A Deep Convolutional Neural Network to Limit Virus Spread Using Facial Mask Segmentation

D. Lefloch and J. M. Wang*

Software Engineering Department, Tiangong University, Tianjin, 300387, China
E-mail: prof.d.le.floch@gmail.com; wangjianming@tiangong.edu.cn
**Corresponding Author*

Received 13 April 2021; Accepted 21 May 2021;
Publication 06 July 2021

Abstract

Due to the recent COVID-19 outbreak the world has experienced many challenges. Limit and control the virus spread rate is one of them. This letter focuses on limiting the speed of virus spreading by monitoring the use of facial mask in crowded public environments such as tourism places, commercial centres, etc. The proposed method first accurately localizes faces using a state-of-the-art approach and, segments facial mask in a second step. The facial mask segmentation allows to distinguish whether the current subject is wearing a facial mask or not but also if it is properly covering the human face. Indeed, most recent face detection algorithms provide as output a set of facial features such as nose tip and mouth corners. By combining these facial features with facial mask segmentation, the proposed method detects real-time subjects that indirectly encourage virus spread in crowded environments. The proposed facial mask segmentation model is trained with pairs of RGB images and its corresponding alpha image created by extending the publicly available real-world masked face dataset. Further, the proposed

Journal of Web Engineering, Vol. 20_4, 1177–1188.
doi: 10.13052/jwe1540-9589.20414

model is pruned and optimized using the TensorRt library to be usable for real-world applications.

Keywords: Mask segmentation, face detection, convolutional neural networks, deep learning.

1 Introduction

Face detection [1–3] has taken a lot of interest in the last decades. It spans various applications from face reconstruction [4] to face recognition [5], which requires precise face detection to operate properly. Furthermore, face detection can be considered more relevant today since it is widely used on smartphones via video. Due to the recent COVID-19 outbreak, the world has experienced many challenges. This research aims to limit and control the virus spread rate by first precisely localizing human faces and classifying each of them by different risk levels. To this end, we propose a novel deep convolution neural network (CNN) that precisely segments facial mask from a single RGB image. The proposed network first relies on a robust face detector and precise facial landmarks localizer. However, most face detection methods suffered less accurate when the target persons are wearing a facial mask [6].

In order to segment a facial mask from the human face, the previous non-learning algorithms detected the object by matching similar patches. For example, Park et al. [7] removed eyeglasses on the human faces using principal components analysis (PCA) reconstruction with an iterative approach that compensates for error.

Recent advances in learning-based methods empowered image segmentation algorithms significantly. Those methods usually outperformed the non-learning methods but required large-scale datasets and/or data augmentation mechanisms [8]. Lin et al. [9] used an improved version of Mask Region-based CNN (Mask R-CNN), usually used to segment different object classes, to segment multiple face instances. Whereas Saito et al. [8] used a popular convolution network for object detection (VGG-16 [10]) coupled with a two-stream deconvolution network (DeconvNet [11]) to robustly segment human faces with strong occlusion.

In order to address the challenging COVID-19 outbreak, many methods recently arise to improve the accuracy of face mask detection. Loey et al. [12] proposed a hybrid method that combines a Resnet50 to localize face features and an SVM classifier to determine whether the localized face is wearing

a face mask or not. Lodh et al. [13] built a real-world application where a MobileNet_V2 was used for the face mask detection as well as their face recognition module. Their system could identify whether a person is wearing a mask or not and send a notification to all persons not wearing a face mask. In [14], an automated system is proposed to limit COVID-19 by finding people who are not wearing any facial mask in public places using monitoring cameras. They use a standard CNN based deep neural network to detect whether a person is wearing a facial mask or not. They have reported a 98.7% accuracy for their face classification model.

The main contributions of this work are as follows:

(1) Train state-of-the-art neural network for face detection and face landmarks localizer to be robust against faces wearing a facial mask with real-time constraint.
(2) Design a novel deep convolutional network to precisely segment facial masks with real-time constraint. The proposed model is able to segment facial masks in real-time with a very high accuracy.
(3) Use the facial segmentation with the facial landmarks to detect whether the facial mask is properly covering risk zone of the face (such as the mouth and the nose regions). Limit and control the rate of virus spread by classifying detected human faces with different risk levels.

