Single shot multibox detector for detecting and classifying masked faces in digital images

Indah Agustien Siradjuddin[1], *Sulaiman*[1*], and *Kurniawan* Eka Permana[1]

[1]Informatics Department, Faculty of Engineering, Universitas Trunodjoyo Madura, Bangkalan, Madura, Indonesia

Abstract. This paper presents a method for detecting and classifying masked faces in digital images using the Single Shot Multibox Detector (SSD). The location of faces in an image and their category, indicating whether the detected face is wearing a medical mask, wearing it improperly, or not wearing one, are found based on the proposed method. This detection is used for security to find the location of faces in a certain category. SSD is a single-stage detector that utilizes various boxes of different sizes in its architecture to detect the location of the target in the image and also the category of the object within the boxes. The various boxes are used to detect different face sizes in the image. The experiments are conducted on three datasets, they are AFLW (Annotated Facial Landmarks in the Wild) with 1865 images, MAFA (Masked Face Analysis) with 3616 images, and MakeML with 853 images. The mAP (Mean Average Precision) of the method in the detection is 41.52%.

1 Introduction

The rapid spread of the infectious disease in 2019, specifically Covid-19, led to the policy of wearing masks in public places to reduce the number of Covid-19 infections among the population. Nowadays, even though Covid-19 has ended, the habit of wearing masks in public places has become more general, as it will prevent the spread of other viruses, such as flu viruses. The habit of wearing a mask is sometimes misused for criminal activities, as the criminal cannot be identified by CCTV since their faces are almost always covered by the mask.

Therefore, monitoring mask usage is still important and essential for ensuring public safety, especially in crowded areas, such as hospitals, malls, offices, and others [1]. However, manual monitoring is time consuming, therefore, automatic detection to detect the faces with mask, without mask, and improper way of wearing mask, is required, and this detection become an important research topic [2].

Face mask detection is now a challenging task, since it is not only to detect the location of the faces that using mask, but also to identify the category of mask-wearing conditions, i.e. whether the face using a mask properly (cover the mouth and half part of nose), without using a mask, and improper way of wearing a mask (not covering the whole mouth area). Several deep learning-based object detection algorithms such as R-CNN[3], [4], YOLO [5], Retina Net, and Single Short Multibox Detector, have been widely used for general object detection, and among them SSD is known for its efficiency, that perform the detection (localization and classification) work simultaneously, making it suitable for real-time scenarios [6], [7].

Despite the increasing number of mask face detection methods, many research studies focus only on binary classification, determining whether the detected objects wear a mask or not [8], our proposed model detects multiclass category.

Therefore, the research contribution in this article consists of two parts. First, in addition to the binary classification of faces within the bounding box (whether the face wears a mask or not), we also consider the possibility of faces using a mask improperly (not covering the entire mouth area). Second, our proposed model is trained using a combined dataset from AFLW, MAFA, and MakeML to increase the dataset size, allowing the model to learn the target pattern from a huge number of samples.

2 Single shot multibox detector

Single Shot Multibox Detector (SSD) is one of the detection algorithms that is based on the One-stage detection, which the region proposal and the detection stages are performed within the same network architecture [6]. This occurs since the target of the network contains information not only the candidate location of the detected object, but also the class of the detected objects.

2.1 SSD architecture

Since the target of SSD contains the complete information of the objects (candidate locations and the class) and cannot be represented with one dimensional vector, therefore in SSD the layers of the fully connected layers are not required, there are only convolution layers in this architecture, as shown in Fig. 1.

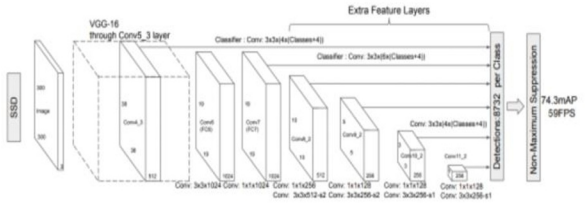


Fig. 1. SSD Architecture [6].

