

Our project is able to detect whether or not people wears face mask in the crowd. This section briefly describes the model architecture and how the proposed system will automatically functions in an automatic manner to prevent the coronavirus spread.

We propose a multi-crowd mask wearing detection method based on SSD (Single Shot MultiBox Detector). This method can automatically detect face masks in real-time video frames or static images. Among many target detection algorithms, SSD has a clear speed advantage compared to Faster RCNN, and has a clear mAP advantage model structure compared to YOLO(See Fig. 3).

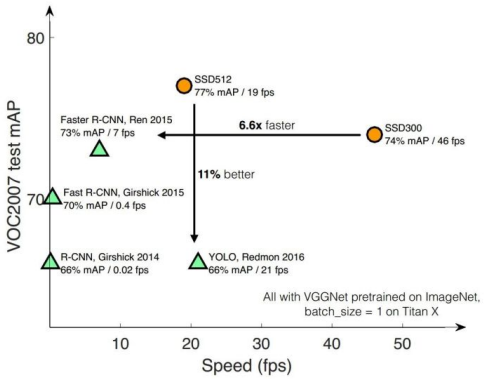

Figure 3: mAP of main object detection models

SSD has the following three characteristics, which make it achieve very good results in object detection tasks.

- Using the idea of converting the detection task into a regression task like YOLO, the target positioning and classification are completed at one time. Compared with the existing methods, both the detection speed and detection accuracy are better.
- Based on the anchor in Faster RCNN, a similar Prior box is proposed(See Fig. 4), that is, some target pre-selection boxes, and then the real target position is obtained through classification and bounding box regression.
- Add detection method based on Pyramidal Feature Hierarchy which can predict targets on feature maps of different receptive fields.

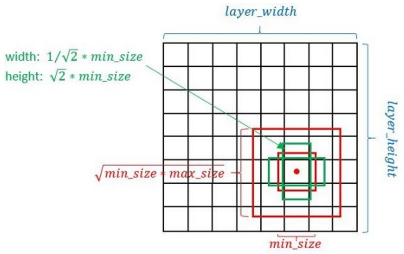

Figure 4: Prior box

However, SSD also has certain shortcomings, such as the need to manually set the min size, max size and aspect ratio values of the prior box, and the basic size and shape of the prior box in the network cannot be obtained directly through learning.

To train a mask detection model based on SSD, the most important thing is to set the size and aspect ratio of the anchor reasonably. So in this experiment, we analyzed the mask face data set used and read the annotation information of all faces. And then calculate the ratio of the height to the width of each face, and obtain the distribution information of the height to the width ratio by statistics. The analysis result shows that the normalized face aspect ratio is concentrated between 1 and 2.5. Therefore, according to the data, we set the aspect ratio of the anchors of the five positioning layers to 1, 0.62, 0.42.

The input image of the mask detection model we used is 260x260, and there are only 28 3x3 convolutions, and the number of channels in each convolutional layer is 32, 64, and 128. After 8 consecutive convolutional pooling layers for feature extraction, it enters the subsequent five positioning and classification layers .The result of the combination of the last five classification layers is used as the output.

We used NMS in our post-processing part to speed up and improve efficiency. NMS's full name is Non-maximum suppression, as the name implies, is to suppress elements that are not maximum values, which can be understood as local maximum search. This part represents a neighborhood, and the neighborhood has two variable parameters, one is the dimension of the neighborhood, and the other is the size of the neighborhood. The general NMS algorithm is not discussed here, but it is used to extract the window with the highest score in target detection. For example, in pedestrian detection, the sliding window extracts features, and after classification and recognition by the classifier, each window will get a score. However, sliding windows will cause many windows to contain or mostly overlap with other windows. At this time, you need to use NMS to select those neighborhoods with the highest scores (the highest probability of pedestrians), and suppress those windows with low scores. NMS has very important applications in the field of computer vision, such as video target tracking, data mining, 3D reconstruction, target recognition, and texture analysis.

In the training process, we use labeled mask data for binary classification model training. We use the preset size box to predict whether the input data image is wearing a mask. The binary classification task is relatively easy to recognize a single object, but our method needs to be able to adapt to multi-target scenarios. In actual use, multi-target recognition is necessary, and the method we propose can fulfill this requirement well (see the experimental part).

Experiment

A. Dataset

The proposed method uses a custom data set consisting of face images with different types of face masks which are labeled and used for the training of our model.We use the existing background subtraction(Mohamed, Tahir, and Adnan 2010)(Piccardi 2004) algorithm in a pre-processing step. Face Mask Dataset contains 7959 images, and its faces are annotated with either with a mask or without a mask. However, Face Mask Dataset a combined dataset made up of Wider Face and MAsked FAces (MAFA) dataset. Wider Face contains 32,203 images with 393,703 normal faces with various illumination, pose, occlusion etc. MAFA contains 30,811 images and 34,806 masked faces. We train our model using the 80% images of the dataset and the remaining 20% for testing. Some samples are shown in Fig. 1, including faces with masks, faces without masks, faces with and without masks in one image.