2 Method

The proposed method consists of three parts: face localization, facial mask segmentation, and risk classification.

The face localization used in this letter is fully based on the multiple cascaded convolution neural networks (MTCNN) proposed by Zhang et al. [3]. The face detector model was trained with two datasets: the WIDER FACE dataset [15] and the CelebA dataset [16]. However, the state-of-the-art MTCNN net did not perform reliably as a face and facial landmark detector on the real-world masked face dataset (RMFD) [6]. The two datasets were coupled with an extension of the RMFD dataset where five additional face landmarks, including facial masks, were manually annotated with an additional binary image representing the facial mask segmentation (see Figure 1 for an overview of the extended dataset). Note that facial landmarks detection is challenging for masked faces but can be deduced with enough precision based on face characteristics (such as face orientation given by eyes and nose, face size, and chin position).

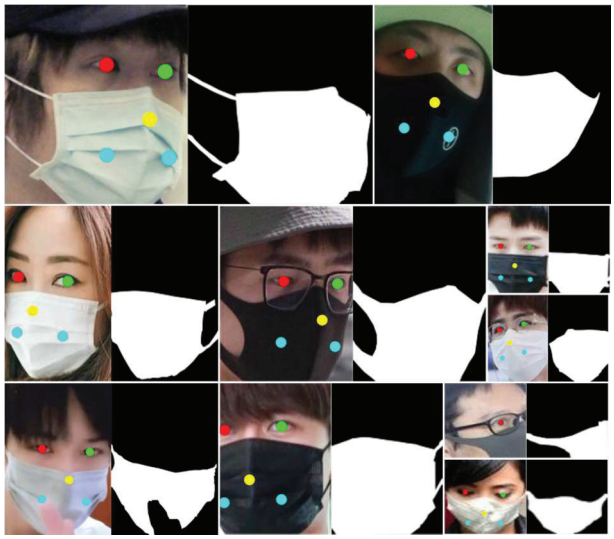

Figure 1 Overview of the RMFD dataset extended with binary images and landmarks. Each RGB image is paired with its corresponding binary mask image, labeling all facial mask pixels and five face features (eyes, nose tip, and mouth corners) precisely.

Like [3], the face and facial landmark detector were trained in three stages with random image patches with the corresponding losses (cross-entropy loss for the face classification, Euclidean loss for the bounding box regression and the facial landmark localization). Once trained, the models were pruned and optimized with half float precision using TensorRt without any quality loss maintaining real-time face detection rate on 1080p resolution on a computer equipped with a GeForce RTX 2080 GPU.

The proposed facial mask segmentation employs a variant of the popular U-NET architecture [17] (See Figure 2). This model architecture provides high-quality segmentation results via skip connections. Also, it consists of standard layers (such as Convolution layers, Max pooling layer, Concatenation layers, Deconvolution layers, and Activation layers) that can be easily pruned and optimized. The modified architecture adopts a set of convolutions and deconvolutions with a filter size of 3×3. Each convolution and deconvolution are followed with a Relu activation function and a batch normalization (BN) layer. Only the last convolution that predicts the facial mask segmentation image does not require a BN layer and use a sigmoid activation layer. Similar to [8], the input to the network is a $128 \times 128 \times 3$ RGB image for training. However, the prediction is not a simple dense 2-class

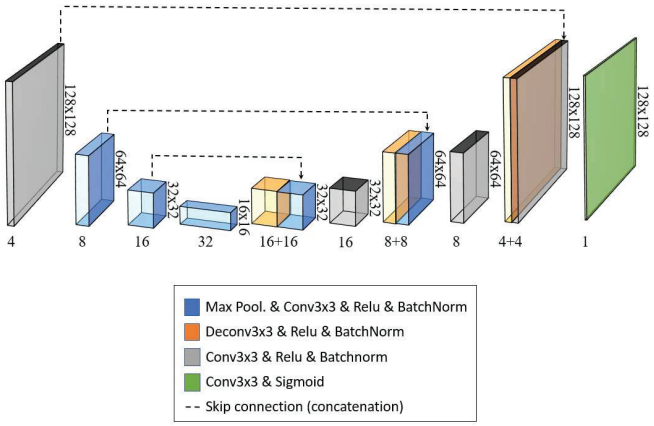

Figure 2 The modified U-NET architecture for facial mask segmentation.

