Real-Time Multi-view Face Mask Detector on Edge Device for Supporting Service Robots in the COVID-19 Pandemic

Muhamad Dwisnanto Putro(✉), Duy-Linh Nguyen(✉), and Kang-Hyun Jo(✉)

Department of Electrical, Electronics, and Computer Engineering,
University of Ulsan, Ulsan, Korea
{dputro,ndlinh301}@mail.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. The COVID-19 pandemic requires everyone to wear a face mask in public areas. This situation expands the ability of a service robot to have a masked face recognition system. The challenge is detecting multi-view faces. Previous works encountered this problem and tended to be slow when implemented in practical applications. This paper proposes a real-time multi-view face mask detector with two main modules: face detection and face mask classification. The proposed architecture emphasizes light and robust feature extraction. The two-stage network makes it easy to focus on discriminating features on the facial area. The detector filters non-faces at the face detection stage and then classifies the facial regions into two categories. Both models were trained and tested on the benchmark datasets. As a result, the proposed detector obtains high performance with competitive accuracy from competitors. It can run 20.60 frames per second when working in real-time on Jetson Nano.

1 Introduction

The technology of robots is developing rapidly in the industrial and medical fields. The Industrial Revolution 5.0 supports to encourage the implementation of robots in the public area. Service robots are one type used by humans to help with daily activities [11]. This robot has human-like abilities that can walk, see, talk, and understand the environment. Since the emergence of COVID-19 spread like a pandemic globally, prevention of this virus is the first step to reduce its impact by wearing face masks. It is useful for protecting the transmission of the virus through droplets in the mouth, and nose area [4]. So everyone to be required to wear a mask when in a public environment. This situation recommends service robots can detect and classify face masks in public areas. It is useful for warning people who don't use it.

Several previous studies have succeeded in classifying face masks. Ejaz et al. used the Principal Component Analysis (PCA) to recognize the face masks [3]. This study uses statistical differences of accuracy to measure the performance

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 507–517, 2021.
https://doi.org/10.1007/978-3-030-73280-6_40

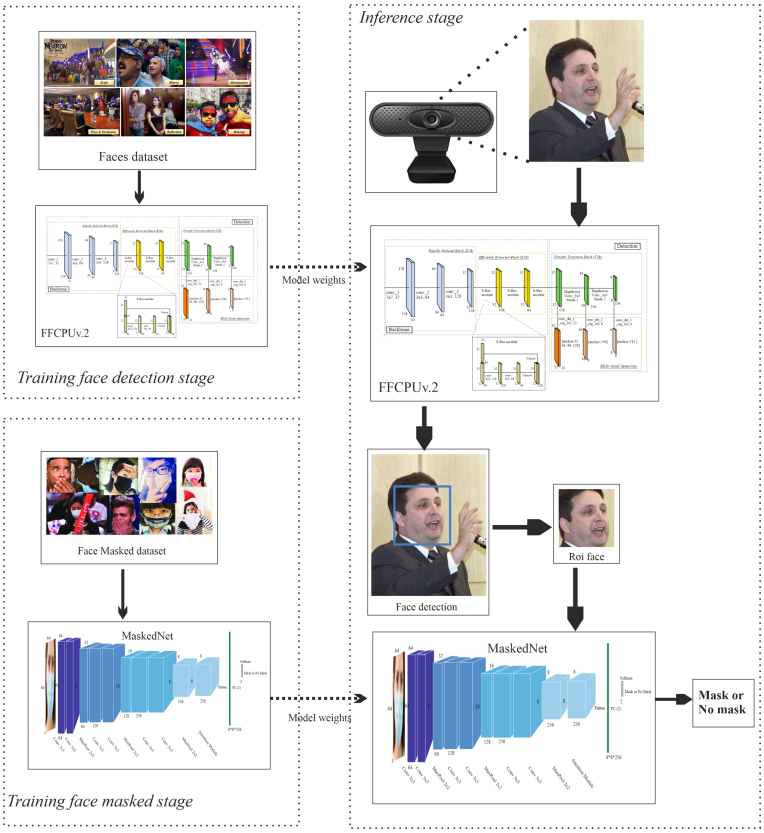

Fig. 1. Flowchart of overall face masks detector.

of the classification system. In addition, the face detection of Viola-Jones was employed for detecting the face region. Another work is applying the Gaussian Mixture Model (GMM) to build a faces model [1]. This system predicts face masks by analyzing and learning typical facial features in the eyes, nose, and mouth area. This model was tested on various face challenges using accessories such as respirators and sunglasses as an evaluation. However, both works are weak when classifying face masks on non-frontal faces. It is caused by the limitation of face detection and feature extraction.

