Detecting and Counting People without Mask with Deep Neural Network

Bowen Zhang¹ and Masahiro Tanaka²

¹Graduate School of Natural Science and Technology, Konan University, Hyogo 658-8501, Japan
²Department of Intelligence and Informatics, Konan University, Hyogo 658-8501, Japan
E-mail: zhangbowen930303@gmail.com, m_tanaka@konan-u.ac.jp

Abstract In this work, we develop a real-time computer vision system to detect people and judge whether each person is wearing a mask or not. We construct a 2-stage algorithm based on deep convolutional neural networks, where the masks are treated as objects in an image. Furthermore, in order to improve the accuracy of recognizing masks when the human face occupies a large area of the image, we adopt the dilated convolution algorithm to solve this problem. Based on the recent research of COVID-19 for infection danger, this system can send dangerous signal level 1-3 due to the proportion of masked people in the captured image. Due to the report of infection danger criteria, this system can send an alarm of three levels with the borders at 20% and 50% of the people without masks in the area, which can notify people in the area as a safe, a little dangerous, or a particularly dangerous situation.

Keywords: deep neural network, dilated convolution, mask detection, image classification

1. Introduction

Recently, De Kai et al.[1] showed that there was a significant impact on the prevention of COVID-19 when at least 80% of people are wearing masks, which is demonstrated through individual agent-based modeling (ABM) Monte Carlo simulation. In contrast, the control work is only minimally impacted on pandemic when 50% or less people are wearing masks.

In this work, we design a real-time processing system based on two-stages of deep learning algorithms to detect people with or without masks in image sequences, and count them. Based on the indicated above, we define thresholds at the number of people without masks with 20% for a significant impact for COVID-19 and 50% for a minimal one, respectively. Therefore, this system could warn people who are in a safe, a little dangerous, or a particularly dangerous environment matched by dangerous level 1, 2 or 3. Furthermore, we describe the experimental details on training the proposed model. Finally, we show that our proposed model attains the performance of 94.8% on the test set of the Real-World Masked Face dataset[2].

2. Model

Deep Convolutional Neural Networks (DCNNs) have recently been the mainstream methods for various computer vision tasks, such as image classification and object detection. In order to achieve the mask recognition work of this study, we built a model whose structure is a cascade of two DCNN modules, a face detector and a mask detector by applying object detection and classification, respectively.

2.1 Face detector (first step)

Object detection is a method to locate numerous kinds of objects such as people, vehicles and human faces, etc. Here, the primary face detection algorithm is constructed with 2-class (“face” or not) object detection structure.

We first compare the properties of AdaBoost-based Haar-like features cascade classifier[3] (we call this “AdaBoost” hereafter) and PyramidBox algorithm[4]. The former is the standard face detector of computer vision, and it was superior in the recognition speed to the latter. However this was not robust enough especially when the human faces were concealed by masks. The latter one worked effectively on small, blurred and
partially occluded faces for detection purposes. Thus, we adopt the PyramidBox algorithm as the face detector.

2.2 Mask detector (second step)

The rectangle image detected as a face in the first step is sent to this mask detector as shown in Fig. 1. We create a DCNN structure to deal with the 2-class (with or without a mask) image classification problem. The structure of DCNN consists of six-layers. The 1st layer consists of 64 channels of $3 \times 3$ convolution kernels (Conv.), 2nd layer consists of 128 channels of $3 \times 3$ Conv., 3rd layer consists of 256 channels of $3 \times 3$ Conv., 4th layer consists of 512 channels of $3 \times 3$ Conv., 5th layer consists of $N \times 50$-sized linear layer where $N$ is the size of the 4th layer output feature map, 6th layer is a $50 \times 2$-sized linear layer whose feature map is concatenated and sent to the last Softmax layer. Herein we briefly describe the convolution and linear calculation. Convolution calculation is to multiply each value of its kernels with the corresponding pixel values of the image and sum them up. Linear calculation is to perform a weighted summation of all pixel values of the image, and the summed value is passed to the nonlinear activation function (ReLU), and finally passed to the next layer.

We use the max pooling layers after the 2nd and 4th layers. Furthermore, we concatenate a dropout layer to the output of the 4th layer which has the largest number of weights for preventing overfitting, while reducing some calculations[5].

