Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.020586
Article

Tech Science Press

Real Time Feature Extraction Deep-CNN for Mask Detection

Hanan A. Hosni Mahmoud, Norah S. Alghamdi and Amal H. Alharbi*

Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11047, KSA
*Corresponding Author: Amal H. Alharbi. Email: ahalharbi@pnu.edu.sa
Received: 30 May 2021; Accepted: 16 July 2021

Abstract: COVID-19 pandemic outbreak became one of the serious threats to humans. As there is no cure yet for this virus, we have to control the spread of Coronavirus through precautions. One of the effective precautions as announced by the World Health Organization is mask wearing. Surveillance systems in crowded places can lead to detection of people wearing masks. Therefore, it is highly urgent for computerized mask detection methods that can operate in real-time. As for now, most countries demand mask-wearing in public places to avoid the spreading of this virus. In this paper, we are presenting an object detection technique using a single camera, which presents real-time mask detection in closed places. Our contributions are as follows: 1) presenting a real time feature extraction module to improve the detection computational time; 2) enhancing the extracted features learned from the deep convolutional neural network models to improve small objects detection. The proposed model is a lightweight backbone CNN which ensures real time mask detection. The accuracy is also enhanced by utilizing the feature enhancement module after some of the convolution layers in the CNN. We performed extensive experiments comparing our model to the single-shot detector (SDD) and YoloV3 neural network models, which are the state-of-the-art models in the literature. The comparison shows that the result of our proposed model achieves 95.9% accuracy which is 21% higher than SSD and 17.7% higher than YoloV3 accuracy. We also conducted experiments testing the mask detection speed. It was found that our model achieves average detection time of 0.85s for images of size 1024 × 1024 pixels, which is better than the speed achieved by SSD but slightly less than the speed of YoloV3.

Keywords: Mask detection; classification; neural network; texture features; transfer learning

1 Introduction

The World Health Organization was informed of cases of Covid19 in Wuhan China at the end of year 2019 [1]. Millions of cases have been confirmed in the world. Many countries have taken public measures including mask wearing, social distance surveillance to fight Covid19 pandemic. In an effort to limit the spreading of the virus, medical experts recommend mask wearing in public [2].

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Many crowded places such as Supermarkets, malls and medical facilities can face high risk of infection. Not wearing masks in such places can cause hazards to public health. Therefore, in this research, we propose a real-time mask wearing detection to detect whether attenders of these places are wearing masks. We propose a lightweight SSD algorithm for mask wearing detection utilizing separable convolution in spatial dimension as well as proposing a Feature Enhancement process to enhance the mask wearing detection procedure [3,4].

Manual checking for face masks is very limiting with inadequate resources, and can be prone to errors. There is an instantaneous necessity for a better solution to manage the virus spread by educating the perfect social distancing standards. This includes solutions for distance detection between people and also solutions for face mask detection to govern the safety of the public by examining that sufficient distance is sustained and that face masks are in use. These models can assist in a varied range of safety keeping in public places with cameras such as in malls and shopping centers and different surveillance applications [5,6].

This paper is presented as follows: Section 2 investigates the previous research in literature. Section 3 presents the dataset of mask-wearing for COVID-19. Our proposed model is depicted in Section 4. The experiments and result analysis are depicted in section 5. While conclusions are displayed in section 6.

2 Related Work

Great advances have happened in the field of Deep- learning paradigm such as Deep-CNN which leads to significant improvement in computer vision object detection [2]. Deep-CNN gained a lot of success in large object detection. On the contrary, Deep-CNN Did not perform well on small object detection. In complex situations of crowd surveillance, the objects in the images or video frames are usually small and distant [3,4]. The system proposed by the authors [5] is an incorporation of deep transfer learning CNN and machine learning model. They utilized ResNet-50 by replacing the last layer with SMV machine learning techniques. They utilized an ensemble of SVM and decision trees to enhance the system performance. The authors utilized two datasets, a dataset includes a large number of faces wearing masks and other fake masks. The only drawback is the time consumption related to other datasets throughout the training phase. There is also no reported accuracy according to related works for this type of dataset.

