A review of the prevalent ICT techniques used for COVID-19 SOP violation detection

Talha Ikram
National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science, Islamabad, Pakistan
tikram.bese17seecs@seecs.edu.pk

Abdullah Saeed
National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science, Islamabad, Pakistan
asaeed.bese17seecs@seecs.edu.pk

Noor Ul Ayn
National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science, Islamabad, Pakistan
nayn.bese17seecs@seecs.edu.pk

Muhammad Ali Tahir
National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science, Islamabad, Pakistan
ali.tahir@seecs.edu.pk

Rafia Mumtaz
National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science, Islamabad, Pakistan
rafia.mumtaz@seecs.edu.pk

2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET) | 978-0-7381-0527-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/HONET50430.2020.9322821

***Abstract*—COVID-19 is a disease that has adversely impacted the health and daily lives of people worldwide, therefore there must be measures in place to control the spread of such diseases. Standard Operating Procedures (SOPs) such as wearing masks and maintaining social distancing are enforced by the Government and healthcare authorities. These SOP's mitigate the spread of COVID-19, but it has been observed with concern that people do not generally follow them. This survey work explores classical techniques that can be used to detect SOP violations. These are primarily computer vision based methods used in object detection and distance estimation. We will also explore deep learning based techniques used for object detection to detect SOP violations.**

Keywords—COVID-19, SOPs, Deep learning, Distance estimation

I. Introduction

The COVID-19 virus has been spreading rapidly around the world since December 2019. After 10 months there still seems to be little clear evidence on how to stop the spread of COVID-19. This is in part due to the limited funding that researchers get and in part due to the limited resources available to authorities to ensure that people follow SOPs. Vaccines are now starting to become available in first world countries but it will take a while before they become common in developing countries.

SOPs are the basic safeguard against the spread of this disease but people widely ignore them in both developing and developed countries. Furthermore, if we look at history, such diseases at least occur once a century on a large scale; yet there are still few tools available to ensure that they do not spread.

Since the spread of COVID-19, there has been a lot of research on rapid diagnosis of COVID-19. In essence, the research done related to COVID-19 has been in two major areas. The first is the identification of COVID-19 using X-ray scans [1]. The second is the predictive modeling of the spread of COVID-19 using different approaches like network information [3]. Detecting SOP violations through various means is a research area that has been largely neglected.

The two basic SOPs are to maintain a 6 feet distance and to wear a face mask. These can easily be detected using closed-circuit television (CCTV) infrastructure which is present at many places where crowding can occur. However, it has been noticed that there is little research in this regard. Current computer vision models require high-end devices with graphics processing units (GPUs) that can handle deep learning models. This makes large scale deployment infeasible.

There are three basic parts of detecting the aforementioned SOP violations using CCTV video streams. The first is person detection using object detection algorithms. Humans can be detected in CCTV footage using a variety of deep learning models or classical techniques. The second part is the estimation of the distance between two detected humans to ensure that there are at least 6 feet apart. The third part is the detection of facemasks.

Research is being carried out in the above mentioned areas, but few studies have tried to determine the most suitable techniques for COVID-19 SOP violation detection. In this survey, the prevalent techniques used in this research area are described in Section II (classical), Section III and IV (deep learning based). In Section IV we also highlight why our approach is necessary by comparing previously used giving a brief analysis of their capabilities and their suitability to the research problem. In Section V we propose a new approach using a lightweight architecture i.e. you only look once (YOLO).

II. Classical Human And Object Detection Techniques

In this section, we will summarize some of the techniques that have been used before deep learning became popular. Before 2015 there has been limited research on detecting facemasks from images; largely due to the complexity of the task, while only statistical techniques were available.

A. Human Detection and Object Tracking Based on Histogram of Gradients (HOG)

Initially, it was proposed by Zhang and Wang [4] to implement the feature descriptor of HOG with a support vector machine (SVM) classifier. They improved this by making the use of multi-scale HOG features, narrowed down by the AdaBoost algorithm which selects the main features. They are then boosted (combined) to produce strong classifiers.