There are six type of feature map's size, they are, 38 × 38, 19 × 19, 10 × 10, 5 × 5, 3 × 3, and 1 × 1. Feature maps with size 38 × 38, 3 × 3, and 1 × 1, have four bounding boxes in each cell. Therefore, there are in total 5776 bounding boxes for 38 × 38, and 36 bounding boxes for 3 × 3 feature map, and four boxes for 1 × 1 feature map. Meanwhile for the feature maps with size 19 × 19, 10 × 10, and 5 × 5 has six bounding boxes for each cell. Therefore, there are 2166, 600, and 150 bounding boxes for each type of feature map's size.

In an SSD architecture, there is a base network before the main convolution layers, i.e., VGG-16, as seen in Fig. 1. However, only convolution layers of VGG-16 are used in SSD as seen in Fig. 2.

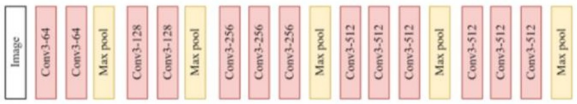


Fig. 2. VGG-16 architecture for the SSD backbone.

2.2 Default Box

One-stage detection approach means that the approach provides a mechanism to produce the region proposal to indicate the candidate location of the target object. SSD has the default box for the region proposal. The default box will be the anchor that will scan the feature maps to produce the candidate location of the object. There are several default boxes sizes based on its scales, as stated in (1) and (2). All default boxes have the same center coordinate as written in (3).

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1) \quad (1)$$

$$w_k^a = {}^{s_k}/_{\sqrt{a^r}} \; ; \; h_k^a = {}^{s_k}/_{\sqrt{a^r}} \; ; \; s_k' = \sqrt{s_k s_{k+1}} \quad (2)$$

$$center = \left(\frac{x + 0.5}{|f_k|}, \frac{y + 0.5}{|f_k|}\right) \quad (3)$$

where:
$s_{max} = 0.9$; $s_{min} = 0.1$; s_k: $scale\ of\ feature\ map$
a_r: $aspect\ ratio$, {1,2,3,1/2,1/3}; s_k': $scale\ for\ a_r = 1$
w_k^a, h_k^a : $width\ and\ height\ of\ default\ box\ each\ a_r$
$k = 1..m$, and m is the number of types of feature map's size; x, y : grid cell's coordinate, and $|f_k|$ is the feature map's size

The size of created default boxes based on the formula is shown in Fig. 3, and the example of the default boxes in each feature map's size is depicted in Table 1. As seen in the table, that the different size of feature maps makes the default boxes detect different size of the target objects. For bigger feature map's size, the default boxes detect the small objects in the image, meanwhile, the smaller size of feature map's size, the default boxes will have the responsible to detect the larger size of the objects in the image.

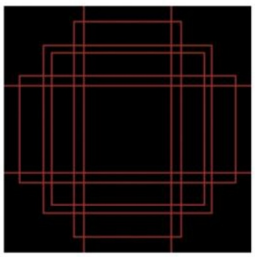
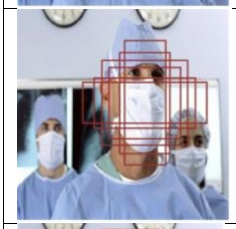
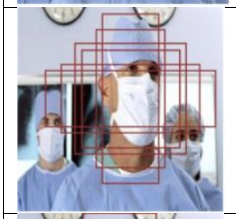
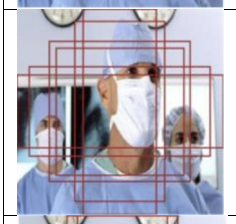
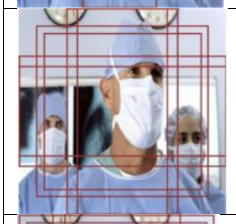
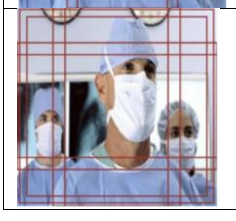


Fig. 3. Default boxes.

Table 1. Default boxes in each feature map's size.