A backbone detection model of heads is employed using Resnet [6–8]. Also a feature pyramid model is utilized for neck predictors [9]. The limitation was the restricted size of the mask dataset which is problematic for machine learning models to learn better features. A self-developed system was proposed by the authors in [10], SocialdistancingNet-19 is a self-developed system for detecting people in video frames and exhibiting labels of safe for social distancing. The distance is computed between people in the frame utilizing centroids. The authors in [11], presented a deep learning CNN utilizing MobileNet and OpenCV for mask detection. This model faced the incorrect categorization of hands over faces as masked. These situations are not appropriate for this model. In [12], the authors utilized the DBSCAN system to compute distances among people and accomplish clustering for distance computation. This model showed better efficiency than other clustering algorithms such as k-means and c-means.

Efforts in detecting small objects have been made [3–10]. The usual model, as depicted in [3] and [5], is to increase the resolution of small size objects by image scaling. But such a technique is time consuming in both the training and testing phases. Other researchers such as the authors in Yang et al. [5–8] utilize multi-scale representation to increase feature resolution by incorporating lower-level features in multi layers. This is not a practical technique due to the complication in the feature dimension. In the following subsections, we are going to focus on the small-size object detection paradigm.

Recent research in analysis of non-linear data has attracted attention in the image analysis and processing field. These approaches are totally unsupervised with reliable one-dimensional signals. These methods also can be applied for texture extraction, which is deemed as a computer vision challenge [7].

2.1 Small-Size Object Detection in Surveillance Setting

Small-size object detection in surveillance settings has been studied extensively in the video processing paradigm. Authors in [13] proposed many algorithms to address this process. Deep-CNN techniques are widely utilized in remote sensing detection especially for small-sized objects due to their high performance. Authors in [11] developed a decomposition CNN technique for object detection in space images, which designed an efficient methodology for feature learning of remote images. Authors in [12] developed a rotation-invariant neural network to identify multiple objects in an optical remote image. While the authors in [13] incorporated a deformation model for small object identification through multiple-stage cascading CNN layers.

2.2 Far Traffic Sign Detection

Self-driving vehicles need many measures to drive correctly. The important measure is far traffic color recognition. Authors in [14] presented a multi-features extraction that preceded the classification process. They utilized a methodology of skipping convolution layers in order to boost recognition. In [15], the authors developed two deep-CNNs, the first one is used to localize objects and then the second CNN classifies traffic signs. The authors in [16], developed a hinge stochastic gradient method for the training phase that enhances the accuracy and leads to stable convergence in less time.

3 Dataset

We utilized a dataset built from public crowd images from the internet especially from datasets published in [17] and [18]. Our dataset contains three types of images, the first set is of people in a crowd with face masks, and the second set is for crowds of people without masks, while the third set is a hybrid of people in crowds where some of them wear masks and some of them without masks. All images were labeled with their set name, some of the images from the three datasets are displayed in Fig. 1.

Images from our dataset include the first set for people in a crowd with face masks, while the second set is for crowds of people without masks. The third set is a hybrid of people in crowds where some of them wear masks and some of them without masks.

The distribution of the face sizes in the three datasets are depicted Fig. 2. Where the average number of faces in the images are plotted against the sizes of the faces in pixel square. It can be extracted from Fig. 2 that most faces are of sizes less than 150-pixel square. While large faces of size more than 500 pixels are very few. The statistical properties of the dataset are depicted in detail in Tab. 1, where we have 3055 images for people wearing masks, while 2920 images with all people in them are not wearing masks. We also have 3100 images for hybrid combinations where some of the people in the same images are wearing masks and some are not wearing masks.

Statistical properties of the dataset such as the mask wearing status, size of the images and number of large and small faces are presented in Tab. 1. The total number of images used are 9075.