Convolutional Neural Network (CNN) has been proven as a robust extractor feature [8]. In general, CNN architecture consists of feature extractor, and classifier [13]. This method can classify various images by updating the filter weights to produce an output that matched the ground truth [2]. However, this reliability is not supported by efficiency when run in real-time. Deep CNN tends to generate a large number of parameters and heavyweights [6]. Loey et al. used Resnet-50 as a backbone for feature extraction followed by decision trees, Support Vector Machine (SVM), and ensemble algorithm as the classifier [9]. The model produces high accuracy but is slow, while the practical application

requires an algorithm to work in real-time on a portable device. Jetson Nano is a mini-computer supported by a 128-core graphic accelerator. This edge device is recommended for use as a robot processor because it supports sensor acquisition and actuator control functions.

This paper aims to build real-time masked face recognition on edge devices implemented in service robots. The contributions of this paper are as follows:

1. Fast face detector was developed to detect multi-view faces and occlusion challenges (FFDMASK). The process helps to get the RoI (Region of Interest) from the face.
2. Slim CNN architecture to fast and accurately classifies the face masks (masks or none). The attention module is applied to improve the quality of shallow feature maps.

As a result, it achieves competitive performance with state-of-the-art algorithms on several datasets and can work in real-time on the Jetson Nano without lack.

2 Proposed Architecture

The proposed detector consists of two-stage, including face detector and classification module. Figure 1 shows the overall diagram of the proposed system. A face detector works to detect faces in an image to produce bounding boxes as the face area [14]. The crop technique is used in each box to generate facial RoI. Then a classification module is employed to extract information and classify masked faces.

2.1 Face Detector

Backbone Module. FFDMASK develops the architecture of FFCPU [10]. This network consists of two main parts, including the backbone and detection block, as shown in Fig. 2. This model functions as a feature extractor to produce a clear feature map on the detection module. Rapidly Reduced Block (R2B) emphasizes reducing the dimension of the feature map with a convolutional layer and an Efficiently Extracted Block (E2B) to separate facial and non-facial distinctive features effectively. FFDMASK employs the S-Res (Split-residual) module to upgrade the E2B. This block divides the feature input into two parts based on the number of channels, then employs a bottleneck convolution on the first part and passes to the end of the module for the other chunk. Three modules are used to increase the detector's performance, which outperforms other competitors.

Detection Module and Anchor. The detection module is responsible for predicting the facial area, which includes three layers. This module employs a depthwise convolution between layers. This block saves computing power and increases speed. Besides, the decrease in accuracy does not have a significant impact. Furthermore, the assignment of anchors with scale variations at each detection layer can adjust the bounding box based on the size of the feature map, large anchors for small feature maps, and vice versa. In order to optimize training process, it still uses balanced loss which refers to the FFCPU.

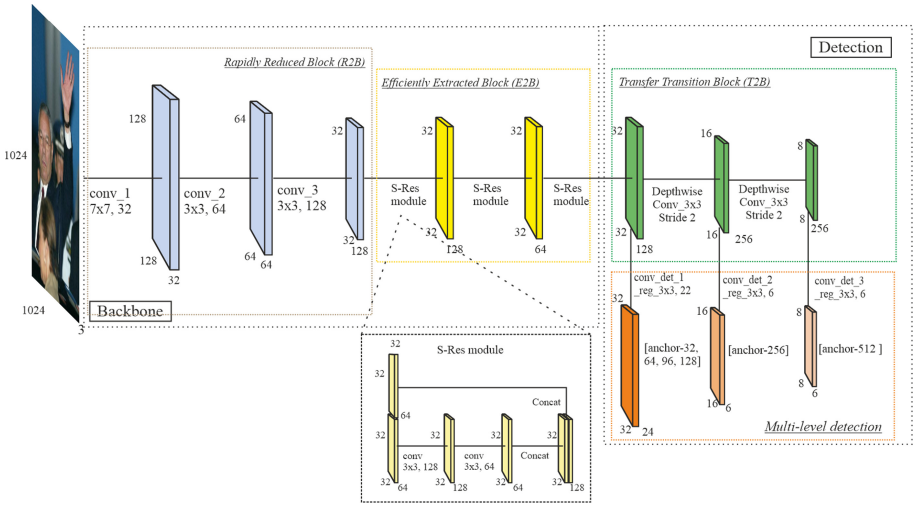

Fig. 2. The proposed architecture of face detector, including Rapidly Reduced Block (R2B), Efficiently Extracted Block (E2B), Transfer Transition Block (T2B), and Multi-level Detection. The training process requires 1024 × 1024 as the input image size.