binary image but a $128 \times 128 \times 1$ alpha image to better model the mixed pixels on the blurry region. Both the input and output of the network are storing floating precision values between [0,1]. Each max-pooling layer uses a 2×2 stride parameter leading to a down-sampled tensor of factor 2. In order to train the facial mask segmentation model, the extended RMFD dataset was used (face samples with mask and face samples without mask) and the annotated binary image representing the facial mask. Note that 80% of the dataset was used for training, and the rest was used for validation. Along training, image patches were cropped randomly and

scaled to a 128×128 image size (bilinear interpolation). The learning rate was halved each time the validation loss remained unchanged or larger for three consecutive times.

The following perturbations were applied to augment the training data:

- Face images are converted to LAB colour space and perturbated uniformly channel A and B on pixels of the facial mask to slightly change its colour. Finally, the image is converted back the image to the original RGB colour space.
- A Gaussian blur with a random kernel size is applied to the face image and the binary image to be more robust to blur due to defocus. Note that the binary image of the facial mask segmentation becomes an alpha image which, is used as ground-truth during training.
- A final gamma correction is applied to be more robust to change in lighting conditions.

The facial mask segmentation is formulated as a regression problem where the following Euclidean loss is minimized:

$$L_i^{alpha} = |\widehat{Y}_i^{alpha} - Y_i^{alpha}|_2^2, \tag{1}$$

where $\widehat{Y}_i^{alpha}$ is the facial mask alpha image obtained from the network and Y_i^{alpha} is the ground-truth alpha image. During inference, the input of the model is an RGB image cropped and centred based on the bounding box given by the face localization network enlarged by a factor of 1.25 and later on scaled to an image resolution of 128×128. In this way, the facial mask segmentation network only operates on a region of interest (improving the performance) that contains a human face with high probability. Note that if several faces are located during the first stage of face localizations, the second stage inference is made with a maximum batch size of 32 faces.

Finally, the proposed processing pipeline classifies all detected faces as possible risks using the output of both previous stages (face and landmark detections and facial mask segmentation). Three different types of risk are distinguished (high-risk, intermediate-risk, and low-risk). A human face is classified as high-risk (no facial mask) if all pixels of the predicted facial mask are smaller than 0.10 (classification: no facial mask). A human face is classified as intermediate-risk if it is not classified as high-risk, and one of the three face mask landmarks (i.e. nose tip, left and right mouth corners) is classified has no facial mask (i.e., the value of the corresponding pixel of the landmark in the facial mask segmentation is smaller than 0.10). Indeed, a facial mask is functional with its maximum efficiency if wear correctly (see the recommendation from the Centers for disease control and prevention [18]). A human face is classified as low-risk if it is not classified as high- and intermediate-risk, meaning that all three face landmarks belong to pixels in the facial mask segmentation. Figure 3 gives an example of such three risk

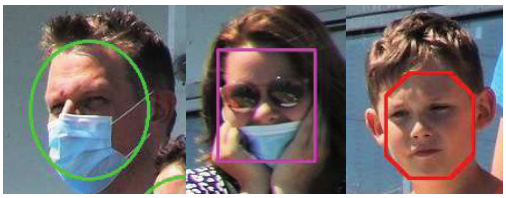

Figure 3 Face risk classification examples (from left to right: low-risk, intermediate-risk and high-risk). These human faces were extracted from the cruse image (source: maxpixel.net).

levels. A face with a circle denotes a low-risk face, the one with a rectangle denotes the intermediate-risk face, and the one with an octagon denotes the high-risk face. For a complete visualization of facial landmarks and facial mask segmentation, refer to the figures in the results section.

