DOI: https://doi.org/10.24113/ijoscience.v7i1.350

RESCNN: A Deep Learning Approach for Unmasking Face Mask

Bhusra Fatima
M.Tech Scholar
Sagar Institute of Research & Technology
Bhopal, M. P, India
bushrafatima94@gmail.com

Arun Kumar Jhapate
Professor
Sagar Institute of Research & Technology
Bhopal, M. P, India

Abstract: **Due to outbreak of COVID-19 pandemic, the trend of wearing mask is rising all over the world. Before such pandemic people wear mask only to protect themselves from pollution. While other people are self-conscious about their looks, they hide their emotions from the public by hiding their faces. But in current scenario, after pandemic, it is compulsory to wear mask everywhere as researchers and doctors have proved that wearing face masks works on impeding COVID-19 transmission. Nowadays, all attendance system or surveillance systems, etc. are integrated with AI technology in which face recognition is considered as input variable. So, there is need to determine all facial landmarks to recognize an individual. In this research work, Residual Convolution Neural Network (ResCNN), network is designed and simulated which unmasks the face mask present on face and restore mask area and recognize an individual. The result analysis is performed in three different cases or scenario, one normal frontal facial region with mask, in another case the masked face is tilted and in third case the noisy masked face is taken as input. The noise in image occurs due to many physical conditions. The dataset for training of ResCNN is prepared by masking facial images taken from CelebA dataset and MFR datasets to prove the efficiency of the proposed model.**

Keywords*:* COVID-19, Face Mask, Detection, Face Unmasking, CNN, Residual Learning.

I. INTRODUCTION

Face detection has become a very interesting problem in image processing and computer vision. It has a range of applications from facial motion capture to facial recognition which initially requires face detection with efficient accuracy. Face detection is more relevant today because it is not only used on images but also in video applications such as real-time surveillance and face detection in videos. High-precision image classification is now possible with the advancement of convolutional networks. Pixel-level information is often needed after face detection, which most face detection methods do not provide[1]-[3]. Getting pixel-level detail was a difficult part of semantic segmentation. Semantic segmentation is the process of assigning a label to each pixel of the image. In our case, the labels are either facial or non-facial. Semantic segmentation is then used to separate the face by classifying each pixel of the front or background image. Also, most of the widely used face detection algorithms tend to focus on front face detection[4][5].

II. LITERATURE REVIEW

Trend of wearing masks in public is growing in recent years all over the world. Previously, some people wear masks to guard themselves from pollution, while some people are self-conscious about their look and they want to hide their face and emotions from the public. But nowadays occurrence of COVID-19 is another cause to wear mask while we are in public places. To address this task, early non-learning-based works [2] erase unwanted object and synthesize the missing content by matching similar patches from the remainder of the image. Some of the noteworthy contribution in this field are discussed as below:

Din et al. [6] proposed a face unmasking technique aimed at removing masked objects in facial images. The problem is solved in two stages: detecting the mask object and deleting mask area. The first step of our model automatically creates a binary segmentation for the mask area. Then the second step removes the mask and synthesizes the affected area with fine detail while maintaining the overall texture of the face structure. To this end, a GAN-based network was used which uses two discriminators, one discriminator that helps to learn the overall facial structure and then another discriminator used to focus learning on the deeply missing region.

Dong et al. [7] proposed GAN network architecture to generate damaged region of facial region. This work restores the damaged region by using radar data restoration.

Li et al. [8] established a method, HGL, to address head posture classification by applying color consistency analysis to images and online portraits. The proposed HGL method combines the H-channel of the HSV color space with the facial portrait and grayscale image and trains the CNN to extract the characteristics for classification. Evaluation of the MAFA dataset shows that the proposed method performed better than

algorithms based on facial mark detection and convolution network.

Hussain et al [9] used real-time deep learning classification and facial emotion recognition. They used VGG-16 to classify seven facial expressions. The proposed model was trained on the KDEF dataset and achieved an accuracy of 88%.

