IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 12, No. 1, March 2023, pp. 384~393
ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i1.pp384-393

Multiple face mask wearer detection based on YOLOv3 approach

Cheng Xiao Ge[1], Muhammad Amir As'ari[1,2], Nur Anis Jasmin Sufri[1]
[1]School of Biomedical Engineering and Health Sciences, Faculty of Engineering, Universiti Teknologi Malaysia, Johor, Malaysia
[2]Sport Innovation and Technology Centre (SITC), Universiti Teknologi Malaysia, Skudai Johor, Malaysia

Article Info

Article history:

Received Jan 30, 2022
Revised Jul 15, 2022
Accepted Aug 13, 2022

Keywords:

DarkNet-53
Face mask detection
Object detection
ResNet-50
You only look once version 3

ABSTRACT

The coronavirus disease 2019 (COVID-19) is a highly infectious disease caused by the SARS-CoV-2 coronavirus. In breaking the transmission chain of SARS-CoV-2, the government has made it compulsory for the people to wear a mask in public places to prevent COVID-19 transmission. Hence, an automated face mask detection is crucial to facilitate the monitoring process in ensuring people to wear a face mask in public. This project aims to develop an automated face and face mask detection for multiple people by applying deep learning-based object detection algorithm you only look once version 3 (YOLOv3). YOLOv3 object detection algorithm was concatenated with different backbones including ResNet-50 and Darknet-53 to develop the face and face mask detection model. Datasets were collected from online resources including Kaggle and Github and the images were filtered and labelled accordingly. The models were trained on 4393 images and evaluated based on precision, recall, mean average precision and the detection time. In conclusion, DarkNet53_YOLOv3 was chosen as the better model compared to ResNet50_YOLOv3 model with its good performance on accuracy with a mAP of 95.94% and a fast detection speed with a detection time of 50 seconds on 776 images.

Corresponding Author:

Muhammad Amir Bin As'ari
School of Biomedical Engineering and Health Sciences, Faculty of Engineering,
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Malaysia
Email: amir-asari@utm.my

1. INTRODUCTION

A newly discovered disease, called the Coronavirus disease 2019 (COVID-19) was first identified in Wuhan City, Hubei Province in China. It is a highly infectious disease caused by a novel coronavirus (SARS-CoV-2) and has caused approximately 173 780 789 confirmed cases worldwide with 3 737 818 deaths (till 5th June 2021) worldwide [1]. According to WHO, transmission of the virus can occur through droplet transmission, close contacts, airborne transmission, and fomite transmission [2]. During droplet transmission, the respiratory droplets or secretions will be exhaled from the infection patient when he or she coughs or sneezes. When another person comes into close contact (within 1 meter) with the infected patient, they will be transmitted into the eyes, nose or mouth of the person which leads to infection [3]. Airborne transmission normally occurs during aerosol-generating medical procedures [2]. People also will get infected by coronavirus by touching the eyes, nose, or mouth after touching the contaminated surface. To minimize the transmission of SARS-CoV-2, various preventive measures has been implemented including surface disinfection, improving hand hygiene, and practicing social distancing. As COVID-19 transmits through droplet transmission and close contact, wearing a mask especially in public is crucial to break the chain of

transmission of the virus. In general, the types of mask used in fighting the COVID-19 are surgical mask, filtering facepiece respirator (FFP) and non-medical masks. To protect the people from the infection, governments has made compulsory for wearing mask in the publics in many countries. Surveillance on the public in ensuring people are wearing face mask in accordance with the law enforcement is essential to reduce the risk of transmission efficiently. However, human surveillance is burdening to monitor large group of people on wearing of face masks in the public as well as maintaining a contactless manner between the surveillant and the public. Therefore, it is necessary to develop an automated face mask detection in public especially in crowded places to facilitate the monitoring process in ensuring people are wearing a face mask.

Deep learning is a subfield of machine learning and has achieved a significant development with its deep artificial neural network in recent years [4]. The vast improvement of deep learning has brought notable performance for visual recognition systems involving image classification, image localization and object detection [5]. In fact, deep neural networks are unable to tackle multidimensional input efficiently such as an image, as the parameters during training will be massive which is impractical [6]. As an alternative, convolutional neural network (CNN) was introduced to interpret image data and perform image classification related tasks in several area such as algriculture [7], [8], medical imaging [9]–[12] and security and surveillance [13]–[16]. It has been extensively used in visual recognition with its great capability in feature extraction and classification with single-object-image [17]. Instead of image classification, there are several deep learning-based object detection approaches formulated CNN as a core network in identifying the class of detected object. For example, region with CNN (R-CNN) [18], fast region with CNN (Fast R-CNN) [19], faster region with CNN (Faster R-CNN) [20], you only look once (YOLO) [21] and single shot detector (SSD) [22]. R-CNN requires high computation to perform object detection and is time-consuming. Both Fast R-CNN and Faster R-CNN performed well in accuracy, however, speed is compromised which is not suitable for real time detection [23]. For YOLO object detection, it is performed with a single neural network in a single evaluation and is fast and accurate [17]. In YOLOv2, the model has outperformed the other detection frameworks including Fast R-CNN, Faster R-CNN and SSD, with mean average precision (mAP) of 78.6 and speed of 40 frames per second (FPS), providing excellent results between speed and accuracy [23]. In comparison with YOLOv2, YOLOv3 improved its architecture proving a better accuracy and better performance on small-sized object detection [24], [25].

