Journal of Multi Disciplinary Engineering Technologies, Vol. 14, Issue 2

Augmented Reality Model for the Virtualisation of the Mask

Jyoti Arora[1, #], Mayank Grover[1], Kartik Aggarwal[1]
[1]*Department of Information Technology, Maharaja Surajmal Institute of Technology, GGSIPU*
[#] Corresponding author, Email: *jyotiarora@msit.in*

***Abstract*- Virtual try on systems for fitting new in shop products have attracted increasing research attention. When the world is in a phase where a virus can spread just through physical contact, automation comes into effect. In this work, we propose Augmented Reality Model for the Virtualisation of the Mask (ARMVM), the image based try-on system for fitting a mask on a person's face. The human face detection seems to be a difficult problem in computer vision due to its high degree of variability which is why the human face can be called a dynamic object. A desirable model will transform the target mask into the most fitting shape. This transformation is done using the geometric augmentation of images. An optimized method of face detection is performed using the deep convolutional neural network along with the extraction of facial features which further help to place the mask accordingly. ARMVM is suitable for a wide range of mask images.**

***Keywords*- face detection; facial features; geometric augmentation; virtual model.**

1. INTRODUCTION

With the occurrence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that started from Wuhan, China from December 2019, has been spread throughout the world. The transmission of the disease is air-borne which affects the people who breathe or come in contact with the infected droplets that remain infectious when suspended in air over a long period of time [1]. Due to airborne transmission, the use of face masks has become ubiquitous for limiting the spread of the virus. A large number of countries have laid down regulations which include wearing of face masks in public areas. Also, the guidelines given by WHO (World Health Organization) have recommended the practice of covering your mouth and nose in order to avoid those droplets which might contain the coronavirus. These droplets are released when an individual coughs, sneezes or speaks in a close proximity [2]. This virus can spread from different entry points i.e. nostrils, mouth and skin touch and two exit points i.e. mouth and nostrils. Masks cover both nostrils and mouth. So, a person whether affected from COVID-19 or not should wear a mask to prevent further spread. The widespread use of masks in Taiwan has linear impact towards successful COVID-19 response [3].

Nowadays "Face masks are the symbol of the pandemic era" a virtual metaphor for the tiny unseen foe which has made the human race to think about their survival chances, lead to an existential crisis, and forced them to adapt to the new normal. While a mask use has been made a priority, there is insufficient information in the public domain about the role masks play and the ability of the mask to protect the wearer from infectious particles [4]. Researchers are working on designing different types of effective mask. R. Panda et al. [5] proposed "Modified anatomical face mask (M-AFM)" for the purpose of an effective alternative of N-95 mask.

There is no proper training, lack of guidance about the effective way to check the suitable mask one can use before purchasing [6]. Some opt for a scarf wrapped around their face; others make do with the surgical masks. The more creative hook colourful homemade varieties around their ears, while the lucky few wear N-95 respirators. The effectiveness of the face mask needs to be taught to people to prevent the virus from spreading. To solve the purpose, this paper proposes an augmented reality model that provides a virtualisation towards an effective way of wearing a mask.

The augmented reality model has been used in many different applications like glasses, clothes, etc. [7,8]. Most of the models use a 3-D object classification CNN model but we have used a 2-D object classification model as it is able to provide more efficiency to the model as explained by [9] for large datasets . Image augmentation is performed to the dataset to provide the model with a more robust dataset. The image augmentation method used is geometric data augmentation as used by [10].

Further the paper is organised in the following sections: Section 2 states the background related to the proposed model. Section 3 describes the details of the proposed model. Section 4 illustrates the experimental results and discussions. The paper is concluded in section 5.

2. Background

2.1 Object Classification

There exist many different methods for 2-D object classification using Convolutional Neural Networks which mainly work on matching different features. The features in 2D image classification are mainly the contours of the object. In this paper we have extracted facial features and the contours of face to perform object classification using a CNN model. The application of the article Robust Real-time Object Detection [5] is used to identify the faces in an image which is further based on Viola Jones [11] algorithm. Viola-Jones algorithm is based on the principle of scanning a sub-window in order to detect faces in given inputs. Shaukat Hayat et al.[9], describes the basic stages of recognition of feature points, classification model and labels used, using a combination of deep learning techniques to improve the efficiency of the system. Similarly, presented by Tae Joon Jun et al.[12] states the use of CNN to classify ECG arrhythmia by using 2D images as the dataset. They presented a model with an average accuracy of 99% using eleven layers.