2.2 Face Masks Classification Module

Baseline Module. The baseline module uses the convolutional neural network to extract distinctive features, and the max-pooling layers are used to shrink the feature map size. The ReLU and Batch Normalization are used to prevent saturation of the network. In order for the detector to work fast, this slim architecture emphasizes the shallow layers and narrow channels, as shown in Fig. 3. Each stage uses two 3 × 3 convolutions and a pool. This convolution has proven that it effectively separates distinctive features from the background (VGG) [12]. The input of this module is an RGB image with a size of 64 × 64. It produces 8 × 8 at the last of the feature map. Furthermore, this module employs an attention module to increase the discrimination power of facial features that are covered by masks and normal faces. Fully connected focuses on vectors with two categories and generates the final probability of predictions.

Attention Module. The shallow layer CNNs tend to produce low-level features. Even this architecture is underperforming when it discriminates against complex features [7]. This problem can be solved by employing the attention module to improve the quality of the feature map representing the global context of an input image. This technique can highlight the differences in facial features covered and without a mask. The position attention module is applied to capture context-based information and separate between interest and useless facial features [5].

The first step is to apply the 1 × 1 convolution as a simple buffer for the feature map output of the baseline module ($fm_{(b)}$), as illustrated by Fig. 4.

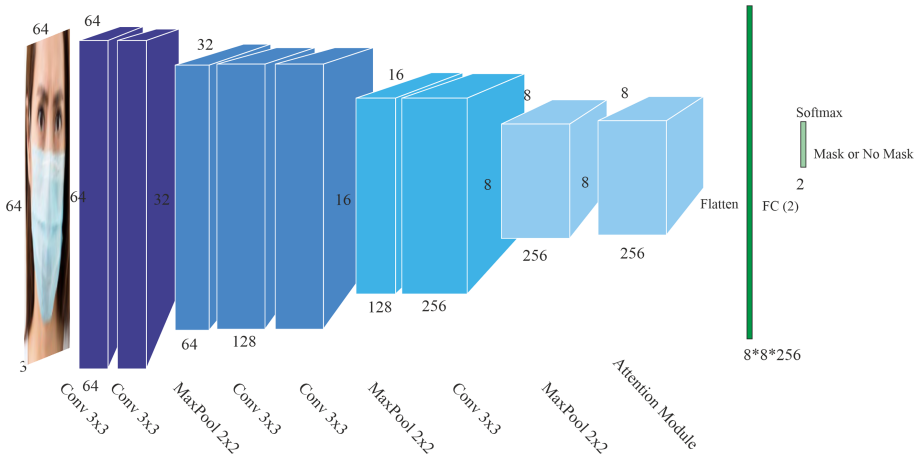

Fig. 3. The slim architecture of the face masks classification. This module consists of seven convolution layers, three max-pooling layers, an attention module, and a fully connected layer at the end of the network.

This operation generates new feature maps $fm_{(k)}$ and $fm_{(l)}$ with size H × W × C. Reshaping technique is required to obtain a single sized feature map (HW × HW). The probability of spatial weights is obtained to represent global information on a spatial scale, as shown in the following equation:

$$Att = fm_{(b)} + \frac{exp(fm_{(k)} \cdot fm_{(l)})}{\sum exp(fm_{(k)} \cdot fm_{(l)})} \cdot fm_{(m)}, \tag{1}$$

where Att measures each position pixel of the local features map with aggregate results from spatial attention and original maps. Furthermore, the module bottleneck is used as a simple feature extractor (1 × 1) without adding a significant amount of computation.

$$Att_{full} = W_{c2} ReLU(LN(W_{c1} Att)), \tag{2}$$

where it takes two convolutions (C_1 and C_2). The linear activation and normalization layers are only placed at the initial convolution. This module shrinks the channel size in the middle and then restores at the end of the module.