In general, deep learning-based approach for face mask can be generalized into: i) image classification; and ii) object detection. In image classification, CNN is established to distinguish between facemask and non-facemask image. For instance, Militante and Dionisio [26] used the CNN model called InceptionV3 through tranfer learning approach for classifying the facemask image together with social distancing. A dataset of 20,000 images consists of face mask wearing and non-face mask wearing was collected from Bing Search application programming interfaces (APIs) and taken in University of Antique. The training achieved an accuracy of 97% and the system was able to signal the alarm upon the detection of person who did not wear mask and observe social distancing. Other than that, the study in [27] also proposed InceptionV3 for facemask classification. The simulated masked face dataset (SMFD) dataset with a total of 1570 images were used, consists of 785 simulated masked facial images and 785 unmasked images. In this study, the last layer of the InceptionV3 model was removed and replaced by 5 more layers to the network. The last layer was the softmax activation function for the classification of mask wearing or non-mask wearing person. A comparison had been done between the proposed model and other machine learning and deep learning models including VGG-16, VGG-19, Xception, MobileNet and MobileNetV2. InceptionV3 outperformed other models and achieved accuracy and specificity of 100% and 100% during testing. Loey *et al.* [28] proposed a hybrid between deep transfer learning and machine learning. ResNet-50 was used as first component for feature extraction in the proposed model and classical machine learning including decision trees, support vector machine (SVM) and ensemble methods were used for classification purpose. Three datasets were used which are real-world masked face dataset (RMFD), simulated masked face dataset (SMFD) and labeled faces in the wild (WLF), each consists of 10,000 images (With and without masks), 1570 images (With and without masks) and 13,000 images (With masks only). As a result, SVM classifier outperformed the other classifier as it achieved 99.64% testing accuracy in RMFD, 99.49% in SMFD and 100% in labeled faces in the wild (LFW), with least consumption of time.

In object detection, facemask location is detected from the image based on the established deep learning-based object detection. For example, Roy *et al.* [29] used different object detection algorithms including YOLOv3, YOLOv3Tiny, SSD and Faster R-CNN were used for the detection of face and face mask. The models were trained using a dataset of 3000 images, where 678 images are from Kaggle datasets of medical face mask, 757 images consist of close-up faces and the rest 1565 images are from Google. As a result, YOLOv3Tiny was the perfect fit for the application, with a mAP of 56.27% and FPS of 138, where accuracy and speed are well-balanced. Loey *et al.* [30] used ResNet-50 deep transfer learning model for feature extraction process and YOLOv2 for medical face mask detection. A total of 1415 images were used for training with 853 images are from face mask dataset and 682 images from medical mask dataset. The

average precision achieved 81% and Adam optimizer is integrated in this study. Other than that, the works in [31], [32] proposed a face mask detection based on YOLOv3. Bhuiyan *et al.* [31] used a total of 600 images with 300 of masked and 300 of non -masked images achieved an accuracy of 96% with mean average precision score of 0.96. However, Ren and Liu [32] proposed YOLOv3 based model called Face_mask Net and perform the comparison of four loss functions which are Intersection over union (IoU), generalized IoU (GIoU), distance IoU (DIoU) and complete (CIoU). The dataset used has of a total of 9056 images consisting masked and non-masked images collected from WIDERFace, masked faces (MAFA), RMFD, web crawlers and video screenshots. Face_mask Net performed better than other networks in multi-target detection and is able to maintain accuracy at 99%.

Based on previous studies [29], [31]–[33], improvements can still be made on improving the accuracy as well as training a more robust model using bigger dataset. In additional, no studies have been done in comparing YOLOv3 model with different backbones such as DarkNet-53 and ResNet-50 on face mask detection. Thus, this project will be focused on formulating YOLOv3 for developing a more robust detection model using bigger and reliable dataset to achieve higher accuracy and speed as well as comparing the performance of both models with different backbones. As compare to the previous study, this study used the image samples obtained from six different dataset or studies which are face mask detection [34], medical mask [35], face mask dataset [36], face mask detection mask dataset [37], COVID face mask detection dataset [38] and correctly masked face dataset [39].