2.3 Dilated convolution for large receptive fields

Dilated convolution[6] enables the DCNN to get huger receptive fields while keeping the number of convolution variables the same. Therefore, it can keep the feature map resolution while keeping the calculation speed unchanged. There is a hyperparameter of dilated convolution called “dilation rate”. We can adjust it such as raising kernel size by filling the gaps with zeros. In this case, the larger objects than the annotated objects can be recognized. In our case, it adapts to the case when the human faces are very close to the camera.

In this system, we enlarge all the dilation rates from 1 to 2 as shown in Fig. 2. Second row shows the case when the dilation rate= 1 as shown in $3 \times 3$ kernel (green). And this $3 \times 3$ kernel fills all the gaps between the 9 parameters with zeros like the first row one (i.e., dilation rate= 2). At the same time, receptive field has successfully been expanded from 9 pixels to 25 per convolution layer as shown with pink color. Besides, the resolution of output remains the same as general convolution at $5 \times 5$ by calculating Eq. (1), where $O_w$ is the width of output size, $I_w$ is the width of input size, $p$ is the padding size, $s$ is the stride size, $k$ is the kernel size, and $r$ is the dilation rate. The amount of convolution calculations are kept intact as $3 \times 3 = 9$.

$$O_w = \frac{I_w + 2p - (k + (k - 1)(r - 1))}{s} + 1 \quad (1)$$

2.4 Masks counter

We also exploit the outputs of our model to count the number of people who are wearing masks or not in each frame from live camera to correspond to the problem mentioned in [1]. In this experiment, we issue
messages as “safe”, “a little dangerous” or “particularly dangerous” with borders at 20% and 50% for the number of people without masks in the image.

2.5 Preprocessing

We illustrate our image preprocessing method before training the mask detector through a simple example in Fig. 3. Given an original RGB image, we first convert it to a gray scale image, and then resize the image to $100 \times 100$ for maintaining the DCNN’s computational efficiency. Finally, we normalize the image value, where each of its original pixel value $v \in \{0, 1, \ldots, 255\}$ is converted to a decimal $d$ by $d = v/255$, thus yielding $d \in [0, 1]$.

![Fig. 3 Operations of image preprocessing](image)

3. Experimental results

We applied the built-in function of Paddlehub which is a library for various AI applications belonging to PaddlePaddle deep learning framework developed by Baidu, Inc. for face detector module and Keras deep learning framework for mask detector module on Python3. We trained our model on the part of RMFD dataset[2], consisting of 1521, 381 and 212 images for training, validation and testing, respectively. The performance was measured in terms of mask detector’s accuracy.

We trained our model on the 2-class classification work by the cross entropy loss function, and we use a mini-batch of 32 images and initialize the weights with “He initialization”[7]. Then we used Adam optimization[8] with $\alpha=0.001$, $\beta_1=0.9$, $\beta_2=0.999$, $\epsilon=10^{-8}$ and $L_2$ regularization with a weight decay rate of $3 \times 10^{-4}$. In the end, we initialized the learning rate with $10^{-4}$ to start the training. The loss after 50 epochs was about 0.01 on the training set and about 0.24 on the validation set. Accuracy of training, validation and test set were about 0.997, 0.945 and 0.948 respectively as illustrated in Fig. 4. Here, the training time of 50 epochs was about 227.77 minutes on a single 4 core-CPU by using Keras deep learning framework.

3.1 AdaBoost vs PyramidBox in face detector

In the case of AdaBoost, especially when the human faces were very small, or when people worn masks, the recognition results were very easy to make mistakes. So, only approximately 33% of the human faces were detected. The insufficient detection of faces with masks causes the next step’s mask detector to fail to predict correctly. On the other hand, nearly 90% faces can be detected in the case of PyramidBox on our test set even with the above problems.

3.2 General convolution vs dilated convolution in mask detector

We used 23 images as the test set. For instance, when the faces occupied large areas of the left 2 images in Fig. 5, masks were not recognized and the confidence scores were only 64.96% and 76.30%. However in the right 2 images, masks were correctly detected, and they achieved 100% and 97.64% confidence scores.