Default Boxes	Feature Map's Size
	38 × 38
	19 × 19
	10 × 10
	5 × 5
	3 × 3
	1 × 1

2.3 Encode Bounding Box

During the training phase of SSD, the target consists of candidate object locations. These candidate locations are defined as the offset or delta between the ground truth and the determined bounding boxes or known as encode the bounding boxes. Since in this research, the detection target consists of three classes (categories), i.e., face with a medical mask, face without a medical mask, and a face with improper use of the medical mask, therebefore, the encode bounding box is represented with $(c_1, c_2, c_3, c_4, \widehat{g^{cx}}, \widehat{g^{cy}}, \widehat{g^{w}}, \widehat{g^{h}},)$ where $c_1 - c_4$ is the classes (included with the background class). The offset between groundtruth and the default boxes is written in (4) and (5).

$$\widehat{g^{cx}} = (g^{cx} - d^{cx})/d^{w} \; ; \; \widehat{g^{cy}} = (g^{cy} - d^{cy})/d^{h} \tag{3}$$

$$\widehat{g^{w}} = ln\left(\frac{g^{w}}{d^{w}}\right) \; ; \; \widehat{g^{h}} = ln\left(\frac{g^{h}}{d^{h}}\right) \tag{4}$$

Where $g^{cx}, g^{cy}, g^{w}, g^{h}$ is the center x coordinate, center y coordinate, width, and height of the groundtruth, respectively. $D^{cx}, d^{cy}, d^{w}, and\ d^{h}$ is the center x coordinate, center y coordinate, width, and height of the default box.

3 Masked face detection

In this article, SSD network is used for the mask face detection and classification, to detect the location of the target object, and define the category of the detected object, whether it is a face with a medical mask, a face without any medical mask, and a face with improper way of wearing mask. The diagram for the masked face detection can be seen in Fig. 4.

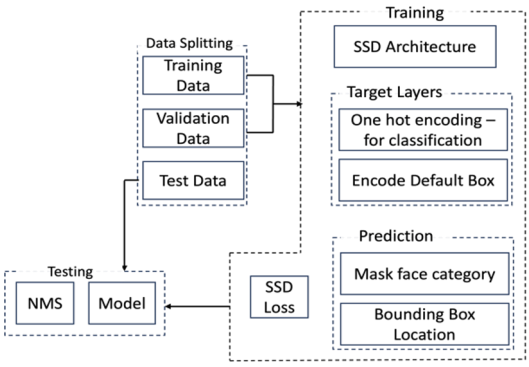


Fig. 4. SSD for the mask face detection.

As seen in the figure, the final stage of the detection is the NMS (Non-Maximum Suppression). The objective of the NMS stage is to eliminate overlaps bounding box based on its confidence value. The example of NMS stage is shown in Fig. 5.

The evaluation of the detection is measured by the mAP (mean Average Precision) metrics. To calculate the mAP, True Positive (TP), False Positive (FP), and False Negative (FN) are defined previously. In this research, we used the threshold of groundtruth and the predicted bounding boxes is 0.45. The example of TP, FP, FN are depicted in Fig. 6.

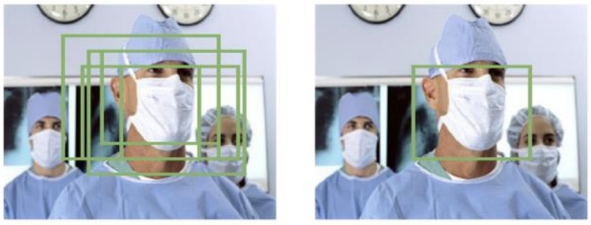


Fig. 5. Non-Maximum Surpression. (Left) predicted bounding boxes from the SSD Model. (Right) is the remain bounding box based on the NMS algorithm.

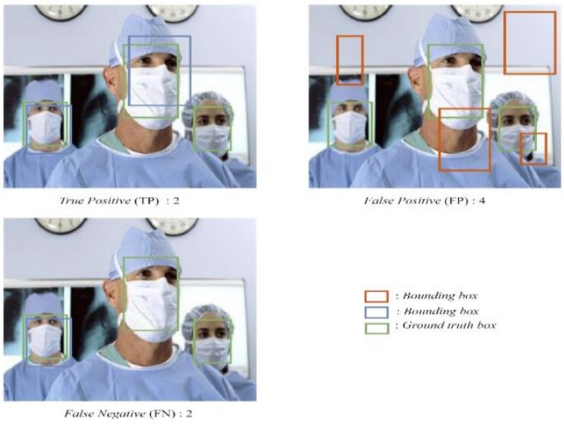


Fig. 6. TP, FP, FN of mask face category.