2. METHOD

The aim is to develop automated face and face mask detection model based on YOLOv3 object detection model using different deep pretrained CNN feature extraction model which are Darknet-53 and ResNet-50. The performance of each model in face and face mask detection is evaluated based on precision, recall, mean average precision (mAP) and detection time. The general method involved were illustrated according to the block diagram in Figure 1.

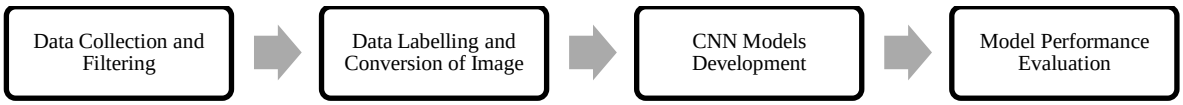


Figure 1. Block diagram of methodology

2.1. Data acquisition

The dataset with classes of masked and non-masked were obtain from online resources including Kaggle and Github as shown in Table 1. In total, there were around 17418 images from the 6 chosen dataset. As some of the images from the dataset were collected from the same source, there will be some duplicated images. Therefore, filtering of images was done beforehand. The selection of data images was based on the type of masks worn by the person, where only person who wears surgical masks, fabric masks, filtering facepiece respirators such as N95 will be selected in this project. Masks that covered full face of the person such as powered air purifying respirator were excluded. In addition, masks with complicated patterns were also filtered out. Some of the images from the dataset also consists of irrelevant images such as close-up facial images and images with low resolution, these images were filtered out as well. The images were filtered manually. Figure 2 represent the sample images that are selected for final dataset.

2.2. Data labelling and conversion of image format

The finalized dataset consists of 5169 images. The images were labelled properly using LabelImg annotation tool by drawing a bounding box on the face from forehead and chin. The annotations were saved in YOLO format and written to a text file (.txt). The images are consisted of various format including .jpg, .jpeg and .png. As the model only support standardized image format, the images with .jpeg and .png format were standardized into .jpg format. Finally, the data were split into 85% training set (4393 images) and 15% testing set (776 images) from the finalized dataset. The training set was used during training the model to fit the parameters of the model, where the model will learn from the training data. The testing set was used to test the trained model, where the model will perform its detection on unseen images, and the performance will be evaluated based on the face mask detection of the testing set.

Table 1. Summary of dataset used in the project

No	Dataset	Total Images	Description
1.	Face Mask Detection [34]	853	This Kaggle dataset consists of 3 classes which are with mask, without mask and mask worn incorrectly. The images consist of single person as well as multiple people. Images from classes with mask and without mask were used for this project.
2.	Medical Mask [35]	6024	This Kaggle dataset consists of 2 classes with 4000+ masked images and 1500+ unmasked images.
3.	Face Mask Dataset [36]	1840	This Kaggle dataset consists of 2 classes which are mask and without mask images. The images were obtained from Google, Bing and other Kaggle datasets.
4.	Face Mask Detection Dataset [37]	7553	This Kaggle dataset consists of 2 classes where 3725 images are with masked and 3828 images are without masks. 1776 images are obtained from Prajna Bhandary's Github account and remaining 5777 images are collected and filtered from Google.
5.	COVID Face Mask Detection Dataset [38]	1006	This Kaggle dataset consists of 2 classes where 503 images are with mask and remaining 503 images are without mask.
6.	Correctly Masked Face Dataset (CMFD) [39]	142	This dataset consists of images with mask only. This dataset is taken from cabani's Github account.