B. Experiment Setup

In the experiments, we employed stochastic gradient descent (SGD) with learning rate $\alpha = 10^{-3}$, momentum $\beta = 0.9$ and we have trained about 200 epochs as optimization algorithm. We trained the model on a NVIDIA GeForce RTX 2080 Ti. The algorithm is developed with PyTorch deep learning framework, each experiment operates 200 epochs.

The post-processing part is mainly non-maximum suppression (NMS). We use a single-type NMS, that is, two types of faces with masks and faces without masks are used as NMS together to improve the speed. To train the target detection model, the most important thing is to set the size and aspect ratio of the anchor reasonably.

In the experiment, the aspect ratio and size of the target object in the data set are generally calculated to set the size and aspect ratio of the anchor. For example, on the mask face dataset we annotated, we read the annotation information of all faces, calculate the ratio of height to width of each face, and obtain the histogram of the distribution of height to width ratio in Fig. 5:

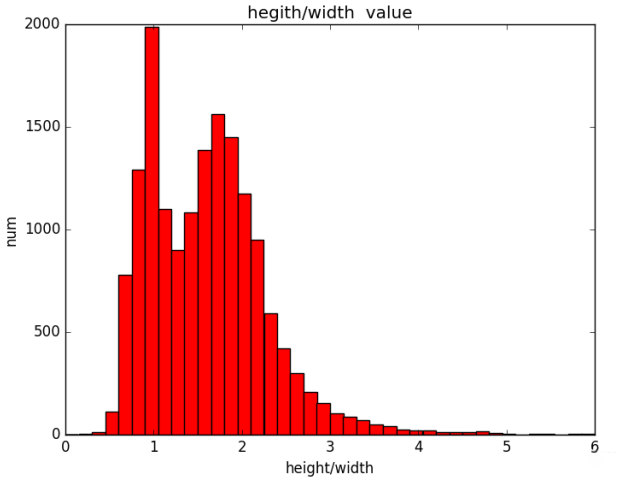

Figure 5: Distribution Histogram of Height and Width Ratio

Because human faces are generally rectangular, and many pictures are relatively wide, such as 16:9 pictures, after the width and height of the face are normalized, the height of many pictures is twice or greater than the width. It can also be seen from the figure above that the normalized face aspect ratio is concentrated between 1 and 2.5. Therefore, according to the data distribution, we uniformly set the aspect ratio of the anchors of the five positioning layers to 1, 0.62, and 0.42. (Converted to aspect ratio, which is about 1, 1.6:1, 2.4:1)

C. Evaluation Metrics

We employed precision and recall as metrics, they are defined as follows:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where TP, FP and FN denoted the true positive, false positive and false negative, respectively.

D. Result and Analysis

The Precision-Recall curve of the final model on the test set is shown in the Fig. 6:

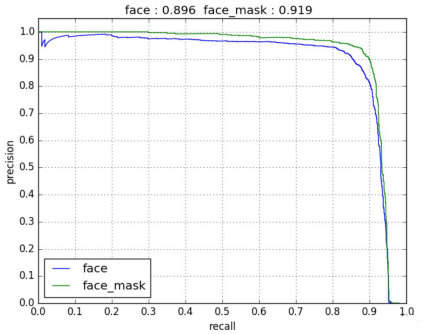

Figure 6: Precision-Recall Curve

Some experiment results are shown in Fig. 7, where the blue and green boxes refer to the face and mask predictions, respectively.

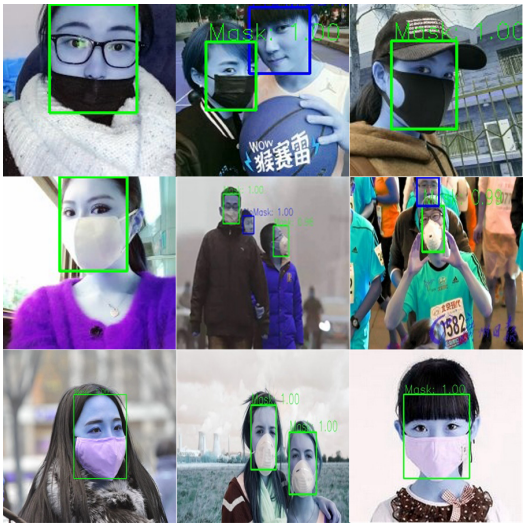

Figure 7: Examples of Detection Results

Conclusion

In this paper, we have proposed a multi-crowd face mask detection method based on SSD(single shot multibox detector), which not only can automatically detect masks in real-time video frames or static images, but also can can possibly contribute to public healthcare. After 200 rounds of training, the accuracy rate of the model in the test set reaches 0.919, which can accurately detect whether people in the crowd do not wear masks. And this is the first time that we have studied this type of topic. There will be not a few shortcomings.

There is no doubt that we can make corresponding improvements in our future work.

Future Work

The method used in this paper can accurately detect the wearing of masks in multiple populations, which is helpful to the prevention and control of epidemic situation. However, due to the limitation of time and equipment, there are still some problems, waiting for further research.