A major disadvantage of HOG is the high dimensionality of the feature vectors which increases the

978-0-7381-0527-7/20/$31.00 ©2020 IEEE

computation cost of SVM classification. The paper addresses this problem by adopting HOG as basic features and then creates much reduced features with the AdaBoost algorithm. A linear SVM is trained for classification with the help of the reduced features. Comparing results on the INRIA dataset, this detector produces similar results with a small feature size resulting in low computation cost and storage requirements. The results are shown in Fig. 1.

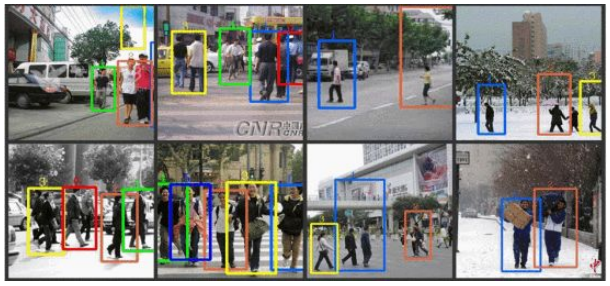

Fig. 1. Experimental results of HOG [4]

A limitation of automated object analysis is that it must be continually identified over time. It is a major fallacy since all the tracking algorithms terminate when subjected to a lack of identity. This problem is enhanced by objects in close proximity, shadows, etc.

B. Human Detection Based on Curvelet Transform

A problem in human detection is change in lighting, pose, clothing, occlusion, and cluttered backgrounds. One method is using curvelet feature extraction on blocks. It is based upon the statistical information extracted from blocks.

The main advantage of the curvelet transform [5] is the sparse representation of the discontinuities in an image. An image can be thought of as a combination of frequency bands. The lowest frequency bands contain lineament edge information whereas higher frequency bands contain finer edge information. The sub-band coefficients are partitioned into equal-sized blocks for which the statistical measures are calculated to get the edge feature vector.

For each block the entropy, energy, standard deviation (SD), mean, minimum value, maximum value, and contrast are selected. Then entropy, energy, SD, max value, and contrast are selected to be the optimal to compose the edge features. The edge features of a human are then constructed by concatenating each edge to one big feature vector.

C. Hausdorff Distance for Target Detection

Three types of noise affect object detection. 1) run time changes in the environment, 2) distortion by the device, and 3) pre-processing distortion such as digitalization, segmentation, etc.

Gastaldo and Zunino [6] describe Hausdorff distance for object detection in static images, which measures the degree of resemblance between two objects. It has increased accuracy because it measures the proximity rather than the exact superposition; which means that small distortions in the image do not affect the performance of the system.

D. Pedestrian Detection by Segmentation and using Virtual Mask

Choo et al [7] present a real-time pedestrian crossing system by using image segmentation and a virtual mask. Image segmentation is used to characterize the regions in the frames which include moving objects. The virtual mask removes areas below a specific threshold size. By only focusing on the regions of interest rather than the whole frame, higher computation speed can be achieved. The HOG feature set along with the SVM classifier is adopted due to their high discrimination capability. To make it feasible to apply to real-time systems, they make use of image segmentation and masking.

E. Illumination-based Background Subtraction

The object detection techniques start from using conventional statistical techniques to detecting humans in images. Apoorva Raghunandan; Mohana, Pakala Raghav, H. V. Ravish Aradhya [8] use a MATLAB for Local Illumination-based Background subtraction method to perform foreground extraction. combined with the Viola-Jones algorithm they can achieve an accuracy of 95% for object detection. The accuracy, however, is only obtained on zoomed static images therefore the technique cannot be used for real-time applications.

III. Pre-COVID-19 Deep Learning Based Object And Face Mask Detection Techniques

In this section, we will explore the techniques that have been prevalent before the spread of COVID-19. Deep learning frameworks are very common, starting from convolutional neural network (CNN) architectures to the more recent YOLO, which takes object detection towards real-time.

A. Object Detection Techniques

Lee et al [9] use an ensemble of CNNs in a two-stage detection setup with a ResNet feature extractor which achieves a mean average precision (mAP) of 0.8421 on the PASCAL VOC 2007 dataset.