Figure 2. Sample images that are selected in the final dataset

2.3. CNN models development and configuration

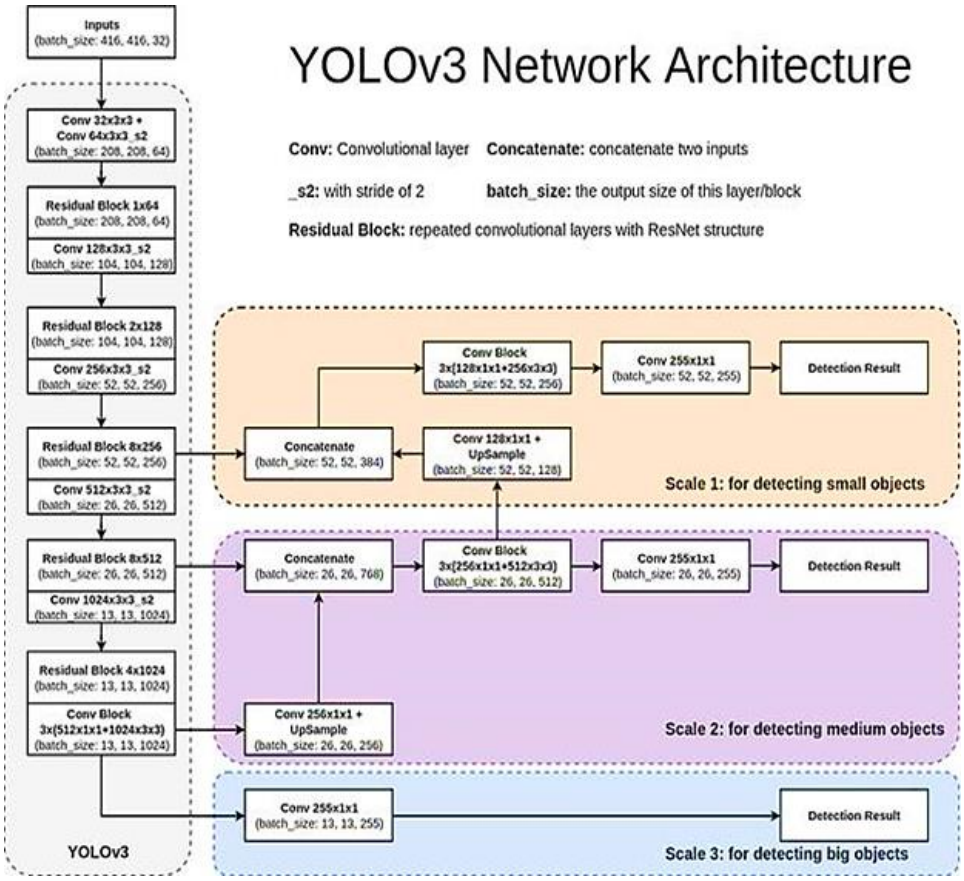
In this project, pretrained ResNet-50 as shown in Figure 3(a) and Darknet-53 model as shown in Figure 3(b) were used where both models were trained on ImageNet dataset with 1000 classes. To concatenate with YOLOv3 as shown in Figure 3(c), the feature extraction layers excluding the average pool, fully connected layer and softmax function from ResNet-50 and Darknet-53 were extracted and saved as separate weight files as feature extractor. The feature extractor models were concatenated with YOLOv3, resulting in ResNet50_YOLOv3 and DarkNet53_YOLOv3 object detection model. Before the training, the models were required to configured accordingly. In YOLOv3, the input image size must be in the multiples of 32 to be able to be downsampled in the detection layers. Thus, in this project, the input images were set at 416×416. The iterations which are the 'max_batches' were set at 4000 as a default where each class which are 'mask' and 'nomask' has to be trained with a minimum of 2000 iterations. As it is impractical to pass all the images from the dataset into training at one iteration, the images were divided into batches. In this case, the batch size was set into 64 where 64 images were fed into training at one iteration. To avoid memory error of the GPU, the subdivision was set to 16 where the GPU will process 4 images at a time as a minibatch. For the training, the learning rate was set at 0.001 at the beginning, decreased to 0.0001 starting at 3200^{th} iteration and 0.00001 starting at 3600^{th} iteration. This allows the model to learn faster at the beginning for gradient descent and proceeds to fine-tuning towards the end of the learning process [40]. The class was set to 2 and the filters number before the detection layer was set according to [(class + 5) * 3] which is 21 in this project. The models were trained with Darknet framework which is an open-source neural network framework written in C and compute unified device architecture (CUDA) and it supports CPU and GPU computation [41]. Overall, the development of the models was done on Google Colab Pro using Tesla P100-PCIE-16GB GPU.

layer name	output size	50-layer
conv1	112 x 112	7 x 7, 64, stride 2
		3 x 3 max pool, stride 2
con2_x	56 x 56	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$
conv3_x	28 x 28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$
conv4_x	14 x 14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$
conv5_x	7 x 7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$
	1 x 1	average pool, 1000-d fc, softmax

(a)

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

(b)



(c)

Figure 3. The model architecture of (a) ResNet-50, (b) Darknet-53, and (c) YOLOv3

2.4. Model performance evaluation

In object detection, the Intersection over union (IoU) is used to measure the performance of the training. It is used as a metric to evaluate how close the prediction bounding box is to the ground truth (the bounding box labelled manually). IoU is defined as the intersection area of predicted region and ground-truth region divided by the union of the two regions where we can determine the similarity among the two regions [42]. During the training, the model will learn and evaluate itself based on the IoU to improve its prediction as close as the ground truth. In this project, the IoU threshold was set at 0.5.