The precision, recall, and Average Precision for mAP are written in (6), (7), and (8).

$$Precision = \frac{TP}{TP + FP} \; ; \; recall = \frac{TP}{TP + FN} \tag{6}$$

$$AP = \frac{1}{11}\sum_{r\in\{0,0.1,0.2,I,1\}} \rho_{interpI} \tag{7}$$

$$\rho_{interpI} = max_{\tilde{r}:\tilde{r}\geq r} precisionI \tag{8}$$

4 Result and discussion

The dataset used for the experiments was collected from three datasets, i.e., AFLW (Annotated Facial Landmarks in the Wild) with 1865 images [9], MAFA (Masked Face Analysis) with 3616 images [10], and MakeML[11] with 853 images. Therefore, the total number of images for training and testing in the experiments was 6334, of which 2717 were faces without a mask, 5332 with a medical mask, and 2123 with an improper mask. The experimental setup is shown in Table 2.

Table 2. Experimental setup.

Name	Information
Image Size	300 ×300
Learning Rate	10^{-4}
Optimizer	Adam
Training Data : Test Data	80 : 20
Epochs	100
Batch Size	64

The architecture of the proposed SSD was built using the pre-trained VGG-16 as a base network. The

whole layers of the implemented SSD is shown in Table 3.

Table 3. SSD network layers.

Layers	Output Shape
Base Network Layers (VGG-16)	19 × 19 × 512
Convolution Layer, kernel 3 × 3, stride 1	19 × 19 × 1024
Convolution Layer, kernel 1 × 1, stride 1	19 × 19 × 1024
Convolution Layer, kernel 1 × 1, stride 1	19 × 19 × 256
Convolution Layer, kernel 3 × 3, stride 2	10 × 10 × 512
Convolution Layer, kernel 1 × 1, stride 1	10 × 10 × 128
Convolution Layer, kernel 3 × 3, stride 2	5 × 5 × 256
Convolution Layer, kernel 1 × 1, stride 1	5 × 5 × 128
Convolution Layer, kernel 1 × 1, stride 1, without padding	3 × 3 × 256
Convolution Layer, kernel 1 × 1, stride 1,	3 × 3 × 128
Convolution Layer, kernel 3 × 3, stride 1, without padding	1 × 1 × 256

For the training, we split the dataset into 5-fold cross validation. The use of cross-validation further supports the model's reliability. The decreasing error loss and validation in each fold, is shown in Fig. 7.

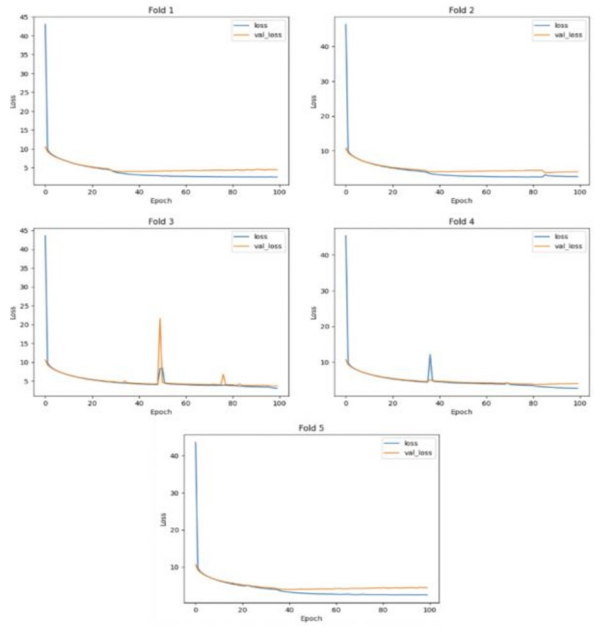


Fig. 7. Training and validation loss during training stage.

As seen in Fig. 7, the loss of training and validation data is decreasing.

The performance of the detection model is shown in Table 4.

Table 4. mAP of the detection model.