3 Implementation Setup

FFDMASK uses the WIDER FACE dataset as a knowledge of facial features to recognize the facial location in a set of images. Meanwhile, the Simulated Masked Face Dataset (SMFD) and the Labeled Faces in the Wild (LFW) are used for the training dataset of the face masks classification model. The detailed configuration of each training stage is shown in Table 1. The training was conducted on the Core I5-6600 CPU @ 3.30 Hz with GTX 1080Ti as an accelerator and Jetson Nano with 128 NVIDIA CUDA as edge devices for testing of the detector.

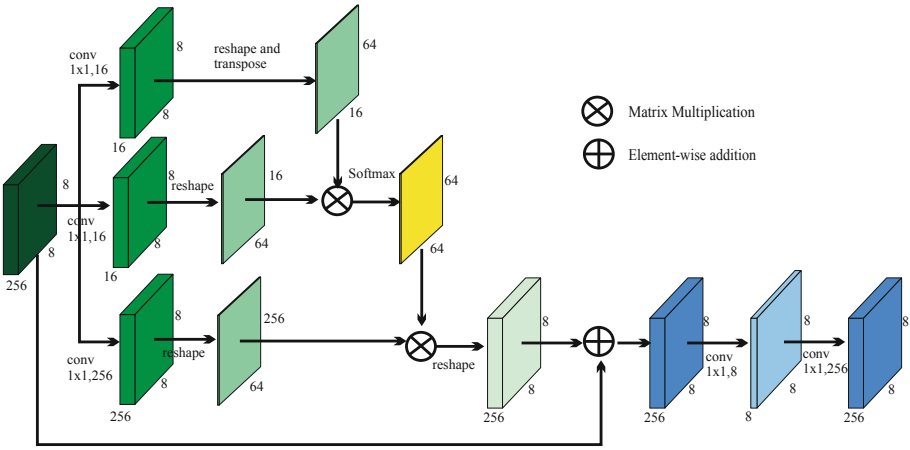

Fig. 4. Attention module.

Table 1. Implementation detail of face detection and face masks classification

Setting	Face detector	Face masks classification
Input image	1024×1024	64×64
Optimizer	Stochastic Gradient Descent (SGD)	Adam
Learning rate	10^{-5}–10^{-3}	10^{-7}–10^{-4}
Batch size	16	8
Total epoch	350	500
Loss function	L1 smooth loss	Categorical Cross-Entropy loss
Epsilon	–	10^{-7}
Weight decay	$5 \cdot 10^{-4}$	–
Momentum	0.9	–
IoU threshold	0.5	–
Framework	Pytorch	Keras

4 Experimental Results

In this section, the proposed architecture of the face detector and face masks classification is evaluated on several datasets. This evaluation shows the qualitative and quantitative results of each dataset. Additionally, another experiment has shown the runtime efficiency of a detector when tested on an edge device.

4.1 Face Detector Results

The FFDMASK detector's evaluation is carried out on the Face Detection Data Set and Benchmark (FDDB) dataset. It is a benchmark dataset consisting of 5,171 faces on 2,845 images. A variety of challenges are provided by this

Table 2. Accuracy of face masks classification on SMFD and LFW datasets

Model	Num of parameter	ACC (%) in SMFD	ACC (%) in LFW
Loey et al. [9]	23,591,816	99.49	100
Proposed	668,746	99.72	100

dataset, including scales, poses, lighting, and complex background. Discrete criteria were chosen as evaluations by comparing the IoU between prediction and ground truth. Figure 5(a) shows that the detector outperforms other competitors (FFCPU and Faceboxes). It is slightly superior to the leading competitors (FFCPU) by 0.2%. Besides, FFDMASK has faster data processing speed on different video input sizes, as shown in Fig. 5(b). Especially for the VGA input size, FFDMASK obtains 51.31 FPS while the FFCPU is 48.27 FPS on the Jetson Nano. These results compare the average speed of each detector when tested at 1000 frames. The quality of the detector performance is also shown in Fig. 5(c). It indicates that the proposed detector can overcome the challenges of occlusion, expressions, accessories, and complex backgrounds.