3 Results

The usability of the proposed method was evaluated in three different challenging situations. For all results, a full overview of the risk detection is first shown on the complete image, and then each of the detected faces is zoomed in to visualize better the quality of the face landmarks and the facial mask segmentation. Each facial landmark is denoted with a bold circle (left eye, right eye, nose tip and mouth corners). The facial mask segmentation is overlayed as light and dark colour if the human face is classified as low-risk and intermediate-risk, respectively.

The **cruse** image includes the problem of strong sunlight conditions. In this use case, all human faces are approximately the same scale since each human is approximately the same depth, and no face is occluded on this image. All faces roughly span the full range of face poses (front face to profile face) with different ages (kid to adult). All eleven human faces are accurately detected, and the face risk is classified correctly. See Figure 4 for a complete quality overview of the proposed method. Note how accurate the face landmarks are even in the presence of the facial mask. All kid faces are accurately labelled by our method; three of them are marked as high-risk since they indeed do not wear any facial mask (face 3, face 6 and face 7); note also how good is the facial detection for face 3 and face 6, both having very challenging head-pose (extreme pan angle); the fourth kid face (face 4) is recognized as an intermediate-risk since the facial mask is placed on the chin. For the adult faces, face 10 and face 11 are correctly detected as intermediate-risk since their facial mask are wrongly placed (under nose).

The second use case (**event-crowd** image) includes a crowded environment with different depth defocus blur, face occlusions, and face scales (see Figure 5). The proposed system can detect precisely three human faces in total. As shown in Figure 5, the detection quality of the extremely blurry human face and the quality of the face landmarks for facial mask are still good. In this image, many other human faces are present, but the proposed face localization is not able to perceive them due to either their strong occlusion and/or amount of defocus blur. The adult face (face 3) is correctly

Figure 4 Quality overview of the proposed framework on the **cruse** use case (source: maxpixel.net).

Figure 5 Quality overview of the proposed framework on the **event-crowd** use case (source: maxpixel.net).

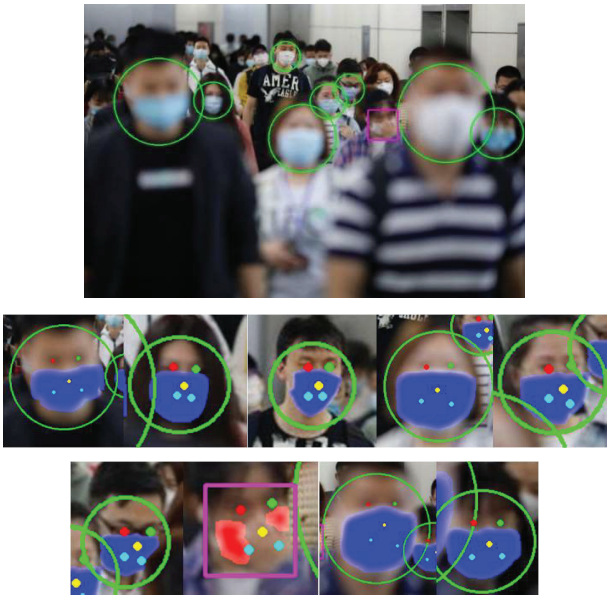

Figure 6 Quality overview of the proposed framework on the **crowd** use case (source: REUTERS/Carlos Garcia Rawlins).

detected as an intermediate-risk since the facial mask is wrongly placed on the chin. The detection quality of the defocus face (face 1) is also correctly labelled as low-risk and the five facial landmarks are also correctly retrieved despite of the extreme blur on that face. It clearly demonstrates the benefit of our proposed model that includes during training many blurry samples to simulate defocus effect.

The third and last use case (**crowd** image) includes a highly crowded environment, which spans a variety of sharp and blurry human faces at different scales with strong occlusions. In this use case, the proposed face localization and landmarks are highly accurate (see Figure 6). The proposed face risk classification is mostly accurate. Indeed, one female subject is partially occluding her face mask by her hand, leading to misclassification. This is one limitation of the proposed method, which cannot be avoided. However, the proposed method is designed to be real-time and thus can operate on image sequences. Thus, as soon as the female subject removes her hand from the occlusion area, the facial mask segmentation will be accurate again, leading to the appropriate risk classification. All other faces are correctly classified as low-risk since all facial masks are correctly placed.