To evaluate the performance for the model, precision, recall and mean average precision (mAP) were calculated. The IoU threshold was set at 0.5 in this study to determine true positive (TP), false positive

(FP) and false negative (FN). If the prediction has an IoU larger than threshold, it will be categorized under TP which means the correct prediction. A prediction will be categorized as FP when the IoU score is lesser than threshold. Another 2 scenarios that will be categorized under FN is either when there is no detection at all (no bounding box) or the classification for the object is wrong. Therefore, using the number of TP, FP and FN, precision and recall can be calculated. Precision is the percentage of correct positives (TP) among all the predictions are made and can be used to determine the reliability of the model in classifying samples as positive [43]. Recall is the percentage of correct positive predictions among all the positives in reality (ground truth) [44]. A high recall indicates that the model is reliable in detecting positive samples. With precision and recall, a precision-recall graph can be plotted and the area under the graph can be calculated to determine the average precision of the class. A mAP is the mean of average precision between two classes which are masked and non-masked people in this project [45].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

Besides measuring the mAP of each models, the detection time was measured as one of the criteria in evaluating a performance of the object detection model. The detection time was determined based on the time needed for the model to perform all the predictions in the testing set images. By evaluating the speed and accuracy of each model, the model with better performance was determined.

3. RESULTS AND DISCUSSION

3.1. Performance evaluation

The models were tested on testing dataset with 776 images and the results were shown in the Table 2. ResNet50_YOLOv3 model obtained higher precision at 0.99 as compared to DarkNet53_YOLO that obtained precision at 0.96. This shows that the percentage of correct detections among all detections are made is higher in ResNet50_YOLOv3 model, indicating that ResNet50_YOLOv3 model had made less false positives in the prediction. However, ResNet50_YOLOv3 model had shown a significant lower recall compared to DarkNet53_YOLOv3 model. DarkNet53_YOLOv3 model achieved a recall of 0.96 while ResNet50_YOLOv3 achieved a recall with merely 0.58. A higher recall indicates that the DarkNet50_YOLOv3 model has better ability in identifying and detecting the objects in the image [46], while ResNet50_YOLOv3 model tends to have miss detections on the objects. Based on the results of average IoU, ResNet50_YOLOv3 shows better results with average IoU of 83.09% over DarkNet53_YOLOv3 model. This indicates that ResNet_YOLOv3 model performs better in localization, where the predicted bounding box are closer to the ground truth than DarkNet53_YOLOv3 model. On the other hand, the average precision of both classes which are mask, and no mask are higher in DarkNet53_YOLOv3, resulting in a higher mAP at 95.94% for the detections of mask and no mask. Meanwhile, the average precision of class no mask in ResNet50_YOLOv3 model has significant lower results at 78.95%, resulting in lower mAP at 84.40%.

Table 2. Performance evaluation comparison for both models

Performance Evaluation	DarkNet53_YOLOv3_4000	ResNet50_YOLOv3_4000
Precision	0.93	0.99
Recall	0.96	0.58
Average IoU	78.65%	83.09%
AP no mask	93.15%	78.95%
AP mask	98.74%	89.95%
mAP@0.5	95.94%	84.40%

Figure 4 summarizes the time needed for each model to perform face and face mask detection. The detection time for each model was determined based on the time taken needed for the model to run through all the 776 images from testing dataset and perform detection. DarkNet53_YOLOv3 shows an outstanding performance with a much shorter time of 50 seconds to complete the detection. On the other hand, ResNet50_YOLOv3 model required a longer time of 2 minutes to complete the detection for 776 images.

Based on the results, both models have their own strength from their performance. ResNet50_YOLOv3 model has better precision result and performed better in localization. However, the performance is poorer when detecting person without mask and results in lower mAP. Although

DarkNet53_YOLOv3 has lower precision, it has a much higher recall with stronger ability to detect person with or without mask and has a better mAP. Looking into the architecture of the feature extractor models, Darknet-53 has deeper and complex networks with over 40 million trainable parameters [47] that guarantee its accuracy comparing to Resnet-50 which has over 23 million trainable parameters [48]. Interestingly, although DarkNet-53 has deeper networks, it still runs much faster than ResNet-50 and gives a promising speed and accuracy performance. To sum up, for this application in face and face mask detection, despite the better localization of ResNet50_YOLOv3 model, it has compromised on its speed of detection. Therefore, DarkNet53_YOLOv3 model is better as it outperforms the ResNet50_YOLOv3 model in its speed of detection, ability of detection and higher accuracy, which is more suitable to be used in commercial camera in the future.