4.2 Face Masks Classification Results

Evaluation of SMFD Dataset. The dataset consists of 1,376 images, 690 for simulated masked faces, 686 for unmasked faces. It is used for the training and testing phases. This face dataset contains portrait images of male and female faces with a variety of poses and sizes. Face detection is applied to obtain a facial RoI measuring 64×64. This process helps the slim model to focus on learning facial features without being affected by background noise. In this dataset, the proposed architecture obtains an accuracy of 99.72%. This result is superior to Loey et al., which only achieves 99.49% on the same dataset. This success is supported by feature extraction suitable for preprocessed datasets. Figure 6(a) shows the qualitative results of the proposed detector. This result proves that the variations in facial poses in the dataset are not a barrier for the model to get high performance.

Evaluation of LFW Dataset. The benchmark dataset contains 13,000 masked faces for celebrities around the round. The training and testing process uses this dataset separately. The face dataset consists of facial images that were manipulated with an artificial mask that referred to work [9], as shown in Fig. 6(b). LFW masked has a size of 128×128, which instantly provides the RoI of a face that is avoided from the background. Our model uses 64×64 as the input size of RoI, which is reshaped from the original size. As a result, the proposed detector achieves perfect and competitive results with Loey et al., as shown in Table 2. The majority of this dataset is in the frontal pose. It is more comfortable than the SMFD dataset. The proposed model explicitly discriminates against nose

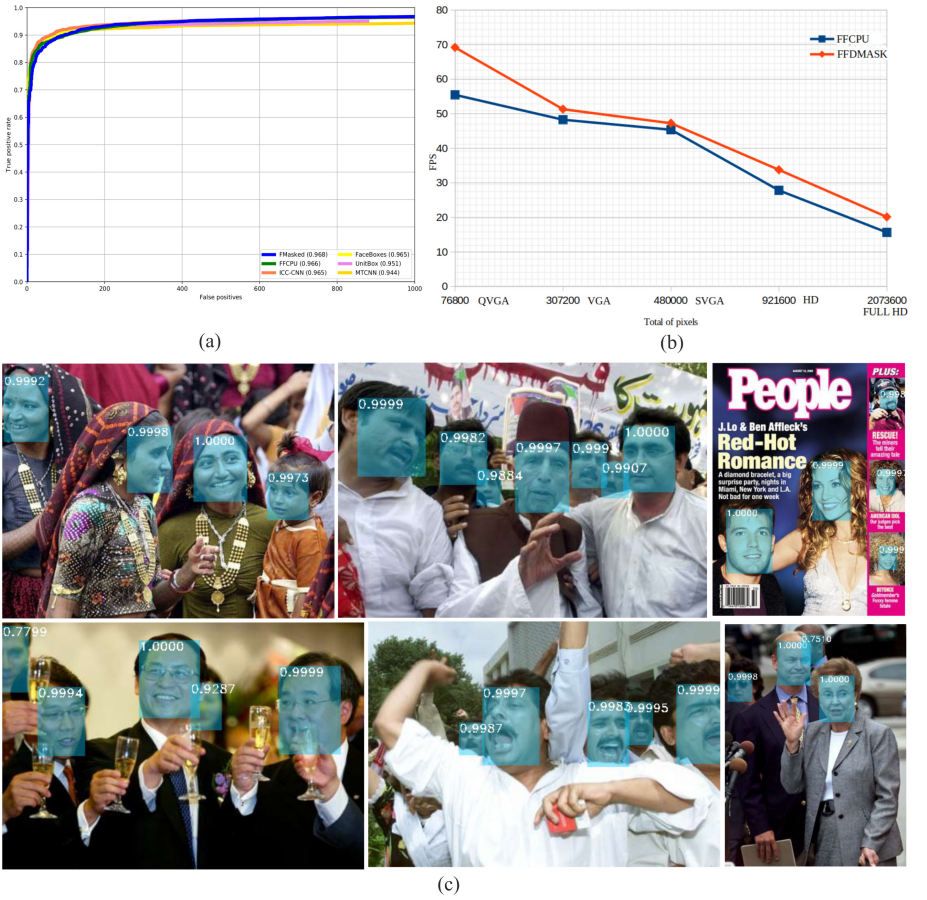

Fig. 5. ROC (Receiver Operating Characteristics) discrete evaluation curve on the FDDB dataset (a), comparison of mean detector speed at different video input sizes (b), qualitative results on FDDB datasets (c).

and mouth features from other features. These features tend to be closed and undetectable when the face is wearing a mask.