Figure 1: Images from our dataset (a) people in a crowd with face masks, (b) crowds of people without masks, (c) is a hybrid of people in crowds (some with masks)

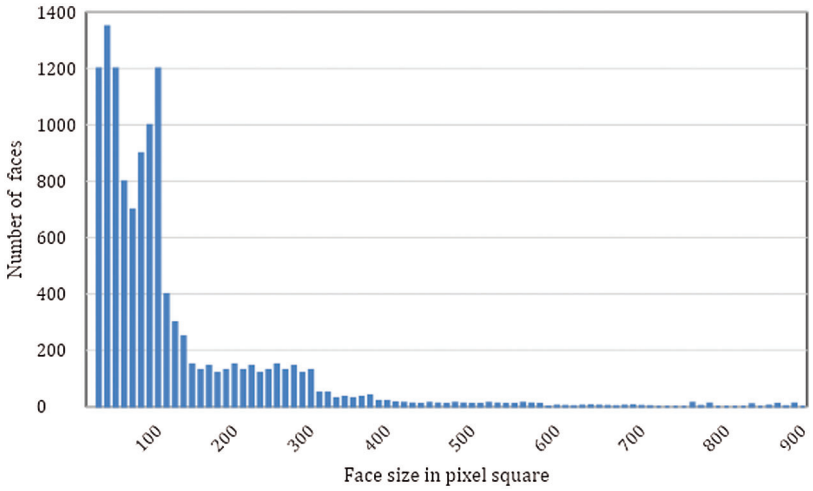

Figure 2: Distribution of face sizes in our dataset

4 Methodology

In this section, we will describe the methodology of our proposed model which modifies the Single Shot detector model (SSD) to enhance its performance [19]. We propose a detection SSD to detect whether people are having masks on their faces. Previous experiments proved that the SSD model has a high miss positive

and positive negative detection rates for small-sized objects. Two problems arise while using SSD in detection of small objects, the first one is the low accuracy and the high detection computational time. We proposed an optimization technique for the SSD. Lightweight backbone CNN and Feature Enhancement algorithms are the basis of our framework. The novelty in our model is the lightweight backbone CNN which enhances runtime performance by reducing runtime to be fit for real time mask detection. Also, accuracy is highly enhanced through introducing a novel optimization technique for the SDD by inserting the feature enhancement module after convolution layers 3 and 5 to enhance the accuracy and decrease the false negative rate as depicted by the experimental results. Our proposed mask wearing detection framework is depicted in Fig. 3.

Table 1: Details of the mask dataset

Mask wearing status	Number of images	Size in Pixels	Faces per image	Number of small faces per image
Wearing a mask	3055	1024x1024	14	6
Not wearing a mask	2920	1024x1024	13	7
Hybrid, some faces have a mask and some doesn't have	3100	1024x1024	15	8

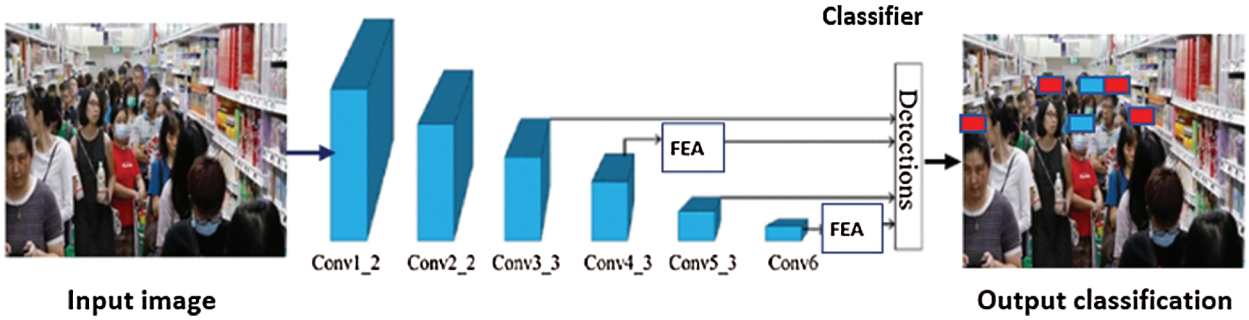

Figure 3: The proposed model (convolution layers are succeeded by the feature enhancement algorithm FEA)