Khan et al [10] presented an interactive method called MRGAN. The method depends on whether you get the user's microphone zone and use the Generative Adversarial Network to recreate that zone.

III. OBJECTIVES

The main objectives of this research work are:

- Automatic removal of mask and restoration of facial feature, i.e. retaining the original structure of the face.
- To reduce error rate while restoration.
- To apply on different facial mask condition i.e. on rotated faces or with noisy image.
- To overcome the data scarcity problem, a new dataset is prepared with and without mask on CelebA and MFR dataset.

IV. METHODOLOGY

The trend of wearing face masks in public is rising due to the COVID19 coronavirus epidemic all over the world. Before Covid-19, People used to wear masks to protect their health from air pollution. While other people are self-conscious about their looks, they hide their emotions from the public by hiding their faces. Scientists proofed that wearing face masks works on impeding COVID-19 transmission [1].

In this research, a methodology is introduced for face detection model with and without mask that is based on deep learning. The proposed model can be integrated with surveillance cameras to impede the COVID-19 transmission by allowing the detection of people who are not wearing face masks.

A. Convolution Neural Network (CNN)

CNN consists of an input layer, an output layer and no. of hidden layers in which alternating sequence of linear and non-linear operations are done. The output of hidden units is termed as feature maps. The convolutional filter with pre-determined size and weight is multiplied with the output of the previous layer. The input image multiplied by the convolutional filter kernel is further modified by activation function. The output of lth layer i.e x_j^l of CNN is given by the following equation.

$$x_j^l = f\left(\sum_{i \epsilon M_j} x_i^{l-1} * w_{ij}^l + b_j^l\right) \quad (1)$$

Where, M_j represents the selection of the input feature maps

$x_i^{l-1}$is the output of the previous layer i.e. feature maps

w_{ij}^l the weight of the convolution kernel of the lth layer

f is the activation function

b_j^l is the bias of the lth layer.

The pooling layers reduce the no. of parameters of the feature maps and it also reduces the number of computations required for network training.

The output size after convolution operation is given by as:

$$Output\ Size = \frac{(N + 2P - F)}{S} + 1 \quad (2)$$

where N ×N is the dimension of an input image

P = Padding depth

F×F = Dimension of the convolutional filter

S = Stride which specifies how many pixels a filter is translated horizontally and vertically. The basic architecture of CNN is given in Figure 1.

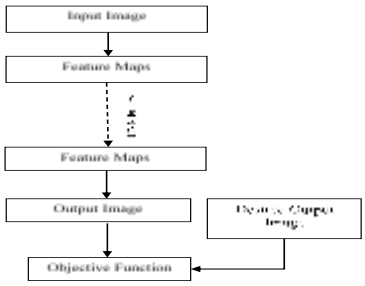

Fig. 1 CNN Architecture

The four key feature of CNN are -:

- Convolutional layer (CL)
- Rectified Linear Unit (ReLU)
- Pooling layer (PL)
- Fully connected layer (FCL)

1. Stride Convolution

The convolution layer is generally used as first layer in CNN architecture which reduce the feature maps of input image according to convolutional operations. TO make output images as same size of input padding is used in before performing convolution [5] which may cause artifact in the input image. So, deconvolution layer is also designed to make the output image of same size as the input image. By using deconvolution layer, artifacts are removed and reduces the computational overhead.

2. Activation Unit

Rectified Linear Unit (ReLU) are generally used as internal layers of CNN architecture whose main function is to activate the network. In this unit the negative co-efficient values are considered to be zero. The mathematical condition that occurs in this layer is calculated as:

$$f(x) = \begin{Bmatrix} 0\ for\ x < 0 \\ x\ for\ x \geq 0 \end{Bmatrix} \quad (3)$$

Some of the neurons dropped since they don't add to advance entry and don't take part in back propagation. Each time an

information is introduced, the neural system breaks down another design, yet every one of these structures share a common weight. This system decreases the complex adjustments of neurons on the grounds that a neuron can't depend on the presence of some different neurons.