3.3 Mask counter and dangerous level

As illustrated in Fig. 6, six persons without wearing masks were detected, and one person with a mask was mistakenly identified as not wearing a mask among 27 persons. Since the 25.9% people were not wearing mask, this situation was judged to the dangerous level 2. Herein the faces were intentionally blurred after computation task for privacy reasons.

On the other hand, because our model are more robust than AdaBoost, very small objects also can be recognized as shown in Fig. 7. This picture was taken at the Sannomiya Station in Kobe, Japan during the rush hour. All the nine persons were successfully detected, of which two were considered to be without masks and the other seven were with masks. In more detail, we can clearly observe that the person who is judged to be without a mask on the right didn’t cover his nose with the mask. Even wearing a mask such as
Fig. 5 Confidence score of 2 samples when the faces are close to the camera: First column: predicted by general convolution. Second column: predicted by dilated convolution.

Fig. 6 Total: 27 persons, mask: 20 persons (green), no mask: 7 persons (red), failure: 1 person (blue), dangerous level 2

this situation has no effect on the spread of the virus, so it needs to be detected.

3.4 Discussion

The computational speed of our system was about 8 FPS (frames per second) on a single 4-core CPU without a GPU parallel acceleration during the edge computing. In other words, the mask recognition task of the surrounding people can be processed every about 0.125 seconds. And according to this report [9], the human average reaction time is about 0.17 seconds for an audio stimulus. Because 0.125 is less than 0.17, when our system alerts people such as “It is now in dangerous level 3, please leave here!” , our system’s speed will not affect people hearing this message.

4. Conclusions

In this work, we presented a structure composed of a cascade of face detector module and mask detector module for mask recognition task. As far as we know, other small systems can only detect one person’s mask condition, while this system is innovative enough to be able to work on a standard laptop PC without GPUs. Through our network structure design and the training strategy, the proposed model yielded a good performance of 94.8% on the test set. And we added a function of giving different prompts to our system according to different dangerous levels. Our approach successfully detected the masks and counted the number of people who were without masks, both of which are practical concerns.

This research was partially supported by MEXT, Japan.

References

[1] D. Kai, G. P. Goldstein, A. Morgunov, V. Nangalia and A. Rotkirch: Universal masking is urgent in the COVID-19 pandemic: SEIR and agent based models, empirical validation, policy recommendations, arXiv:2004.13553, 2020.
[2] Real-World Masked Face Dataset: https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset
[3] P. Viola and M. Jones: Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Vol. 1, pp. I-511-I-518, IEEE, 2001.
[4] X. Tang, D. K. Du, Z. He and J. Liu: Pyramidbox: A context-assisted single shot face detector, Proceedings of the European Conference on Computer Vision (ECCV), pp. 797-813, 2018.
[5] I. Goodfellow, Y. Bengio and A. Courville: Deep Learning, Massachusetts, MIT Press, 2016.
Fig. 7 Total: 9 persons, mask: 7 persons (green), no mask: 2 persons (red), failure: 1 person (blue), dangerous level 2

[6] Liang-Chieh Chen et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs, arXiv preprint, arXiv:1412.7062, 2014.
[7] K. He, X. Zhang, S. Ren and J. Sun: Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification, Proceedings of the IEEE International Conference on Computer Vision, pp. 1026-1034, 2015.
[8] P. D. Kingma and J. Ba: Adam: A method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.
[9] Experiment: How fast your brain reacts to stimuli, https://backyardbrains.com/experiments/reactiontime

Bowen Zhang received his B.Eng. degree from the Department of Mechanical Engineering, Dalian Institute of Science and Technology, China in 2016. Currently, he belongs to Graduate School of Natural Science and Technology, Konan University, Japan. His research interest includes the theory and applications of machine learning, and deep learning applications to image processing.

Masahiro Tanaka received his B.Eng., M.Eng. and D.Eng. degrees from the Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan, in 1979, 1981 and 1990, respectively. He previously worked at Shimadzu Corp., Shiga University, and Okayama University, and he has been a full professor at Konan University, Kobe, Japan since 1999, where he is currently at the Department of Intelligence and Informatics. His research interest includes neural networks, deep learning, computer vision, state estimation and soft computing by using visual sensors including depth sensors. He is a member of IEEE, SICE, ISCIE, IEEJ, RSJ and others.

(Received April 14, 2021; revised May 14, 2021)