4.1 Face Region Detection

Face region detection in surveillance images is very important to apply mask wearing detection methods. We propose a fast technique for real time requirements. The face region technique is based on converting the color image which is an input from the surveillance camera into gray image followed by denoising using a low pass filter. The image is then transformed using window standard deviation and then converted to binary image utilizing adaptive threshold technique. Other techniques can utilize multi-scale principal component analysis techniques, which are usually used in the preprocessing phase to perform denoising [9].

The binary image has protruding boundaries and prominent facial features such as eyebrow, eyes, and mouth. These features are identified by applying morphological operations. The image is then scanned in the vertical direction to locate the probable box covering the face. Our proposed method determines the face region by identifying the eye-brow in the possible face surrounding box. Experiments have been performed on single shot image dataset images and it is giving high accuracy of 94.5% in multi faces images.

The algorithm of the face detection using a single shot image can be summarized in the following steps.

1. Convert the input image into binary image utilizing LWSD.
2. Create a denoised binary image utilizing Adaptive Threshold technique.
3. Identify the face box region utilizing vertical profiling.
4. Locate the top of the face.
5. Compute the box from the top of the face and 1.5 of the width as computed obtained in Step 3 and 4 as the face region.
6. Identify the eye-brows using horizontal profiling.
7. Detect the region in the input image and output it as the face region.

4.2 Lightweight Backbone Network

Our lightweight backbone CNN proposal for mask wearing detection utilizes SSD algorithm and separable spatial CNN layers. Our technique is built such as:

1) The extracted feature map from the VGG-16 network shallow layer includes greater number of features from small-sized objects [18].

2) Reducing the computational cost of the deep-CNN.

Deep-CNN faces real time and space challenges in many applications. The required computational power leads to CPU and GPU that needs a large space, also it leads to poor real-time performance. Therefore, lightweight CNNs such as Mobilenet [19] and EffNet [20], are needed. EffNet utilizes a separable convolution network.

SSC divides the CNN into kernels of smaller size. We present the case where the CNN is divided as depicted in Tab. 2.

Table 2: The CNN layers

Layer	Layer type	Properties
1	Input	1024 × 1024 images
2	Number 1 Convolutional	360 × 6 × 1 convolutions
3	Number 1 First Pooling	Max pooling
4	Number 2 Convolutional	128 (3 × 3 × 32) convolutions
5	Number 2 Pooling	2 × 2 max pooling
6	Number 3 Convolutional	256 (3 × 3 × 16)
7	Number 3 Pooling	2 × 2 max pooling
8	Feature Enhancement	FEA
9	Number 4 Convolutional	256 (2 × 2 × 16)
10	Number 5 Convolutional	256 (2 × 2 × 16)
11	Number 6 Convolutional	256 (2 × 2 × 16)
12	Feature Enhancement	FAE
9	Softmax	Softmax
10	Classifier Output Layer	Two output classes: 1. Wears a mask 2. Non wearing a mask

4.3 Feature Enhancement Algorithm

Small sized objects in complex surveillance possess a difficulty in detection due to low resolution and low data volume. Improvement of the accuracy of detection of small sized objects can be increased by the flow of the Inception layer [21,22]. Therefore, we propose a Feature Enhancement algorithm that is based on feature fusion extracted from convolution layers with various kernel sizes. The Feature Enhancement algorithm for feature fusion is depicted in Fig. 4.