3. Pooling Layer

The pooling layer is used in convolution layer to reduce or shrink the dimension of input data value. This entire calculation in this layer is performed by sliding window termed as kernel with stride level which can estimate the data value collectively for some group of values. This layer can be arranged number of times according to dimension of input image. Sometimes, it can cause time complexity with possibility of overfitting of network.

4. Fully-Connected Layer

Fully connected is considered to be last layer of CNN layers which is required to generate final output or final decision from entire network, i.e. in categories or classes.

5. Loss Function

Loss function measures the difference between output estimation with the help of given ground truth in network through forward propagation. Ordinarily utilized loss function for multiclass order is cross entropy, while mean squared blunder is commonly applied to relapse to ceaseless qualities. A sort of loss function is one of the hyper parameters and should be resolved by the given undertakings.

Recent research has gradually reached out to the structure of misfortune capacity to get an additionally recognizing highlight circulation, which implies the conservativeness of intra-class and the discreteness of between classes at the earliest opportunity. Because of the compact fitting ability of the CNN, these techniques can function admirably and the exactness can be improved. So many researchers have improved the loss function. Due to the benefits of available theory, quick training, & very nice performance, the common cross entropy loss function is very much helpful and used in processing of image. The process of training of neural network is guided by the loss function. To improve the classification of image problem, MSE (L2 loss) and cross entropy loss are very much used.

B. *Proposed Flow Chart*

In this work, the proposed methodology is designed for unmasking the facial region. Many researchers contributed their work for face mask detection after outbreak of COVID-19. In this work, mask detection and face recognition system is developed for noisy facial image. This work used Convolution Neural Network (CNN) for noise removal, mask detection and recognition system.

The Residual Neural Network (ResNet) is modified with leaky ReLU activation layer for training the network with facial features. Residual Neural Network (ResNet) by Kaiming - introduced of a new type of architecture with "skip links" and batch normalization. These skip links are also known as gated units or recurring gated units and have a strong resemblance to the RNNs. This technique will give lower complexity than other CNN networks.

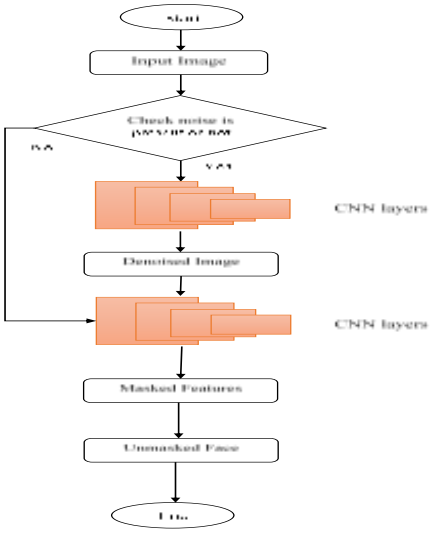

Fig. 2 Proposed Flow Chart

C. *Proposed Algorithm*

Input: Input image, X={x_1, x_2,…..x_n}, Label image, Y={ y_1, y_2,…..y_n }, noise standard deviation (σ), Iteration (itr_1, itr_2)

Output: Output Image ($\hat{X}_{unmask}$)

Procedure: Face Unmasking with noise and without noise

start

If (X=X+ σ)

for epochs 1:N

for iteration 1:itr_1

Conv+LReLU ← FV; (FV1……., FV_{25})

Conv+Clip ← FV_{25};

end for

end for

($\hat{X} = Y - Residue(Y)$)

End if

for iteration 1:itr_2

CNN+LReLU ← $\hat{X}$; (FV_{26})

CNN+LReLU+BN ← FV; (FV_{26}……., FV_{56})

($\hat{X}_{unmask} = Y - Residue(\hat{X})$)