Fold	AP (%)			mAP (%)
	No-Mask	Mask	Improper-Mask	
1	42.07	31.67	36.33	36.89
2	44.71	31.79	40.66	39.05
3	36.82	26.4	33.68	32.3
4	46.69	36.22	41,65	41.52
5	42.04	31.77	33.00	35.60

The example of the detection result is shown in Fig. 8. As seen in the figure, the detector can detect well if the number of faces in the image is one or two, however, if the number of faces is more than one, then the detector fails to detect the target object. We also detect the real image (image that is not in the dataset), and the result is shown in Fig. 9. The result shows that SSD can perform multiclass mask detection with reasonable accuracy, however, the performance degrades when multiple faces appear simultaneously in one image.

Fig. 8. Detection result.

Green box means the face is wearing mask, yellow box is the face with improper mask (only covering half of mouth area), and red box is the face without any mask

Fig. 9. Detection result of images that are not in the dataset.

5 Conclusion

Single Shot Multibox Detector is implemented for mask face detection. In this detection, SSD will give the location of the target object and the category of detected object, whether the face is using a mask, without a mask, and using a mask but improper way. The combined dataset from three dataset that open for public is used to build the detection model. The result shows that the model detect the face objects in the image and distinguish the category of the usage of mask on each face. However, the detector mostly failed in detection when there are many target objects in the image. For the future research, preprocessed is required to create a good quality input for the model, and tuning the hyperparameter of SSD, such as the epochs, learning rate, and others.

The authors would like to thank the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) of Universitas

Trunodjoyo Madura for supporting this research under International Collaboration Grant.

References

1. World Health Organization, Mask use in the context of COVID-19. WHO Interim Guidance. (2020). https://iris.who.int/server/api/core/bitstreams/158d176a-6de8-47be-8c9e-7bb1a3410cd2/content
2. S. Sethi, M. Kathuria, T. Kaushik, Face mask etection using deep learning: An approach to reduce risk of Coronavirus spread. J Biomed Inform. **120** (2021). https://doi.org/10.1016/j.jbi.2021.103848
3. O. Hmidani, E. M. Ismaili Alaoui, A comprehensive survey of the R-CNN family for object detection. 5th International Conference on Advanced Communication Technologies and Networking (CommNet). 1–6 (2022). https://doi.org/10.1109/CommNet56067.2022.9993862
4. I. A. Siradjuddin, Reynaldi, A. Muntasa, Faster Region-based Convolutional Neural Network for Mask Face Detection. 5th International Conference on Informatics and Computational Sciences (ICICoS). 282–286 (2021). https://doi.org/10.1109/ICICoS53627.2021.9651744
5. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767. (2018). http://arxiv.org/abs/1804.02767
6. W. Liu *et al.*, SSD: Single Shot MultiBox Detector. European conference on computer vision. 21-37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
7. P. Mounika, C. L. Narayana, D. Kondapalli, V. Ramana, A Comparative Study of Yolo and Ssd Object Detection Algorithms. International Journal of Creative Research Thoughts – IJCRT. **9**, 11, b821-b827 (2021). https://ijcrt.org/papers/IJCRT2111212.pdf
8. W. Boulila, M. Afifi, A. Alzahem, I. Alturki, A. Almoudi, M. Driss, A Deep Learning-based Approach for Real-time Facemask Detection, 20th IEEE International Conference on Machine Learning and Applications (ICMLA). 1478-1481 (2021). https://doi.org/10.1109/ICMLA52953.2021.00238
9. M. Köstinger, P. Wohlhart, P. M. Roth, H. Bischof, Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. IEEE International Conference on Computer Vision Workshops (ICCV Workshops). 2144–2151 (2011). https://doi.org/10.1109/ICCVW.2011.6130513
10. S. Ge, J. Li, Q. Ye, Z. Luo, Detecting Masked Faces in the Wild with LLE-CNNs. Proceedings of the IEEE conference on computer vision and pattern recognition. 2682-2690 (2017). https://openaccess.thecvf.com/content_cvpr_2017/html/Ge_Detecting_Masked_Faces_CVPR_2017_paper.html
11. M. Pamarthi, S. L. Injam, Md, O. K. Z., & Lakshmi T. Surekha. A Deep Learning Framework for Social Distance Monitoring and Face Mask Detection." In IoT Based Control Networks and Intelligent Systems: Proceedings of 3rd ICICNIS, 613-620 (2022). https://doi.org/10.1007/978-981-19-5845-8_43