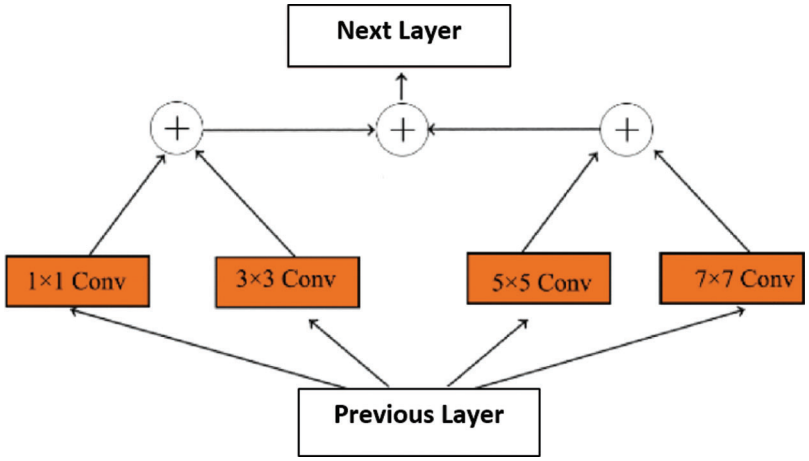

Figure 4: Flow diagram of the feature enhancement algorithm

4.4 SSD Architecture

We build the SSD architecture by designing appropriate convolution layers and appropriate box building, which are needed for small sized object detection.

1. Lightweight CNN network: The Lightweight network is constructed using SSD. We exchanged the CNN layers in reverse order. We also got rid of the other layers to achieve the lightweight purpose and reduce computational complexity. For real-time purposes we converted 'conv1_1' to 'conv6' To SSD network. We also defined convolution layers 3-6 for detection layers.
2. Feature Enhancement algorithm: to improve the features of small sized objects, we present the Feature Enhancement Algorithm. The two layers Conv4_3 and Conv6 are succeeded by the feature enhancement.
3. Boxes measures: To reduce the false negatives rate of small sized object detection, all detection layers have to be able to match the small sized scale. This can be accomplished by setting a number of small boxes. The box dimensions are depicted in Tab. 3.

Table 3: Box dimensions

Mask CNN detection layer	Box scale
Conv3	0.03
Conv4	0.15
Conv5	0.25
Conv6	0.45

4.5 Detection Block Diagram

The process incorporates two phases, the training phase and the detection phase. In the training phase, a mask dataset was utilized for model training to build a mask-wearing detection algorithm. In the first detection phase, images from surveillance video are used, and then the mask detector is used to detect whether the faces in the images are wearing masks or not. The block diagram of the model is depicted in Fig. 5.

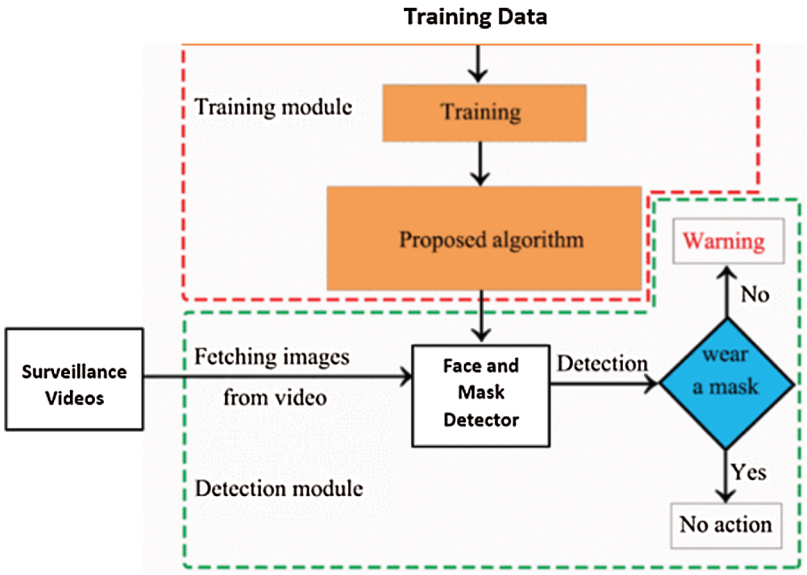

Figure 5: Block diagram of mask wearing detection

The system is composed of two phases: the training phase and the detection phase. The first phase includes transfer learning as indicated before. In the detection phase, images are fetched from the frames of the surveillance videos and used as input to the face and mask detector module. The decision is made after detecting a face and then classifying it as wearing a mask or not wearing a mask. The system will generate warning if the classifier decide that the person is not wearing a mask

5 Experiments

This section presents various experiments that are done on the Mask data set. We first perform the experiments on the architectures present in [19,22,23]. We then perform the experiments on the proposed lightweight CNN and the feature enhancement algorithm to evaluate their performance and their accuracy.