2.2 Image Augmentation

Image augmentation mostly consists of two methods geometric and photometric [13]. We have opted for a geometric method for image augmentation in this paper. Luke Taylor et al.[10] used the Generic Data Augmentation to improve deep learning. The authors used geometric data augmentation by flipping, rotating and cropping schemes which helped them to provide a more robust dataset for training the model. We have used a similar data augmentation technique to provide robustness to our model. Barret Zoph et al.[14] has explained the use of different augmentation methods and their effect on the robustness of the object detection model.

2.3 Virtual Try-on systems

The virtual try-on systems are mainly based on the augmentation of the image or video frames. In this paper augmented reality technique is used through which a system is developed by integrating virtual elements with real world visualizations. The most related virtual reality system proposed are the virtual reality glasses try on system in the faces are modelled before the system puts virtual glasses on the face. The facial landmarks are extracted using Face Tracker after that geometrical augmentation is performed to provide virtual glasses.

2.4 Multi-layer feed forward (MLF) neural networks [15] offers various useful properties and capabilities.

MLF possesses different functionalities over a single layer neural network. They proved to perform better with different advantages, such as:

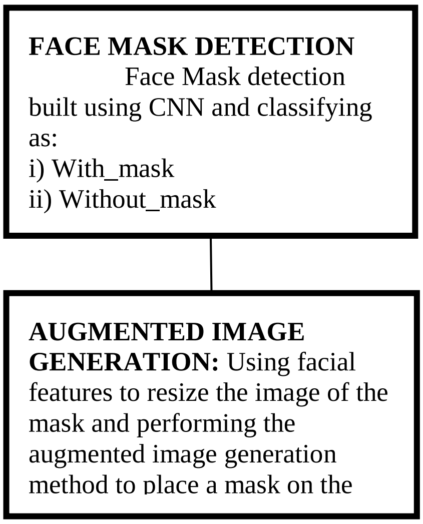
2.4.1 Leaning: MLF are able to adapt without assistance from the user.

2.4.2 Nonlinearity: A neuron is a non-linear device. Consequently, a neural network is itself non-linear.

Nonlinearity is a very important property, particularly, if the relationship between input and output is inherently non-linear.

2.4.3 Input-output mapping: In supervised training, each example consists of a unique input signal and the corresponding desired response. An example picked from the training set is presented to the network, and the weight coefficients are modified so as to minimise the difference between the desired output and the actual response of the network. The training of the network is repeated for many examples in the training set until the network reaches the stable state. Thus, the network learns from the examples by constructing an input-output mapping for the problem.

The proposed model is built for superimposition of face masks in the following stages as shown in Fig 1. First, the face is recognised. Secondly facial features are extracted using *Viola-Jones algorithm* [5]. Third, the mask is superimposed using geometric augmentation technique. Further the model involves a learning process for the precise imposition of the face mask as per the facial features.


FACE MASK DETECTION Face Mask detection built using CNN and classifying as:
i) With_mask
ii) Without_mask

AUGMENTED IMAGE GENERATION: Using facial features to resize the image of the mask and performing the augmented image generation method to place a mask on the

Fig. 1: Workflow of ARMVM

3. METHODOLOGY

Face masks have become the new necessity in the world due to the ongoing pandemic. **Augmented Reality Model for the Virtualisation of the Mask (**ARMVM) aims to build a virtual reality model that can be used to educate people about how to wear masks as well as can be used in the fashion industry to check how different types of masks will look on an individual face.