The research then moves towards R-CNN which is a region-based CNN. It divides the input images into different regions which are fed to the CNN model which only outputs a single outcome per region. Dong and Wang [10] use the R-CNN for pedestrian detection in images from the INRIA dataset and obtained a miss rate of 14.1%. This is done using an NVIDIA Tesla K20 GPU.

A faster R-CNN architecture has emerged [11-12] which uses modification on the region based CNN network. It employs feature extraction first on the entire image and then only passes the feature vectors into the fully connected layers instead of the entire image. This saves computational cost.

This architecture is succeeded by the single shot detectors (SSD) and YOLO architectures. Both of these can achieve real-time multi-class object detection using GPUs but they are still too complex for basic GPUs to handle.

The SSD architecture [13] uses the basic CNN architecture with a different approach. It categorizes the output space into different boxes based on resolution and aspect ratios per feature map. It then outputs only one category per bounding box and combines the output predictions from multiple boxes as well as different resolutions to produce a final output. Using only a limited number of bounding boxes per feature map allows the SSD to reduce the inference time by a great margin.

The YOLO architecture is proposed by Redmon et al [14] and further worked on by [15-16]. It is similar to SSD in the sense that they both have the backbone of a CNN feature extractor. YOLO then downsamples the images according to fixed aspect ratios and performs predictions based on the preset number of bounding boxes in the total image. The predictions on bounding boxes from different feature maps are then aggregated onto the base image by up-sampling the output image using the fixed ratios.

B. *Face Mask Detection Techniques*

Research related to face mask detection remained ignored before 2015 as it was not a critical need. However, there have been many research papers [17-19] since then that tackle this problem. Street crimes using different kinds of masks are among the major reasons for research in this area.

Fig. 2. Cascaded CNN structure results [18]

Deore et al [17] use a simple HOG feature extractor along with the Viola-Jones algorithm. It is a combination of Haar feature selection, integral image creation, Adaboost training, and cascading classifiers. It classifies first the mouth, nose, and ears. If a person's mouth is not detected it is assumed that he is wearing a mask.

Bu et al [18] use a CNN based approach to detect masked faces. Their CNN architecture consists of 3 CNN layers, each of which removes false detection windows making it more robust. Their model achieves 86.6% accuracy on their custom MASKED FACE dataset. The architecture's results can be seen in Fig. 2.

Ejaz et al [19] implemented a principal component analysis (PCA) feature extractor. Their model detects faces using the Viola-Jones algorithm, performs image processing on it, then uses PCA to extract features. The model gives the best accuracy of 73.75% on masked faces.

There has only been primitive research on detecting masked faces before the spread of COVID-19. Most of the techniques for distance estimation are widely available on the internet and they perform well. The most widely used technique is the estimation of the distance between objects using a centroid approach. A centroid is first calculated for the detected object and then its Euclidean distance is retrieved from all other detected objects.

IV. Post-COVID-19 Deep Learning Based Object And Face Mask Detection Techniques

In this section, we give a brief view of computer vision techniques used after the spread of COVID-19. A method called "backbone, neck, and head technique" is popular. The backbone contains pre-trained convolutional neural networks like ResNet, VGG, and DenseNet. The neck is an extra layer. e.g. YOLO v3 uses a feature pyramid network (FPN) as a neck. The head includes the classification and regression.

A. *Object Detection and Human Distancing Techniques*

Punn et al [20] suggest a technique for the detection of objects and human distancing to help in dealing with COVID-19. For object detection, he proposes YOLO v3 [21] alongside Deepsort [22] for tracking pedestrians. The pairwise L2-norm is calculated with the aid of the bounding boxes. To identify the individuals in clusters not following the social distancing, a vectorized representation is presented. This achieves an mAP of 33% on PASCAL VOC and 74% on the MICROSOFT COCO dataset.

Rezaei et al [23] put forward a human detector model based on deep neural networks. This DeepSOCIAL model distinguishes and tracks individuals to prevent the spread of COVID-19. CSPDarkNet53 is used as the backbone (The model's capability is expanded with the increase in the number of parameters in distinguishing numerous objects). It utilizes the spatial attention module (SAM) [24] and spatial pyramid pooling (SPP) as the neck and mish activation function as the head. The model achieves 99.8% accuracy in a real-time environment.