In this example, one can clearly see the robustness of the proposed method against different scales of faces caused by different level of depths leading to different level of defocus blur for each of the faces.

4 Conclusion

In this work, we present a novel real-time method to accurately classified each detected human face by risk levels. To mark human faces as risk levels and thus monitor human behaviour in public areas can be highly valuable to control and limit the rate of virus spread.

The proposed method is shown to be very performant to segment facial mask and detect where it is covering the appropriate parts of the face (mouth and nose). Few false positives were also discussed and are handle properly by our real-time system. We show that our method can operate in real-time constraint using a standard computer equipped with a GeForce RTX 2080 GPU. For future works, we would like to recognize facial mask types since different facial masks have different efficiencies in terms of virus spread. Furthermore, the current method could be extended to the detection of minimal social distancing (e.g., the distance of less than 1-meter radius between persons can be considered as potential risk), which will greatly reduce the rate of the virus spread. And finally, in order to refine the potential risk levels of a person, it would be highly valuable to train our model to have a rough estimation of the age of the person since elder people are persons with the higher risk.

References

[1] P. Viola, M. J. Jones, 'Robust real-time face detection', International Journal of Computer Vision, pp. 137–154, 2004.

[2] D. Chen, G. Hua, F. Wen, J. Sun, 'Supervised transformer network for efficient face detection', European Conference on Computer Vision, pp. 122–138, 2016.

[3] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, 'Joint face detection and alignment using multitask cascaded convolutional networks', IEEE Signal Processing Letters, pp. 1499–1503, 2016.

[4] F. Liu, D. Zeng, Q. Zhao, X. Liu, 'Joint face alignment and 3d face reconstruction', European Conference on Computer Vision, pp. 545–560, 2016.

[5] J. Deng, J. Guo, N. Xue, S. Zafeiriou, 'Arcface: Additive angular margin loss for deep face recognition', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699, 2019.

[6] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, J. Liang, 'Masked face recognition dataset and application', arXiv preprint arXiv:2003.09093, 2020.

[7] J. S. Park, Y. H. Oh, S. C. Ahn, S. W. Lee, 'Glasses removal from facial image using recursive error compensation', IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 805–811, 2005.

[8] S. Saito, T. Li, H. Li, 'Real-time facial segmentation and performance capture from rgb input', European Conference on Computer Vision, pp. 224–261, 2016.

[9] K. Lin, H. Zhao, J. Ly, C. Li, X. Liu, R. Chen, R. Zhao, 'Face Detection and Segmentation Based on Improved Mask R-CNN', Discrete Dynamics in Nature and Society, pp. 1–11, 2020.

[10] K. Simonyan, A. Zisserman, 'Very deep convolutional networks for large-scale image recognition', International Conference on Learning Representations, 2015.

[11] H. Noh, S. Hong, B. Han, 'Learning deconvolution network for semantic segmentation', Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528, 2015.

[12] M. Loey, G. Manogaran, M. H. N. Taha, 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, vol. 167, 2020.

[13] A. Lodh, U. Saxena, A. Motwani, L. Shakkeera, V. Y. Sharmasth, 'Prototype for Integration of Face Mask Detection and Person Identification Model–COVID-19', 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1361–1367, 2020.

[14] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud, J.H. Kim, 'An Automated System to Limit COVID-19 Using Facial Mask Detection in Smart City Network', IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1–5, 2020.

[15] S. Yang, P. Luo, C. C. Loy, X. Tang, 'Wider face: A face detection benchmark', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5525–5533, 2016.

[16] Z. Liu, P. Luo, X. Wang, X. Tang, 'Deep learning face attributes in the wild', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3730–3738, 2015.
[17] O. Ronneberger, P. Fischer, T. Brox, 'U-net: Convolutional networks for biomedical image segmentation', International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.
[18] CDC, 'How to wear masks' https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-to-wear-cloth-face-coverings.html, updated September 2020.

Biographies

D. Lefloch is currently working at Tiangong University. His research includes Artificial Intelligence, Deep Learning.

J. M. Wang is currently working as a Professor at Tiangong University. He is also dean of the Software Engineering Department. His research includes machine learning, signal processing.