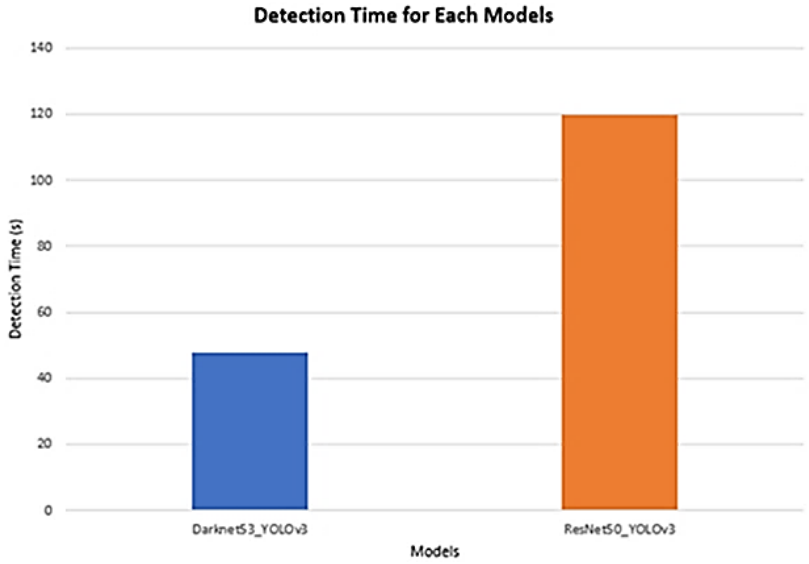
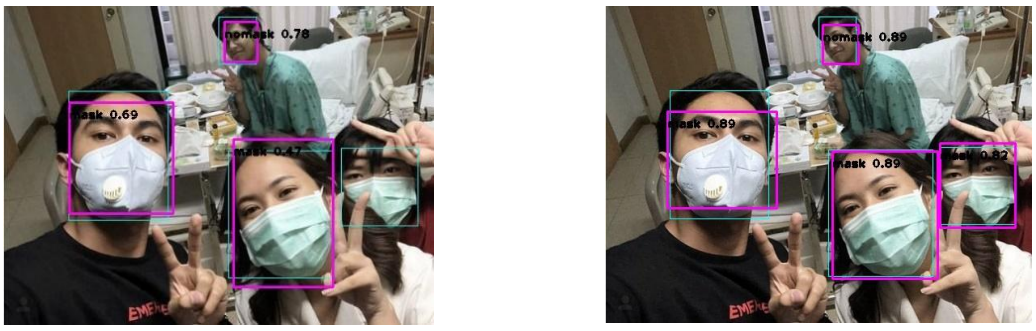


Figure 4. Detection time for both models

3.2. Face and face mask detection

This section shows the example of the output image from the detection of ResNet50_YOLOv3 as shown in Figure 5(a) and DarkNet53_YOLOv3 as shown in Figure 5(b) model. Green bounding box refers to the ground truth while *magenta* box refers to the predicted bounding box by the model. Overall, the models can classify the classes properly. ResNet50_YOLOv3 model tends to have lower confidence while predicting the objects, and there are some miss detections by the model.



(a) (b)

Figure 5. Sample output image detected by (a) ResNet50_YOLOv3 shows miss detection on the girl with mask while (b) DarkNet53_YOLOv3 shows perfect face and face mask detection

4. CONCLUSION

In summary, YOLOv3 based deep learning object detection was applied in this study to develop an automated face and face mask detection for multiple people that can be employed in commercial cameras in the future. The project started with data acquisition where the dataset was collected from online resources, followed by filtering and labelling of the images. The feature extraction layers from pretrained ResNet-50 and Darknet-53 were extracted and concatenated with YOLOv3 classifier and resulted in two models, ResNet50_YOLOv3 and DarkNet53_YOLOv3. Both models were trained with more than 4000 images on Google Colab Pro utilizing the darknet framework. The models were evaluated through precision, recall, mean average precision and detection time. Overall, DarkNet53_YOLOv3 model outperforms the ResNet50_YOLOv3 model, showing a well-balanced speed and accuracy performance, with a mean average precision of 95.94% and a much shorter time to perform the detection on the images. Besides, it also shows better ability to detect the faces on the image compared to ResNet50_YOLOv3 that has lower recall in the results. Thus, DarkNet53_YOLOv3 model was chosen as the better model as it is fast and reliable in detecting the faces with or without mask. This can be an initiative to help the authorities to monitor whether people are wearing face masks in public to minimize the transmission of COVID-19. The fast and accurate automated face mask detection is crucial during the pandemic to replace manual surveillance which is inefficient for detection of large groups of people, as well as providing a contactless monitoring method. This project can be improved by training on additional classes such as incorrectly worn mask where the nose or mouth are not covered by the mask, and the model can be integrated into webcam software or commercial camera for the development of prototype. The evaluation can be done based on frames per second on the live video and make improvement accordingly.

5. ACKNOWLEDGMENT

The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) for supporting this research and Ministry of Higher Education under Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/UTM/02/9)

REFERENCES

[1] World Health Organisation, "WHO Coronavirus Disease Dashboard," *Who.Int*, p. 1, 2022, [Online]. Available: https://covid19.who.int/.%0Ahttps://covid19.who.int/.

[2] World Health Organization, "Infection Prevention and Control during Health Care when Novel Coronavirus (nCoV) infection is suspected," *9789240000919*, vol. 38, no. 1, pp. 71–86, 2020, [Online]. Available: https://www.who.int/publications/i/item/10665-331495.

[3] WHO.c, "Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations," *Geneva: World Health Organization;*, vol. Available, no. March, pp. 1–10, 2020, [Online]. Available: https://www.who.int/publications-detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations.