Yang et al [25] present a technique for observing human distancing in real-time through AI and a monocular camera. The framework uses a deep CNN. Region of interest (ROI) is employed to abstain from crowding by changing the inflow. Both YOLO v4 and faster R-CNN are incorporated for pedestrian detection. This achieves an mAP of 42.7% on faster R-CNN Model and 43.5% on YOLO v4. The analysis is carried out on an Intel Core i7-4790 CPU with Nvidia GeForce GTX 1070Ti GPU using the Ubuntu operating system.

B. *Face Mask Detection Techniques*

The objects of different colors are distinguished better in hue-saturation-value (HSV) color space than red-green-blue (RGB). Li et al [26] present a model that utilizes the HSV color channel for the processing of masked faces. This classifies head pose images with masks through color texture inspection. The RGB color channel is converted into HSV. Mean filtering and binarization are performed on the H-channel. Dong et al [27] propose a method to get the facial contour and feature information. The size of the convolutional kernel and input picture is changed based on AlexNet. This approach shows 0.8% (front) and 2.28% (side) more accuracy than other approaches..

Qin et al [28] build up the identification of masked faces model with the help of image super-resolution with classification network (SRCNet). They utilize MATLAB image handling tools to process raw images. Then the multitask cascaded CNN is used for face detection. The

super-resolution network used in this is made with the residual encoder-decoder (RED) network [29]. RED utilizes convolutional layers as auto-encoder and deconvolutional layers for image upsampling. For identification of masked faces, Mobilenet-v2 is incorporated. This system achieves 98.70% accuracy.

The architecture proposed by Jiang et al. [30] applies the same concept of "backbone, neck, and head". ResNet and MobileNet are used as the backbone that performs feature extraction with CNN. For the neck, the feature pyramid network (FPN) extracts the high-level semantic information and then adds it with the previous layer. A multi-scale detection strategy as SSD is used as a head to predict with multiple FPN feature maps. With each head, a convolutional block attention module (CBAM) is infused to point out specific areas. It achieves 93.4% (ResNet) and 82.3% (MobileNet) accuracy.

TABLE 1. COMPARISON OF PERSON DETECTION MODELS/TECHNIQUES

Model/ Technique	Accuracy (%)/mAP	Dataset	Limitations/ Benefits
HOG [4]	-	INRIA	5.6 FPS on small images
Curvelet Transform [5]	90.17%	INRIA	Low FPS
Hausdorff Distance [6]	90%	custom	Custom dataset
Image Segmentation and Virtual Mask [7]	83.29%	custom	A large number of false positives & negatives
Viola-Jones [8]	95%	custom	Static images only
CNN ensemble [9]	82.4%	PASCAL VOC 2012	Less than real-time
R-CNN [10]	86%	INRIA	Less than real-time
Faster R-CNN [11]	66	PASCAL VOC 2012	Less than real-time
MobileNet+SSD [13]	60.6	Custom COCO subset	3 sec/frame with optimum object-camera distance
YOLOv1 [14]	63.4	PASCAL VOC 2007	fast YOLO 155 FPS
YOLOv3 [21]	57.9	PASCAL VOC 2007	51 ms inference time
DeepSOCIAL [23]	99.8% precision/ 97.6 recall	Oxford Town Centre	Requires heavy GPU
YOLOv4 + R-CNN [25]	41.2	Oxford Town Center etc.	48 ms inference time

TABLE 2: COMPARISON OF FACE MASK DETECTION MODELS/TECHNIQUES

Model/ Technique	Accuracy (%)/mAP	Dataset	Limitations/ Benefits
HOG with Viola-Jones [17]	46.6%	Custom	Low accuracy
CNN [18]	86.6%	Wider Face + Masked Face	
PCA [19]	83%	ORL face + custom	Gives poor performance on masked faces
HSV Color Channel + CNN [26]	90%	MAFA	Requires heavy GPU
SRCNet [28]	98.7%	Medical Masks	Requires heavy GPU
SSD + ResNet + FPN [30]	92% precision/ 94% recall	Public Face Mask	Requires heavy GPU

Table 1 and 2 present tabular comparisons of various techniques that can be used for COVID-19 SOP violation detection.