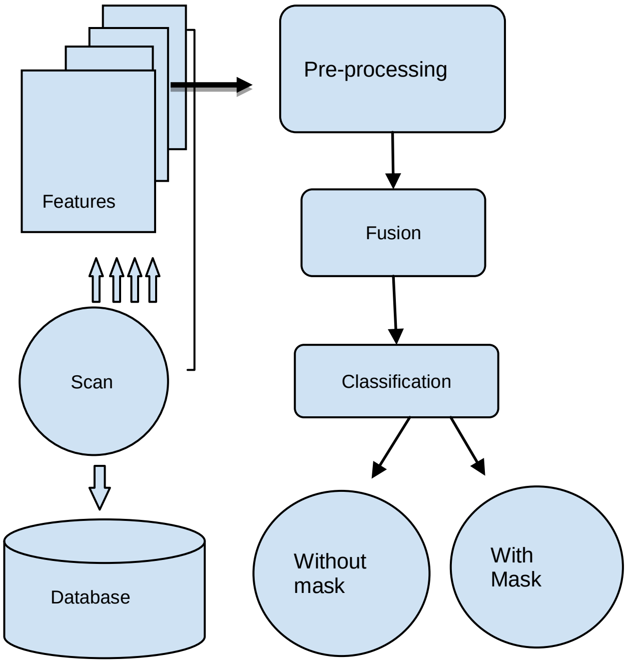


Fig. 2**:** Architecture of Augmented Reality Model of Virtualisation of the Mask

Fig. 2 represents the model which is implemented in the different steps as follows:

3.1 *Scanning the input*: An input is provided to the model, the ARMVM scans the image and detects a face using Robust Real-time Object Detection. Once the face is detected the facial features are extracted which gives the landmarks of the eyes, nose, chin and the facial border.

3.2 *Feature Extraction*: This process involves extraction of face components and their features such as mouth, eyes, and nose etc. from a face detected in the image. The number of Haar-like features is very large in any sub-window, they are far larger than the detected pixels in the image. Face features are extracted using a method based on automatic face segmentation, facial feature extraction and tracking [16].

3.3 *Pre-processing*: The extracted features from the feature extraction stage are then used to calculate: i) distance between the nose and chin. ii) The width of the mask required to fit on the face properly. iii) The angle of rotation of mask image depending on the angle of the face features.

3.4 *Fusion*: The mask image is then blended in the image containing the face using the paste model described in the Basic Operations on Images [17].

3.5 *Classification*: The classification model is built using a Convolutional Neural Network (CNN) for object classification [18] as shown in Fig 3 as the different layers of CNN for the task of object detection. The structure of the CNN consists of the following layers:

TABLE I: Convolutional Neural Network for object classification

INPUT 100 ✖100 image

Layer 1 3✖3 Conv2d,
Activation= Rectified Linear Unit (ReLU)
Pooling = Maxpool2d(2,2)

Layer 2 3✖3 Conv2d, Activation= Rectified Linear Unit (ReLU) Pooling : Maxpool2d(2,2)
Layer 3 Flatten Layer
Layer 4 Dropout Layer with a dropout rate of 0.5
Layer 5 Dense Layer with activation Relu
Layer 6 Dense Layer with activation Softmax

In Layer 1, Conv2d is used to scan images which do not have a moving frame like a video. It is used to create a kernel that is wind with layers input which helps produce a vector of outputs. These vectored outputs are the input to the further layers in the network. It is supported by the Rectified Linear Unit (ReLU) activation function. ReLU is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero.

Flatten Layer [19] is used to flatten and convert the data into a 1-dimensional array for inputting it to the next layer. We flatten the output of the convolutional layers to create a single long feature vector.

Dropout [20] is a technique where some neurons are dropped or ignored during training in order to avoid repeated feature training. They are “dropped-out” randomly. Dense layer represents a fully connected layer, which means that all the individual neurons in a single layer are connected to the neurons present in the next layer.

Max pooling [21] layer is used for operation for 2D spatial data. It down samples the input representation by taking the maximum value over the window defined by pool size for each dimension along the features axis.

4. Mask Detection Using Convolutional Neural Network (CNN)

Mask detection is performed using the Application of Deep Learning for Object Detection [22] model. The output of the Convolutional neural network implemented classifies the image into i) with_mask and ii) without_mask.