[4] S. Ram, S. Gupta, and B. Agarwal, "Devanagri character recognition model using deep convolution neural network," *Journal of Statistics and Management Systems*, vol. 21, no. 4, pp. 593–599, 2018, doi: 10.1080/09720510.2018.1471264.

[5] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of Deep Learning for Object Detection," *Procedia Computer Science*, vol. 132, pp. 1706–1717, 2018, doi: 10.1016/j.procs.2018.05.144.

[6] D. Ravi *et al*., "Deep Learning for Health Informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017, doi: 10.1109/JBHI.2016.2636665.

[7] A. Jebelli and R. Ahmad, "Efficient Commercial Classification of Agricultural Products using Convolutional Neural Networks," *IAES International Journal of Robotics and Automation (IJRA)*, vol. 10, no. 4, p. 353, 2021, doi: 10.11591/ijra.v10i4.pp353-364.

[8] S. I. Prottasha and S. M. S. Reza, "A classification model based on depthwise separable convolutional neural network to identify rice plant diseases," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 4, pp. 3642–3654, 2022, doi: 10.11591/ijece.v12i4.pp3642-3654.

[9] R. A. Pratiwi, S. Nurmaini, D. P. Rini, M. N. Rachmatullah, and A. Darmawahyuni, "Deep ensemble learning for skin lesions classification with convolutional neural network," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 563–570, 2021, doi: 10.11591/ijai.v10.i3.pp563-570.

[10] O. A. Fagbuagun, O. Nwankwo, S. A. Akinpelu, and O. Folorunsho, "Model development for pneumonia detection from chest radiograph using transfer learning," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 3, p. 544, 2022, doi: 10.12928/telkomnika.v20i3.23296.

[11] I. S. Masad, A. Alqudah, A. M. Alqudah, and S. Almashaqbeh, "A hybrid deep learning approach towards building an intelligent system for pneumonia detection in chest x-ray images," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5530–5540, 2021, doi: 10.11591/ijece.v11i6.pp5530-5540.

[12] H. Zanddizari, N. Nguyen, B. Zeinali, and J. M. Chang, "A new preprocessing approach to improve the performance of CNN-based skin lesion classification," *Medical and Biological Engineering and Computing*, vol. 59, no. 5, pp. 1123–1131, 2021, doi: 10.1007/s11517-021-02355-5.

[13] A. F. Y. Althabhawee and B. K. O. C. Alwawi, "Fingerprint recognition based on collected images using deep learning technology," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 81–88, 2022, doi: 10.11591/ijai.v11.i1.pp81-88.

[14] A. Kherraki and R. El Ouazzani, "Deep convolutional neural networks architecture for an efficient emergency vehicle classification in real-time traffic monitoring," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 110–120, 2022, doi: 10.11591/ijai.v11.i1.pp110-120.