We utilized Keras Deep-CNN platform for the training and testing of our model. We also utilized Adam which is a Moment Estimation platform. The optimization technique for our model training starts with a learning rate of 0.002 and reduces the rate to 0.0002 after 25k iterations. We ceased the training phase after 90K iterations. The last iteration output is utilized to assess the performance of the small-sized object detection on the test dataset.

We performed various experiments on other similar algorithms in the literature. We conducted a comparison of the performance of our model and other similar models.

The comparison is depicted in Tab. 4. We can conclude from the results of our model from Tab. 4, as compared to other models, that our model accomplishes high detection accuracy as well as real-time

requirements on the Mask dataset. The performed Experiments depict results of our proposed model to achieve 95.9% accuracy which is 21% higher than SSD and 17.7% higher than YoloV3 accuracy.

Table 4: Performance evaluation of mask detection for object detection algorithms for small sized objects

Model	mAP	Wear a mask	No mask	Run Time (seconds)
R-CNN [22]	75.6	71.6	77.9	0.23
Single-Shot Detector (SSD) [19]	73.9	69.7	76.8	0.205
Yolo [23]	74.7	70.2	79	0.08
Our proposed model	95.9	92.3	93	0.085

We also conducted experiments testing the mask detection speed. It was found that our model achieves average detection time of 0.85s for images of size 1024 × 1024 pixels, which is better than the speed achieved by SSD but slightly less than the speed of YoloV3.

For further performance and effectiveness evaluation of the proposed model, we performed experiments using the ablation test. This testing process incorporates several testing procedures as depicted in Tab. 5. Test 1 is for the first unchanged SSD, and test 2 is for the altered SSD with 4 detection layers, without the spatial convolution and Feature Enhancement algorithm. Test 3 modifies test 2 by exchanging original convolution into spatial convolution. Test 4 incorporates Feature Enhancement algorithm into Test 3.

Table 5: Ablation testing results

Test	mAP %	Run time (seconds)
Test 1	74.2	0.21
Test 2	88.9	0.149
Test 3	87.7	0.11
Test 4	94.5	0.85

The comparison in Tab. 5, proves that the altered SSD in test 2 accomplishes 89.3% in mAp which is 6 times better than the original single-shot detector (SDD). This is due to using boxes scales that are more appropriate for the mask datasets with smaller objects. In test 3, the spatial separable module reduces the mAP but also reduces the runtime to a good extent. For test 4 we added the Feature Enhancement algorithm for training, and this helped the mAP is enhanced to 91.9% and runtime was reduced to reach 0.85 s. The experiments depict that our model helps the real-time requirement for detection to a great extent.

In Tab. 6, we compared the accuracy, sensitivity, specificity of our model using the 4 tests: test 1 to test 4. Also, we incorporated the comparison with the SDD model and the YoloV3 model. All our scenarios from Test 1 to Test 4 outperformed SDD and YoloV3 models.

Test 1 is for the original SSD and it has mask detection accuracy of 90.17%. Test 2, which utilize the altered SSD with four detection layers, outperformed Test 1 and has mask detection accuracy of 92.59%. We have to note that test 1 and 2 do not utilize the spatial convolution CNN or the Feature Enhancement algorithm. Test 3 modifies Test 2 by exchanging original convolution into spatial convolution and it

achieves mask detection accuracy of 94.91%. Test 4, which incorporates Feature Enhancement algorithm into Test 3, has the highest mask detection accuracy of 96.99%.