4.1 *Object detection algorithm for mask detection*

Input $100X100$ image, is convolved by a $3X3$ kernel, applying 200 filters to each kernel. Rectified Linear Unit (ReLU) function is defined as:

$$f(x)=0 \quad if \quad x<0$$

$$f(x)=x \quad if \quad x>0 \tag{1}$$

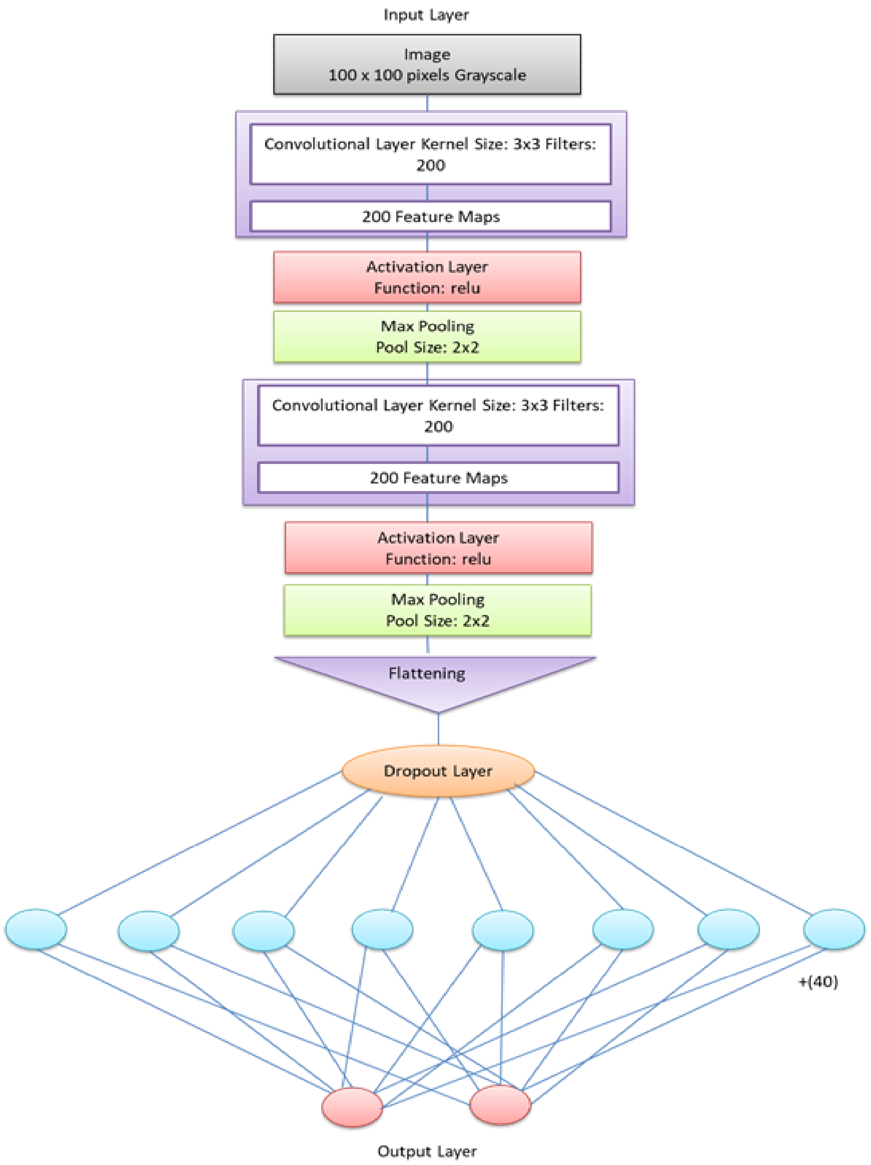


Fig. 3: The Structure of the CNN for object Classification

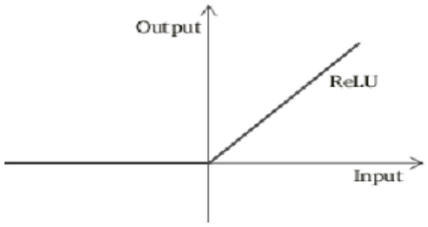


Fig. 4: ReLU function is linear for values greater than zero and zero for values less than zero.

ReLU activation acts as the default, when in the process of development of a CNN and a multilayer perceptron. The pooling layer (Max pooling layer) calculates the maximum value for each patch from the feature map. Extracted features encode the relating presence of various patterns over the different sub windows of the feature maps. It proves to be beneficial to extract more information at the areas having different types of features. Flatten layer is used to transform a two-dimensional matrix of features into a vector that can be fed to the dropout (fully connected) network.