Return → $\hat{X}_{unmask}$

Exit

D. *Leaky ReLU Activation Layer*

In the image classification models, accuracy is improved by replacing parameter free ReLU with a learned activation unit. The LReLU is defined in eq 4. The coefficient a_i is the learnable parameter. Eq 4 can be represented as:

$$P(y_i) = \max(0, y_i) + a_i min(0, y_i) \quad (4)$$

When a_i small and fixed ($a_i = 0.001$). The motivation of LReLU is to avoid zero gradients. However, LReLU has a negligible impact on accuracy as compared with ReLU. In LReLU, a small number of additional parameters are added. The number of extra parameters is equal to the total number of convolution layers, which is negligible when considering the total number of weights. So, this network doesn't have overfitting risk. In a channel-shared variant of LReLU, the coefficient a_i is shared by all channels of one layer. This variant only introduces a single extra parameter into each layer. The network embedded with the LReLU layer can be trained by backpropagation and optimized simultaneously with other CNN layers. The update equation of a_i is derived from the chain rule. The gradient of a_i with respect to loss function is given by chain rule,

$$\frac{\partial C}{\partial a_i} = \sum_y \frac{\partial C}{\partial P(x_i)} \frac{\partial P(y_i)}{\partial a_i} \quad (5)$$

where the loss function is denoted by C. The term $\frac{\partial C}{\partial P(y_i)}$ is the gradient propagated from the deeper layer. The gradient of the activation is given by differentiating eq 6. The summation over x_i is carried out at all positions of the feature maps. In the channel shared variant of the LReLU network, the gradient of $a_i$is given by

$$\frac{\partial C}{\partial a_i} = \sum_i \sum_{x_i} \frac{\partial C}{\partial P(y_i)} \frac{\partial P(y_i)}{\partial a_i} \quad (6)$$

Where, summation over i denotes all channels of the layer. The time complexity introduced by LReLU is negligible in both forward and backward propagation. The a_i is updated by the momentum method which is given by

$$\Delta a_i = \mu \Delta a_i + \varepsilon \frac{\partial C}{\partial a_i} \quad (7)$$

In eq 7, μ = momentum

ε = learning rate.

The l_2 regularisation is not used while updating a_i as weight decay tends to push a_i to zero and therefore it biases LReLU towards ReLU.

V. Result Analysis And Discussions

A. Implementation Details

This chapter comprises with an analytical and numerical description of proposed algorithm for face mask detection which is simulated to obtain the performance of the proposed algorithm. In order to evaluate the performance of proposed algorithm scheme, the proposed algorithm is simulated in following configuration:

Software Requirement

MATLAB-8.3.0 Platform

32/64 bit Windows Operating System

Hardware Requirement

Intel Core i5-3210M CPU @ 2.50GHz

4 GB RAM

512 GB Hard Disk

B. Description of Dataset

There is no publicly available dataset that contains facial image pairs with and without mask object to train the proposed model in a supervised manner. A dataset is prepared with 500 images with and without mask from publicly available CelebFaces Attributes Dataset (CelebA). Another used dataset is Masked Faces in Real-World (MFR) dataset for facial recognition. The dataset is processed in terms of face alignment and image dimensions.

C. Result Analysis

The experimental result is performed and tested on different facial images with different conditions such as normal masked facial images, noised masked images and rotated masked images. All these images, are created from celebA dataset. Figure 3 shows some examples dataset images that are used to verify the effectiveness of the proposed algorithm.

Masked Facial Images

Unmasked Facial Images

Noisy Masked Facial Images

Rotated Masked Facial Images

Fig. 3 Dataset Sample Images used for Training

D. Performance Parameters

In this research work two performance parameters are used for image quality assessment. These parameters are:

1. Peak Signal to Noise Ratio (PSNR)

PSNR represents the degradation of the image with reference images. It is expressed as a decibel scale. Higher the value of PSNR higher the quality of image. PSNR is represented as:

$$PSNR = 10log10(\frac{(X * Y)}{MSE} \quad (8)$$

Where,

X and Y are height and width respectively of the image.

MSE= Mean Square Error between restored image and reference images

2. Structural Similarity Index (SSIM)

One of the quality assessment methodology in digital image processing is structural similarity (SSIM). This index is used to represent the systematic quality assessment of any type of image inputs. SSIM is used to find and measure of similarity between images as well as videos.