Table 6: Detection experiment results

Model	Accuracy	Sensitivity	Specificity	Mask detection accuracy	No-Mask detection accuracy
SDD	86.21	84.51	85.25	87.26	84.34
YoloV3	87.22	85.44	86.65	88.92	85.85
Our proposed model: Test 1	89.67	87.37	87.92	90.17	87.78
Our proposed model: Test 2	91.32	91.62	91.18	92.59	92.47
Our proposed model: Test 3	93.41	91.33	91.93	94.91	91.91
Our proposed model: Test 4	95.42	93.32	93.88	96.99	93.87

We also measured the Performance by computing the area under the curve in the ROC Curve (Receiver Operating Characteristics). Area under the curve indicates better performance as it reaches 1 (better classification). ROC plots the true positive and the false positive rates. Fig. 6 depicts the ROC area under the curves for the proposed CNN with and without the FEA. The ROC curves in Fig. 6 prove that the CNN which is combined with the FEA performs better than the CNN alone.

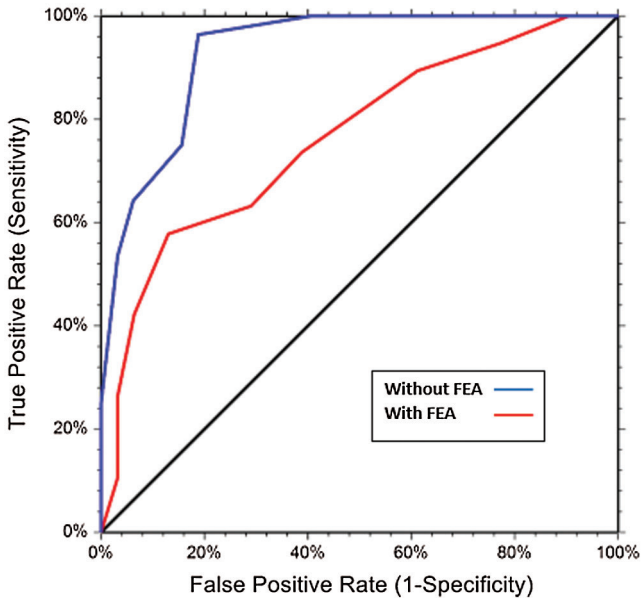

Figure 6: ROC curve for the proposed model with and without FEA

6 Conclusions and Future Work

In this article, we presented an altered SSD model to identify persons in surveillance videos with no masks. We were concerned about accuracy and real time requirements. For real time requirements, we presented a backbone network with lighter convolution layers. For better accuracy especially for small-sized faces we proposed a Feature Enhancement Algorithm (FEA). FEA was proved to enhance the

accuracy of the mask detection effect of our model. We performed various experiments to compare our model to the SDD and YoloV3 neural network models. Our proposed model accomplishes 95.9% accuracy which is 21% higher than SSD and 17.7% higher than an YoloV3 accuracy. We also tested the mask detection speed. It was perceived that our model has average mask detection time of 0.85s for images of sizes 1024 × 1024 pixels, which is better than the speed achieved by SSD but slightly less than the speed of YoloV3.

As future work, we will investigate other denoising methods such as the multiscale principal component analysis algorithm. This denoising algorithm can lead to better accuracy for detecting faces, especially small faces in a crowded image [24,25].

Funding Statement: This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Hui, I. AzharE and T. Madani, "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan China," in *Proc. of International Journal of Infectious Diseases*, Cleveland, Ohio91, pp. 264–266, 2020.

[2] Y. Liu, P. Sun and M. Highsmith, "Performance comparison of deep-learning techniques for recognizing birds in aerial images," in *Proc. IEEE Third Int. Conf. on Data Science in Cyberspace (DSC)*, Athens, Greece, pp. 317–324, 2018.

[3] X. Chen, K. Kundu and Y. Zhu, "3d object proposals for accurate object class detection," *Advances in Neural Information Processing Systems*, vol. 2, no. 1, pp. 424–432, 2015.