[15] M. A. Uthaib and M. S. Croock, "Multiclassification of license plate based on deep convolution neural networks," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5266–5276, 2021, doi: 10.11591/ijece.v11i6.pp5266-5276.
[16] W. Chen, Q. Sun, J. Wang, J. J. Dong, and C. Xu, "A Novel Model Based on AdaBoost and Deep CNN for Vehicle Classification," *IEEE Access*, vol. 6, pp. 60445–60455, 2018, doi: 10.1109/ACCESS.2018.2875525.
[17] J. Du, "Understanding of Object Detection Based on CNN Family and YOLO," *Journal of Physics: Conference Series*, vol. 1004, no. 1, 2018, doi: 10.1088/1742-6596/1004/1/012029.
[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014, doi: 10.1109/CVPR.2014.81.
[19] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.
[22] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.
[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.
[24] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, [Online]. Available: http://arxiv.org/abs/1804.02767.
[25] N. D. Nguyen, T. Do, T. D. Ngo, and D. D. Le, "An Evaluation of Deep Learning Methods for Small Object Detection," *Journal of Electrical and Computer Engineering*, vol. 2020, 2020, doi: 10.1155/2020/3189691.
[26] S. V. Militante and N. V. Dionisio, "Deep Learning Implementation of Facemask and Physical Distancing Detection with Alarm Systems," *Proceeding - 2020 3rd International Conference on Vocational Education and Electrical Engineering: Strengthening the framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering, ICVEE 2020*, 2020, doi: 10.1109/ICVEE50212.2020.9243183.
[27] G. Jignesh Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face Mask Detection Using Transfer Learning of InceptionV3," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12581 LNCS, pp. 81–90, 2020, doi: 10.1007/978-3-030-66665-1_6.
[28] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement: Journal of the International Measurement Confederation*, vol. 167, 2021, doi: 10.1016/j.measurement.2020.108288.
[29] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, and T. Das, "MOXA: A Deep Learning Based Unmanned Approach For Real-Time Monitoring of People Wearing Medical Masks," *Transactions of the Indian National Academy of Engineering*, vol. 5, no. 3, pp. 509–518, 2020, doi: 10.1007/s41403-020-00157-z.
[30] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection," *Sustainable Cities and Society*, vol. 65, 2021, doi: 10.1016/j.scs.2020.102600.
[31] M. R. Bhuiyan, S. A. Khushbu, and M. S. Islam, "A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3," *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225384.
[32] X. Ren and X. Liu, "Mask wearing detection based on YOLOv3," *Journal of Physics: Conference Series*, vol. 1678, no. 1, 2020, doi: 10.1088/1742-6596/1678/1/012089.
[33] M. S. M. Suhaimin, M. H. A. Hijazi, C. S. Kheau, and C. K. On, "Real-time mask detection and face recognition using eigenfaces and local binary pattern histogram for attendance system," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 1105–1113, 2021, doi: 10.11591/EEI.V10I2.2859.
[34] P. Kataria, "Face Mask Detection," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 06, no. 05, 2022, doi: 10.55041/ijsrem12757.
[35] W. Intelligence, "Face Mask Detection Dataset," 2020, [Online]. Available: https://www.kaggle.com/wobotintelligence/face-mask-detection-dataset.
[36] A. Purohit, "Face Mask Dataset (YOLO format)," 2022, [Online]. Available: https://www.kaggle.com/aditya276/face-mask-dataset-yolo-format.
[37] O. GURAV, "Face Mask Detection Dataset," 2020, [Online]. Available: https://www.kaggle.com/omkargurav/face-mask-dataset.
[38] Prithwiraj Mitra, "COVID Face Mask Detection Dataset," 2020, [Online]. Available: https://www.kaggle.com/datasets/prithwirajmitra/covid-face-mask-detection-dataset?select=New+Masks+Dataset.
[39] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "MaskedFace-Net – A dataset of correctly/incorrectly masked face images in the context of COVID-19," *Smart Health*, vol. 19, 2021, doi: 10.1016/j.smhl.2020.100144.
[40] I. Goodfellow, Y. Bengjo, and A. Courville, "Deep Learning (Adaptive Computation and Machine Learning series) Illustrated Edition," *Cambridge Massachusetts*, p. 429, 2016.
[41] AlexeyAB, "Github," 2020, [Online]. Available: https://github.com/AlexeyAB/darknet#how-to-train-to-detect-your-custom-objects.
[42] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10072 LNCS, pp. 234–244, 2016, doi: 10.1007/978-3-319-50835-1_22.
[43] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, [Online]. Available: http://arxiv.org/abs/2010.16061.
[44] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
[45] R. Jie Tan, "Breaking Down Mean Average Precision (mAP)," *Towards Data Science*, 2019, [Online]. Available: https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52.

[46] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
[47] Z. Li, L. Zhao, X. Han, and M. Pan, "Lightweight Ship Detection Methods Based on YOLOv3 and DenseNet," *Mathematical Problems in Engineering*, vol. 2020, 2020, doi: 10.1155/2020/4813183.
[48] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, "IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture," *Computer Networks*, vol. 171, 2020, doi: 10.1016/j.comnet.2020.107138.

BIOGRAPHIES OF AUTHORS

Cheng Xiao Ge received a Bachelor Of Engineering (B.Eng) in Biomedical Engineering from Universiti Teknologi Malaysia (UTM). She is currently working as RnD Engineer in Bbraund Malaysia that focused on designing the intravenous catheter. Her research interest is in Deep Learning and Machine Learning. She can be contacted at email: xiao.ge@graduate.utm.my

Muhammad Amir As'ari holds a PhD in Biomedical Engineering from the Universiti Teknologi Malaysia. His PhD's work was in the field of Assistive Technology, Computer Vision and Image Procesing and his work focused on developing a novel 3D shape descriptor for recognizing the activities of daily living (ADLs) based on Kinect-like depth image. Amir pursued his master's degree and bachelor's degree at the Faculty of Electrical Eng, Universiti Teknologi Malaysia, majored in Electronic Engineering. His master's degree and bachelor's degree projects were also related to computer vision and image processing for security and surveillance. Currently, he is working on automated human action recognition based on wearable sensor and context-aware modality for assistive technology and sport technology. He can be contacted at email: amir-asari@utm.my

Nur Anis Jasmin Sufri received a Bachelor of Engineering (B.Eng.) in Biomedical Engineering from Universiti Teknologi Malaysia (UTM). She is currently a PhD student in Biomedical Engineering under School of Biomedical Engineering and Health Sciences, Faculty of Engineering, UTM. Her research areas of interest include Artificial Intelligent specifically in Deep learning, Image Processing, Computer Vision, Assistive Technology, Banknote Detection and Classification. She can be contacted at email: najasmin2@live.utm.my