V. Proposed Approach

Our approach combines the detection of face masks and social distancing in a lightweight system that does not require GPUs. The benefit of such a system is its local deployability on mobile devices without requiring constant connection to a cloud server. We propose a tri-partied approach. The first module will be human detection. For this purpose, we propose to use a lightweight computation model like [16] which will initially partition the image. This will be followed by selecting the appropriate features using CNN with the help of the RFB module which aggregates multiple convolutions onto each other.

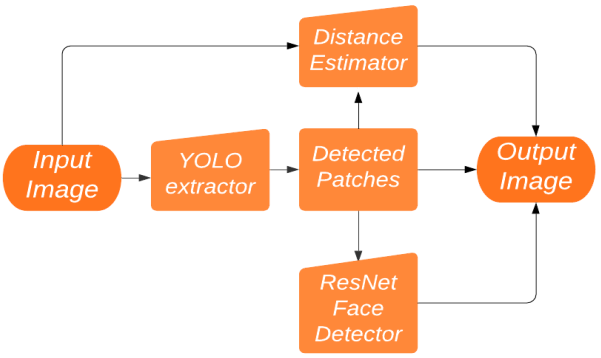

Fig. 3. Proposed system architecture

The second module is the detection of face masks and the classification of masked and unmasked faces. CNN's with transfer learning approach is used. We propose to use ResNet as the pre-trained model that will perform feature selection and output generation.

The third module is based upon the distance calculation between humans using Euclidean distances between centroids. We shall calculate the L2 norm in a three-dimensional feature space based on a set of bounding boxes with the ID of each person. The closeness will be judged upon the violation of a certain threshold. An overview of the architecture is shown in Fig. 3.

VI. Conclusion

While individual modules are well developed and achieve reasonable accuracies, there has been a lack of research towards formulating a combined tool that can help the authorities classify SOP violations. Moreover, the face mask detection techniques that have been employed are all done on somewhat outdated architectures instead of the new lightweight deep learning algorithms available now.

The task now is to focus on building a tool that solves both the problems of object detection (humans) and face mask detection using a lightweight architecture to save computational costs. Additionally, these tools will only

become viable if they can run along with the existing infrastructure of CCTVs present worldwide.

Preliminary research shows that this is possible using optimized tiny-YOLO or slim-SSD architectures which are continuously being modified to make them even lighter than before allowing them to run on non-GPU-based computers giving a decent FPS.

REFERENCES

[1] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu and C. Zheng, "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT," *IEEE Transactions on Medical Imaging,* vol. 39, no. 8, pp. 2615-2625, Aug. 2020.

[2] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19-19 Detection in Chest X-Rays," *IEEE Access*, vol. 8, pp. 115041-115050, 2020.

[3] A. A. R. Alsaeedy and E. K. P. Chong, "Detecting Regions At Risk for Spreading COVID-19 Using Existing Cellular Wireless Network Functionalities," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 187-189, 2020..

[4] S. Zhang and X. Wang, "Human detection and object tracking based on Histograms of Oriented Gradients," in *Ninth International Conference on Natural Computation (ICNC),* Shenyang, pp. 1349-1353, 2013.

[5] Hong Han, Youjian Fan and Zhichao Chen, "Human detection based on Curvelet transform," in *International Conference on Multimedia Technology*, Hangzhou, pp. 356-359, 2011.

[6] P. Gastaldo and R. Zunino, "Hausdorff distance for target detection," in *IEEE International Symposium on Circuits and Systems*, Phoenix-Scottsdale, AZ, USA, pp. V-V, 2002.

[7] C. Y. Choo, K. Lee, H. Q. See, Z. J. Tan and S. W. Khor, "Pedestrian detection with image segmentation and virtual mask," in *3rd International Conference on Computer Science and Information Technology,* Chengdu, China, pp. 23-27, 2010.