Loss function: A loss function is used to optimize the parameter values in a neural network model. CrossEntropy used here is defined as:

Cross Entropy Loss:

$$-\left(y\log(p)+(1-y)\log(1-p)\right) \qquad (2)$$

Where y is a binary indicator 0 or 1 (without_mask or with_mask) and p - predicted probability observation.

Optimiser: To optimise the loss calculated using the cross entropy loss function is to be optimized by changing the weights and learning rate of the neural network in order to reduce the losses. The optimiser used here is the Adam Optimiser with a learning rate of 0.01.

4.2 Virtual augmented mask Generation

The virtual mask generation model is superimposed after face detected with the help of the geometric augmentation of the mask image. Any image of the mask with a transparent background can be used in this proposed model. Features are extracted using the *face segmentation, facial feature extraction and tracking* [16] from the face detected using *Robust Real-time Object Detection* [5] as shown in Fig. 6. These features are used to perform the calculations to perform the augmentation of the mask image.

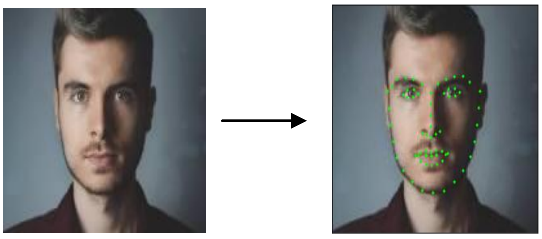


Fig. 5**:** Predicting the features in the given image

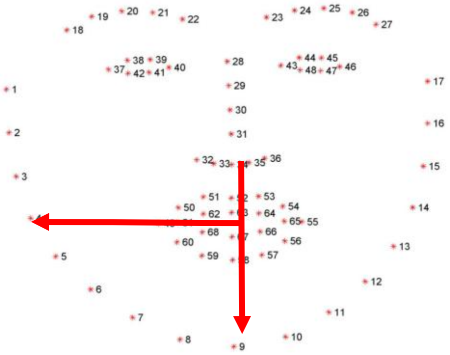


Fig. 6**:** The Facial landmarks pattern detected in a face

i) In Fig 7, let $(x1, y1)$ be the nose centre points from Y, $(x2, y2)$ chin bottom point and (x, y) be the chin left point.

ii) The mask image is split into the left part and the right part as shown in Fig 8.

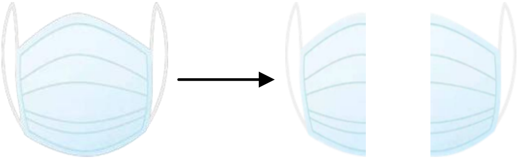


Fig. 7**:** Splitting the left and right half of the mask

iii) Calculating the distance between the left chin point and the line from $(x1, y1)$ to $(x2, y2)$.The distance between a point and a line is calculated as :

$$D = Ax + By + C$$

$$\sqrt{\left((x2-x1)^2+(y2-y1)^2\right)} \qquad (3)$$

where $A = (x - x1)$, $B = (y - y1)$, C is calculated as

$$y = \frac{(y2-y1)x}{(x2-x1)} + C$$

(Equation of Line)

$$C = y(x2 - x1) - x(y2 - y1)$$

$$d = \frac{(x2-x1)x + (y2-y1)y + y(x2-x1) - x(y2-y1)}{\sqrt{\left((x2-x1)^2+(y2-y1)^2\right)}} \qquad (4)$$

Height_req = $y2 - y1$ (5)

(nose point - chin bottom point)

iv) Width_left (d) = width of the mask using the point to line distance formula and upscaling it by a factor of 1.2

Width_right (d) = width of the mask using the point to line distance formula and upscaling it by a factor of 1.2

v) The mask image generated is of width = *Width_left* + *Width_right* , height = *Height_req*

vi) Angle_of_rotation of the new mask

$$\left(\frac{\tan^{-1}(y2-y1)/(x2-x1)}{180}\right)\text{radian} \qquad (6)$$

TABLE II: Pseudo code Of The ARMVM

1. Input : X=image
2. **Pre-processing :** Normalization: x=x/255 Conversion of colour : RGB → Grayscale Resizing : Size=100*100
3. Classification: X (input) → CNN model → y=1(With mask) Or y=0(Without mask)

4. **Augmented Image Generation:**
If(y=0) mask_img=mask_generator (mask)
final_image=masked_face (mask_img,x)

5. Experimental Results And Analysis:

In this section, the performance of the ARMVM is analysed with respect to the accuracy detection of the image with the mask and without the mask. Further mask with the different geometric shapes are augmented on the detected face without the mask.