[4] W. Liu, D. Anguelov and D. Erhan, "SSD: Single shot detector," in *Proc. of European conf. on computer vision*, Paris, France, pp. 21–37, 2016.

[5] F. Yang, W. Choi and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Cairo, Egypt, pp. 210–221, 2019.

[6] C. Szegedy, S. Ioffe and V. Vanhoucke, "Inception-v4 inception-resnet and the impact of residual connections on learning," *International Journal of Artificial Intelligence*, vol. 2, no. 1, pp. 123–134, 2020.

[7] M. Sadiq, X. Yu, Z. Yuan, Z. Fan, A. Rehman *et al.,* "Motor imagery EEG signals classification based on mode amplitude and frequency components using empirical wavelet transform," *IEEE Access*, vol. 7, pp. 127678–127692, 2019.

[8] G. Cao, X. Xie and W. Yang, "Feature-fused SSD: Fast detection for small objects," in *Proc. Ninth Int. Conf. on Graphic and Image Processing*, New York, NY, vol. 10615, pp. 106–113, 2018.

[9] M. Sadiq, X. Yu, Z. Yuan, F. Zeming, A. Rehman *et al.,* "Motor Imagery EEG Signals Decoding by Multivariate Empirical Wavelet Transform-Based Framework for Robust Brain-Computer Interfaces," in *IEEE Access*. Vol. 7, pp. 171431–171451, 2019.

[10] G. Takacs, V. Chandrasekhar and S. Tsai, "Unified real-time tracking and recognition with rotation-invariant fast features," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Bern, Germany, pp. 934–941, 2020.

[11] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 2016.

[12] G. Cheng, P. Zhou and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[13] W. Ouyang, X. Wang and X. Zeng, "Deepid-net: Deformable deep-convolutional neural networks for object detection," in *Proc. of the IEEE conf. on computer vision and pattern recognition*, Ostrava, CZ, pp. 2403–2412, 2020.

[14] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Proc. of Int. Joint Conf. on Neural Networks*, Napoli, Italy, pp. 2809–2813, 2019.

[15] Z. Zhu, D. Liang and S. Zhang, "Traffic-sign detection and classification in the wild," *Computer Vision and Pattern Recognition*, vol. 2, no. 3, pp. 2110–2118, 2016.

[16] J. Jin, K. Fu and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.

[17] C. Sagonas, E. Antonakos, G. Tzimiropoulos and S. Zafeiriou, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, no. 6, pp. 3–18, 2016.

[18] A. Cabani, K. Hammoudi, H. Benhabiles and M. Melkemi, "Masked Face-Net-A dataset of correctly/incorrectly masked face images in the context of COVID-19," *Smart Health*, vol. 19, no. 1, pp. 125–137, 2020.

[19] W. Liu, D. Anguelov and D. Erhan, "SSD: Single shot multibox detector," *Computer Vision*, vol. 1, no. 2, pp. 21–37, 2019.

[20] S. Zhang, G. He and H. Chen, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 864–868, 2019.

[21] A. Howard, M. Zhu and B. Chen, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *Computer Vision*, vol. 2, no. 1, pp. 214–224, 2020.

[22] I. Freeman, L. Roese-Koerner and A. Kummert, "Effnet: An efficient structure for convolutional neural networks," *Image Processing*, vol. 2, no. 1, pp. 310–321, 2019.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better faster stronger," *Computer Vision and Pattern Recognition*, vol. 2, no. 1, pp. 117–129, 2019.

[24] M. Sadiq, X. Yu, Z. Yuan and Z. Aziz, "Motor imagery BCI classification based on novel two-dimensional modelling in empirical wavelet transform," *Electronics Letters*, vol. 56, no. 25, pp. 1367–1369, 2020.

[25] M. Sadiq, X. Yu and Z. Yuan, "Exploiting dimensionality reduction and neural network techniques for the development of expert brain-computer interfaces," *Expert Systems with Applications*, vol. 164, no. 10, pp. 114031, 2021.