1. False Mask False Detection: In the detection process, occasionally, when people cover their mouth and nose with hands or other occlusions, the model will be misjudged as wearing masks. In order to solve this problem, in the follow-up work, some occlusion objects are made on the face data set, labeled with no mask, and added to the training set to enhance the sensitivity of the model to the "false mask", and improve the accuracy and reliability of mask wearing detection.
2. Multi Angle Missing Inspection: At present, the data set of multi population masks is limited, and there is a situation of missing detection of side face masks. In order to solve this problem, we need to further collect or make more data sets of multi angle face masks, and use the common data expansion methods such as flip, rotation, scaling, clipping to expand our data set, so as to achieve better detection results.
3. Cough And Sneeze Detection: For assisting epidemic prevention and control work, multi population mask wearing detection is only one of the measures, which can be further improved and expanded on the basis of this model, such as cough and sneezing detection. According to the guidelines of the World Health Organization, chronic cough and sneezing are one of the main symptoms of covid-19 infection and one of the main ways of disease transmission.

References

Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; and Zafeiriou, S. 2019. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641* .

Fang, Y.; Nie, Y.; and Penny, M. 2020. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. *Journal of medical virology* 92(6): 645–659.

Farfade, S. S.; Saberian, M. J.; and Li, L.-J. 2015. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 643–650.

Felzenszwalb, P.; McAllester, D.; and Ramanan, D. 2008. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Feng, S.; Shen, C.; Xia, N.; Song, W.; Fan, M.; and Cowling, B. J. 2020. Rational use of face masks in the COVID-19 pandemic. *The Lancet Respiratory Medicine* 8(5): 434–436.

Ge, S.; Li, J.; Ye, Q.; and Luo, Z. 2017. Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2682–2690.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Glass, R. J.; Glass, L. M.; Beyeler, W. E.; and Min, H. J. 2006. Targeted social distancing designs for pandemic influenza. *Emerging infectious diseases* 12(11): 1671.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* .

Lee, D.-H.; Chen, K.-L.; Liou, K.-H.; Liu, C.-L.; and Liu, J.-L. 2020. Deep learning and control algorithms of direct perception for autonomous driving. *Applied Intelligence* 1–11.

Leung, N. H.; Chu, D. K.; Shiu, E. Y.; Chan, K.-H.; McDevitt, J. J.; Hau, B. J.; Yen, H.-L.; Li, Y.; Ip, D. K.; Peiris, J. M.; et al. 2020. Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nature medicine* 26(5): 676–680.

Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; and Pietikäinen, M. 2020a. Deep learning for generic object detection: A survey. *International journal of computer vision* 128(2): 261–318.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Y.; Gayle, A. A.; Wilder-Smith, A.; and Rocklöv, J. 2020b. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* .

Memish, Z. A.; Zumla, A. I.; Al-Hakeem, R. F.; Al-Rabeeah, A. A.; and Stephens, G. M. 2013. Family cluster of Middle East respiratory syndrome coronavirus infections. *New England Journal of Medicine* 368(26): 2487–2494.

Mohamed, S. S.; Tahir, N. M.; and Adnan, R. 2010. Background modelling and background subtraction performance for object detection .

Oro, D.; Fernández, C.; Saeta, J. R.; Martorell, X.; and Hernando, J. 2011. Real-time GPU-based face detection in HD video sequences. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 530–537. IEEE.

Park, K.; Jang, W.; Lee, W.; Nam, K.; Seong, K.; Chai, K.; and Li, W.-S. 2020. Real-time Mask Detection on Google Edge TPU .

Piccardi, M. 2004. Background subtraction techniques: a review .

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ren, X.; and Liu, X. 2020. Mask wearing detection based on YOLOv3. *Journal of Physics: Conference Series* 1678: 012089. doi:10.1088/1742-6596/1678/1/012089. URL https://doi.org/10.1088%2F1742-6596%2F1678%2F1%2F012089.

Ristea, N.-C.; and Ionescu, R. T. 2020. Are you wearing a mask? Improving mask detection from speech using augmentation by cycle-consistent GANs .

Rota, P. A.; Oberste, M. S.; Monroe, S. S.; Nix, W. A.; Campagnoli, R.; Icenogle, J. P.; Penaranda, S.; Bankamp, B.; Maher, K.; Chen, M.-h.; et al. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *science* 300(5624): 1394–1399.

Savita, K.; Hasbullah, N. A.; Taib, S. M.; Abidin, A. I. Z.; and Muniandy, M. 2018. How's the Turnout to the Class? A Face Detection System for Universities. In *2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, 179–184. IEEE.

Viola, P.; and Jones, M. 2001. 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*, volume 1, I–I. IEEE.

Wang, Z.; Wang, G.; Huang, B.; Xiong, Z.; Hong, Q.; Wu, H.; Yi, P.; Jiang, K.; Wang, N.; Pei, Y.; et al. 2020. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093* .

Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; and Wu, X. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems* 30(11): 3212–3232.

Zhu, X.; and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, 2879–2886. IEEE.