[8] A. Raghunandan, Mohana, P. Raghav and H. V. R. Aradhya, "Object Detection Algorithms for Video Surveillance Applications," in *International Conference on Communication and Signal Processing (ICCSP), Chennai*, pp. 0563-0568, 2018.

[9] J. Lee, S. Lee and S. Yang, "An Ensemble Method of CNN Models for Object Detection," in *International Conference on Information and Communication Technology Convergence (ICTC),* Jeju, pp. 898-901, 2018.

[10] P. Dong and W. Wang, "Better region proposals for pedestrian detection with R-CNN," *Visual Communications and Image Processing (VCIP)*, Chengdu, pp. 1-4, 2016.

[11] T. Liu, H. Y. Fu, Q. Wen, D. K. Zhang and L. F. Li, "Extended faster R-CNN for long distance human detection: Finding pedestrians in UAV images," in *IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, pp. 1-2, 2018.

[12] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp. 1440-1448, 2016.

[13] S. Kanimozhi, G. Gayathri and T. Mala, "Multiple Real-time object identification using Single shot Multi-Box detection," in *International Conference on Computational Intelligence in Data Science (ICCIDS),* Chennai, India, pp. 1-5, 2019.

[14] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, pp. 779-788, 2016.

[15] W. Lan, J. Dang, Y. Wang and S. Wang, "Pedestrian Detection Based on YOLO Network Model," *IEEE International Conference on Mechatronics and Automation (ICMA)*, Changchun, pp. 1547-1551, 2018.

[16] R. Huang, J. Pedoeem and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," in *IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, pp. 2503-2510, 2018.

[17] G. Deore, R. Bodhula, V. Udpikar and V. More, "Study of masked face detection approach in video analytics," in *Conference on Advances in Signal Processing (CASP)*, Pune, pp. 196-200, 2016.

[18] W. Bu, J. Xiao, C. Zhou, M. Yang and C. Peng, "A cascade framework for masked face detection," in *IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM),* Ningbo, pp. 458-462, 2017.

[19] M. S. Ejaz, M. R. Islam, M. Sifatullah and A. Sarker, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition," in *1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, pp. 1-5, 2019.

[20] N. S. Punn, S. K. Sonbhadra and S. Agarwal, "Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLOv3 and Deepsort techniques," *arXiv preprint arXiv:2005.01385*, 2020.

[21] J. R. A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018

[22] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime," *arXiv preprint arXiv:1703.07402*, 2017.

[23] M. Rezaei and M. Azarmi, "DeepSOCIAL: Social Distancing Monitoring and Infection Risk Assessment in COVID-19 Pandemic," *Applied Sciences,* vol. 10, no. 21, 7514, 2020.

[24] S. P. Woo, J., J.-Y. Lee and I. S. Kweon, "Convolutional Block Attention Module," *arXiv preprint arXiv:1807.06521*, 2018.

[25] D. Yang, E. Yurtsever, V. Renganathan, K. A. Redmill and U. Ozguner, "A Vision-based Social Distancing and Critical Density Detection System for COVID-19," *arXiv preprint arXiv:2007.03578*, 2020.

[26] S. Li, X. Ning, L. Yu, L. Zhang, X. Dong, Y. Shi and W. He, "Multi-angle Head Pose Classification when Wearing the Mask for Face Recognition under the COVID-19 Coronavirus Epidemic," in *International Conference on High Performance Big Data and Intelligent Systems*, 2020.

[27] D. Xiaoli, "Research on technologies of machine artistic portrait based on facial images," *University of Chinese Academy of Sciences*, 2018.

[28] B. Qin and D. Li, "Identifying Facemask-Wearing Condition Using Image Super-Resolution with Classification Network to Prevent COVID-19," *Sensors*, vol. 20, no. 18, 5236, 2020.

[29] X.-J. Mao, C. Shen and Y.-B. Yang, "Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections," *arXiv preprint arXiv:1606.08921*, 2016.

[30] M. Jiang, X. Fan and H. Yan, "RETINAFACEMASK: a face mask detector,", *arXiv preprint arXiv:2005.03950v2,* 2020.