This index is completely reference based and determines the noisy or compressed or distorted images with some reference images. SSIM is designed to improve on traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE).

The SSIM index is calculated on various windows of an image. The measure between two windows x and y of common size N×N is:

$$SSIM\ (x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (9)$$

Where, μ_x= mean of x

μ_y = mean of y

σ_x^2 = variance of x

σ_y^2= variance of y

σ_{xy}= co-variance of x and y

c_1 and c_2 are variables to stabilize the division with weak denominator

E. Performance Evaluation

Table below shows the performance of proposed methodology with respect to average PSNR and SSIM indices for masked facial images. Three conditions are considered in this evaluation, i.e. normal masked facial image, noisy masked facial images and rotated masked facial images.

Case-I: This condition is considered that there is no distortion in facial features with masking. In this case face is restored after unmasking and its performance is evaluated in terms of PSNR and SSIM. Table 1 represents the PSNR and the SSIM.

Table 1: PSNR Evaluation of Proposed Methodology for Case I

Input Image	PSNR	SSIM
1	30.66503	0.934148
2	34.86233	0.962005
3	34.23507	0.962096
4	35.70338	0.951554
5	31.22913	0.931742
6	31.89186	0.940194
7	30.09416	0.960055
8	32.81404	0.961616
9	34.1339	0.9598
10	35.53663	0.947617
Average	33.11655	0.951083

Case-II: This condition is considered that there is rotation in facial features with masking. In this case face is restored after unmasking and its performance is evaluated in terms of PSNR and SSIM. Table 2 represents the PSNR and the SSIM.

Table 2: PSNR Evaluation of Proposed Methodology for Case II

Input Image	PSNR	SSIM
1	27.49	0.79
2	30.36	0.88
3	32.35	0.91
4	28.10	0.78
5	29.14	0.84
Average	29.49	0.84

Case-III: This condition is considered that there is distortion in facial features with masking. In this case face is restored after denoising and unmasking and its performance is evaluated in terms of PSNR and SSIM. Table 3 represents the PSNR and the SSIM.

Table 3: PSNR Evaluation of Proposed Methodology for Case III after denoising masked image

Input Image	PSNR	SSIM
1	28.71	0.91
2	31.18	0.92
3	32.36	0.90
4	33.95	0.91
5	33.28	0.92
Average	31.90	0.91

Table 4: Output examples generated by Proposed Methodology

If PSNR value is high then the properties of image are preserved. In this proposed methodology, on an average of 31 PSNR is achieved after enhancement which shows the effectiveness of proposed methodology.

Case-IV: This condition is considered that there is distortion in mask region only. In this case face is restored after denoising and unmasking and its performance is evaluated in terms of PSNR and SSIM. Table 4 represents the PSNR and SSIM. Similarly, table 5 shows some output for case IV.

Table 4: PSNR and SSIM Evaluation of Proposed Methodology for Case IV after unmasking

Input Image	PSNR	SSIM
Average (100 images)	30.67	0.88

Table 5: Output example generated by Proposed Methodology for Case IV

Input Noisy Masked Image	Detected Mask Region	Output Unmasked Image

F. Comparative Performance Evaluation

In this sub-section, comparisons with existing work is performed with proposed ResCNN based face mask detection technique on a static scene. Table 6 and Figure 4-5 represents the comparative performance of proposed methodology as compared to existing methodology.