4.3 Runtime Efficiency

The practical application recommends a computer vision method to run real-time on portable devices. In general, service robots use mini-computers to process intelligent algorithms and computer vision. It requires a small accelerator to process the computation of the algorithm. Therefore, the proposed architecture is tested on a Jetson Nano to find out the efficiency of the model. Table 3 shows that face detector and face masks classification models have less computing power than competitors. FFDMASK produces a smaller number of parameters than

Fig. 6. Qualitative results on SMFD (a), LFW dataset (b), and running real-time (c).

Table 3. Comparison of detector speeds with competitors on the Jetson Nano.

Model	Input size	Num of parameter	Accuracy (%)	Speed(FPS)
Face detector				
FFCPU [10]	640 × 480	715,844	96.60	48.27
FFDMASK	640 × 480	602,310	96.80	51.31
Face masks classification				
Loey et al. [9]	64 × 64	23,591,816	99.49	5.80
Proposed	64 × 64	668,746	99.72	20.60

FFCPU. It also impacts the speed of the detector. The S-Res module emphasizes computational savings for the residual method without compromising the quality of feature extraction.

Furthermore, the proposed model of face masks classification produces a faster speed than Loey et al. This competitor uses the Resnet-50 backbone module, which generates many parameters and heavyweights. Meanwhile, the proposed module only requires a slim architecture to obtain superior results. Table 3 shows that this model produces 668,746 as of the number of parameters and achieves 20.60 FPS on the Jetson Nano. These results are an accumulation of face detection (VGA-resolution) and masks classification speed (RoI size of 64). Real-time detectors use training data on the SMFD dataset, which tends to have a variety of poses. As a result, the system achieves high performance when it recognizes multi-view masked faces, as shown in Fig. 6(c). Besides, this detector also obtained satisfactory results for the challenge of various colored masks.

5 Conclusion

This paper presents a real-time multi-view face masks recognition system applied to service robots. The two-stage module is used to focus feature extraction on facial RoI. Face detection is responsible for filtering non-face areas, while face masks classification is used to classify Roi faces into two categories. Light and slim architecture do not prevent the detector from obtaining high performance. As a result, two CNN modules can outperform competitors in accuracy and speed. Additionally, the system achieves 21 FPS when running on the Jetson Nano. In future work, the augmentation can increase the dataset varieties and solve the disturbance of lighting and extreme poses.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (2020R1A2C2008972).

References

1. Chen, Q., Sang, L.: Face-mask recognition for fraud prevention using Gaussian mixture model. J. Vis. Commun. Image Represent. **55**, 795–801 (2018). http://www.sciencedirect.com/science/article/pii/S1047320318302050
2. Ejaz, M.S., Islam, M.R.: Masked face recognition using convolutional neural network. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1–6 (2019)
3. Ejaz, M.S., Islam, M.R., Sifatullah, M., Sarker, A.: Implementation of principal component analysis on masked and non-masked face recognition. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1–5 (2019)
4. Fadare, O.O., Okoffo, E.D.: Covid-19 face masks: a potential source of microplastic fibers in the environment. Sci. Total Environ. **737** (2020). http://www.sciencedirect.com/science/article/pii/S0048969720338006
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141–3149 (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
8. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017). http://www.sciencedirect.com/science/article/pii/S0925231216315533
9. Loey, M., Manogaran, G., Taha, M.H.N., Khalifa, N.E.M.: A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the Covid-19 pandemic. Measurement **167**, 108288 (2021). http://www.sciencedirect.com/science/article/pii/S0263224120308289
10. Putro, M.D., Jo, K.: Fast face-CPU: a real-time fast face detector on CPU using deep learning. In: 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE), pp. 55–60 (2020)

11. Putro, M.D., Jo, K.-H.: Real-time multiple faces tracking with moving camera for support service robot. In: Nguyen, N.T., Gaol, F.L., Hong, T.-P., Trawiński, B. (eds.) ACIIDS 2019. LNCS (LNAI), vol. 11432, pp. 639–647. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14802-7_55
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
13. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
14. Zhang, S., Wang, X., Lei, Z., Li, S.Z.: Faceboxes: a CPU real-time and accurate unconstrained face detector. Neurocomputing **364**, 297–309 (2019). http://www.sciencedirect.com/science/article/pii/S0925231219310719