5.1 Mask detection using Convolutional Neural Network (CNN)

The images of the detected face are analysed for the presence and absence of the mask. The mask detection using CNN classifies the given image into with_mask and without_mask categories with an accuracy of **92.02 %.**

Accuracy is defined as:

$$\frac{TP+FN}{TP+FN+FP+TN} \qquad (7)$$

Where, True Positives (TP): a measure of outcomes in which the model is able to correctly classify the positive class.

True Negatives (TN): a measure of outcomes in which the model is able to correctly classify the negative class.

False Positives (FP): a measure of outcomes in which the model is able to incorrectly classify the positive class.

False Negatives (FN): a measure of outcomes in which the model is able to incorrectly classify the negative class.

Training accuracy is defined as the accuracy of a model on examples it was constructed on.

Validation accuracy is defined as the accuracy of a model on a new set of data. The graph of loss with respect to the number of epochs is obtained as shown in the Fig 8

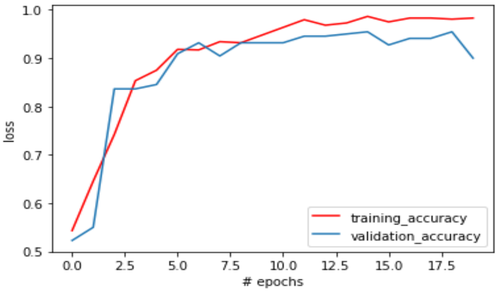


Fig. 8: Training accuracy vs. Validation accuracy

The degree of over fitting is measured by the gap between training and validation accuracy in Fig 8. The larger the gap, the higher the over fitting. The above graph shows that the validation accuracy increases with increase in the number of epochs in the model.

5.2 Virtual augmented mask Generation

Features are extracted using the *face segmentation, facial feature extraction and tracking* [16] from the face detected using

Robust Real-time Object Detection [5]. The visual representations of various types of masks are applied on the detected face as shown in Fig 9.

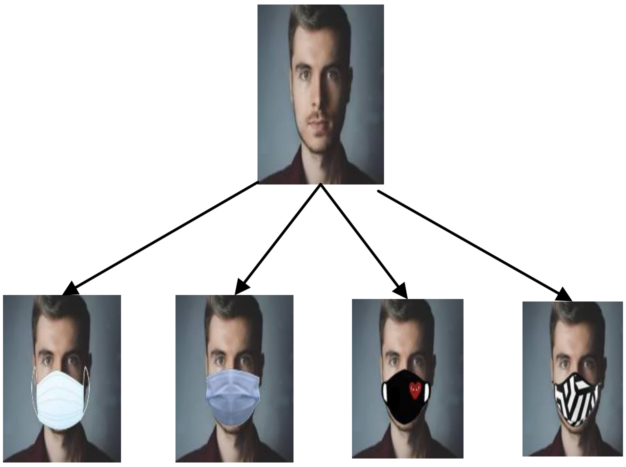


Fig. 9: Samples generated from the ARVRM model

6. CONCLUSION

This paper proposes a model which is used to visualize virtual 3d objects such as masks on the face detected in the input image. ARMVM augments an image with another one to generate a morphed image and perform its optimization using the technique of geometric image augmentation. This is done to adjust the user's facial features using the technique of geometric image augmentation. After optimization, convolutional neural network (CNN) model is used to classify the given image as “with mask” or “without mask”. At the end, a final image is generated showing the virtual mask which is well positioned as well as properly scaled on the user's face.