Table 6: Comparative Performance Evaluation of SSIM and PSNR Values

Techniques	SSIM	PSNR
Implemented	0.88	30
Existing [6]	0.86	26

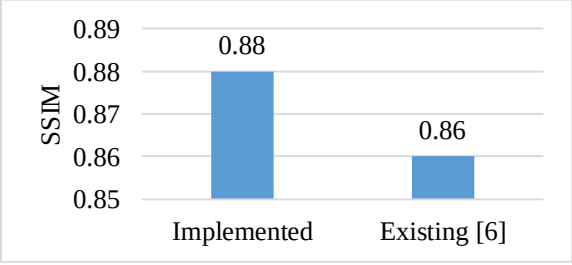

Fig. 4 SSIM Comparative Performance Evaluation

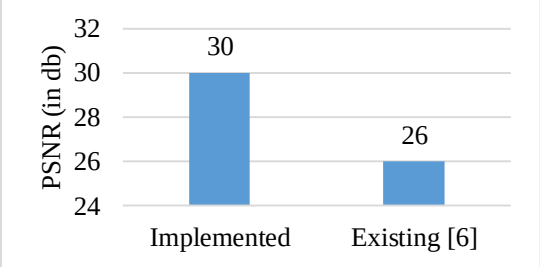

Fig. 5 PSNR Comparative Performance Evaluation

VI. CONCLUSION

In this research work, a face mask detection and unmasking is proposed using residual Convolution Neural Network (CNN). This model is designed for three cases, case I (normal masked faces), case II (rotated masked faces) and case III (noisy masked faces). The simulation result is performed on MATLAB platform in which two datasets is taken for reference i.e. celebA and MFR datset. From these datasets, facial images were taken and masking is performed for training the ResCNN network. The training is performed to determine the mask is either present or not. If mask is present then unmasking is performed and facial regions are restored. The simulation is performed on three cases and achieved PSNR values of ~33, ~29 and ~31 respectively. So, overall PSNR value obtained is approx. 31 and shows about 16% of improvement over existing work. Similarly, obtained SSIM is 0.94, 0.44 and 0.91 respectively. So, overall SSIM is about 0.9 which is approx. 3% improvement over SSIM of existing work. This model has limitation that it requires facial image as input and then perform unmasking. In future this model can be implemented for real time applications such as CCTV at social places to identify an individual among many peoples. This will then require more training and less time consumption for processing.

REFERENCES

[1] Kimball A, Hatfield KM, Arons M, James A, Taylor J, Spicer K, *et al.*, "Public Health – Seattle & King County; CDC COVID-19 Investigation Team. Asymptomatic and presymptomatic SARS-CoV-2 infections in residents of a long-term care skilled nursing facility - King County", *MMWR Morb Mortal Wkly Rep* 2020;69:377–381.

[2] Radonovich LJ Jr, Simberkoff MS, Bessesen MT, Brown AC, Cummin gs DAT, Gaydos CA, "ResPECT Investigators. N95 respirators vs medical masks for preventing influenza among health care personnel: a randomized clinical trial", *JAMA* 2019;322:824–833.

[3] Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, "SARS-CoV-2 viral load in upper respiratory specimens of infected patients", *N Engl J Med* 2020;382:1177–1179.

[4] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks.

[5] Timo Ahonen and Matti Pietikainen, "Face description using Local Binary Patterns: Application to Face Recognition" inIEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 28, No 12, 2006.

[6] N. Ud Din, K. Javed, S. Bae and J. Yi, "A Novel GAN-Based Network for Unmasking of Masked Face," in IEEE Access, vol. 8, pp. 44276-44287, 2020.

[7] G. Dong, W. Huang, W. A. P. Smith, and P. Ren, ''A shadow constrained conditional generative adversarial net for SRTM data

restoration,'' Remote Sens. Environ., vol. 237, Feb. 2020, Art. no. 111602.

[8] S. Li et al., "Multi-angle Head Pose Classification when Wearing the Mask for Face Recognition under the COVID-19 Coronavirus Epidemic," 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), Shenzhen, China, 2020, pp. 1-5.

[9] S. A. Hussain, A.S.A.A. Balushi, A real time face emotion classification and recognition using deep learning model, J. Phys.: Conf. Ser. 1432 (2020) 012087, doi: 10.1088/1742-6596/1432/1/012087

[10] M.K.J. Khan, N. Ud Din, S. Bae, J. Yi, Interactive removal of microphone object in facial images, Electronics 8 (10) (2019) , Art. no. 10, doi: 10.3390/ electronics8101115.