In future works, we plan to apprehend a dataset with more number of entries for the training and testing part. We would like to consider more attributes to verify our model and add more virtual 3D objects such as sunglasses, beards, hats etc. Along with that, we would like to improve upon the optimization scheme and compare it with our current solution. The goal of this paper is to construct a model which can be used to create an augmented reality interface where users can add 3D virtual objects interactively.

REFERENCES

[1] J. Howard et. al., Face Masks Against Covid-19: An Evidence Review, Medicine and Pharmacology, 2020.

[2] R. Panda, P. Kundra, S. Saigal, D. Hirolli, P. Padhihari, Covid-19 mask : A modified anatomical face mask, Indian J. Anaesth.,64, 2020 S144-5.

[3] H. K. Sra, A. Sandhu, M. Singh, Use of Face Masks in Covid-19, The Indian Journal of Pediatrics, 2020.

[4] Liping Yuan, Zhiyi Qu, Yufeng Zhao, Hongshuai Zhang, Qing Nian, "A Convolutional Neural Network based on Tensorflow for Face Recognition", 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 25-26 March 2017, Chongqing, China.

[5] M. J. Paul Viola, "Robust Real-time Object Detection," International Journal of Computer Vision, pp. 137-154, 2004.

[6] H. A. C. H. a. S. L. Bo Wu, "Fast rotation invariant multi-view face detection based on real Adaboost," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 79-84, 2004.

[7] P. Azevedo, T. O. Dos Santos and E. De Aguiar, "An Augmented Reality Virtual Glasses Try-On System," 2016 XVIII Symposium on Virtual and Augmented Reality (SVR), Gramado, 2016, pp. 1-9, doi: 10.1109/SVR.2016.12.

[8] Wang, B., Zhang, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018). Toward Characteristic-Preserving Image-based Virtual Try-On Network. *ECCV*.

[9] S. Hayat, S. Kun, Z. Tengtao, Y. Yu, T. Tu and Y. Du, "A Deep Learning Framework Using Convolutional Neural Network for Multi-Class Object Recognition," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, 2018, pp. 194-198, doi: 10.1109/ICIVC.2018.8492777.

[10] Taylor, L., & Nitschke, G. (2018). Improving Deep Learning with Generic Data Augmentation. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1542-1547.

[11] Yi-Qing Wang, *An Analysis of the Viola-Jones Face Detection Algorithm*, Image Processing On Line, 4 (2014), pp. 128–148. https://doi.org/10.5201/ipol.2014.104

[12]Jun, T., Nguyen, H., Kang, D., Kim, D., Kim, D., & Kim, Y. (2018). ECG arrhythmia classification using a 2-D convolutional neural network. *ArXiv, abs/1804.06812*.

[13] Shorten, Connor and T. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning.” *Journal of Big Data* 6 (2019): 1-48.

[14] Zoph, B., Cubuk, E., Ghiasi, G., Lin, T., Shlens, J., & Le, Q.V. (2019). Learning Data Augmentation Strategies for Object Detection. *ArXiv, abs/1906.11172*.

[15] Svozil, D., Kvasnicka, V., & Pospíchal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems, 39*, 43-62.

[16] Sobottka, K., & Pitas, I. (1998). A novel method for automatic face segmentation, facial feature extraction and tracking. *Signal Process. Image Commun., 12*, 263-281.

[17] Pajankar, Ashwin. (2017). Basic Operations on Images. 10.1007/978-1-4842-2731-2_4.

[18] Sharma, N., Jain, V., & Mishra, A. (2018). An Analysis Of Convolutional Neural Networks For Image Classification. *Procedia Computer Science, 132*, 377-384.

[19] Culurciello, E., Jin, J., Dundar, A., & Bates, J. (2013). An Analysis of the Connections Between Layers of Deep Neural Networks. *ArXiv, abs/1306.0152*.

[20] “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, *Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan*

[21] A. Giusti, D. C. Cireşan, J. Masci, L. M. Gambardella and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," *2013 IEEE International Conference on Image Processing*, Melbourne, VIC, 2013, pp. 4034-4038, doi: 10.1109/ICIP.2013.6738831.

[22] Pathak, A., Pandey, M., & Rautaray, S. (2018). Application of Deep Learning for Object Detection. *Procedia Computer Science, 132*, 1706-1717.