

PAPER • OPEN ACCESS

Comparative analysis of CNN and Viola-Jones for face mask detection

To cite this article: M SivaKumar *et al* 2021 *J. Phys.: Conf. Ser.* **1916** 012043

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

Comparative analysis of CNN and Viola-Jones for face mask detection

M SivaKumar¹, N Saranprasath¹, N S Sridharan¹ and V Shanmuga Praveen¹

¹Department of Computer Science and Engineering, Sri Krishna College Of Engineering and Technology, Coimbatore, India.
sivakumarm@skcet.ac.in

Abstract. According to the World Health Organization, the Coronavirus (COVID-19) pandemic is causing a worldwide emergency, and one safe way to cover oneself is to wear masks. This pandemic constrained governments everywhere in the world to force lock-downs to avoid the transmission of infection. Reports show that wearing masks at work diminishes the danger of infection. We assemble our model by utilizing the concept of deep neural learning and AI. The dataset comprises pictures with masked faces and non-masked faces. Several computer algorithms are there for face detection. But this analysis centers around two of the most widely recognized procedures: The Viola-Jones algorithm and the Convolution Neural Networks. We will check whether the individual in the image/video wears a mask or not with a CV and Deep neural learning. Not only finding out about face mask detection, but this project also introduced the chance to delve into the field of computer algorithms.

Keywords: Haar-like features, Keras, Tensorflow, Computer Vision, Mask Detection.

1. Introduction

The fast emits of COVID-19 in 2020 urged the World Health Organization to proclaim it a worldwide pandemic. The infection spreads through direct contact and in closely-packed regions. Artificial Intelligence aids us in the battle of Covid-19 from numerous points of view. The habit of wearing a face mask is expanding anywhere. Researchers have claimed that to forestall the transmission of COVID-19, we should put on masks. Several nations imposed rules to encourage their residents to put on the mask. These law-related rules and regulations have been proposed to counter the quick raise in cases and mortality in several nations. The technique for the observation of enormous groups of individuals is hard. Our lives are made simpler by using advancements in technology such as ML and AI to solve many basic issues. For simple human perception, a few procedures are executed with a CV algorithm. From image characterization to video investigation, CV has been demonstrated to be a developmental part of current innovation. With the guide of innovation, 'Work From Home' has subbed our everyday work schedules. Here we show a masked face identification that depends on Computer Vision and deep learning. This is joined with cameras to block the virus transmission by distinguishing individuals who are not wearing a face mask. We utilized two distinct algorithms: The [1]Viola-Jones and [2]Convolutional Neural Networks. A comparison was made between them to decide which algorithm achieved better accuracy with less computational time.

2. Related works



Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

In the recent past, various techniques deployed for face mask detection. It is the process of localizing the object in an frame and classify them. There are many complexities linked with detection algorithms. Various factors affect the performance of model I]Image Quality: An detection of objects needs quality images. Extract the required features from high quality images. The system's reliability is affected by poor picture quality. II] Lighting: Variations in lighting can affect the quality of results. III] Visual Angles: Changes in face angles vary precisely along the camera's optical axis. Types of objection detection techniques are single-stage and multi-stage detectors. The Single Shot Multi-Box Detector (SSD)[3] detects several items in an image with only one shot. First, they extracted the ROI from the image and then the required features from the ROI. RCNN, Fast RCNN, Faster RCNN are the types of Multi-Stage detectors. [4]Faster R-CNN provided better results compared to the previous models. However, it continues to use the Selective Search Algorithm, a time-consuming method that takes about two seconds per image to find objects that aren't appropriate for massive real-time datasets. [5] A model using Deep and classical machine learning based on Pytorch, a machine learning system that is open-source, is proposed for face mask recognition. Train the model using MobileNetV2 architecture after that mask classifier is applied over live video or image. YOLO is a single-stage detector. [6]YOLOv3 divides the input into a grid. From that grid, it will analyze the target object features. YOLOv3 detects real-time objects very fast and has better accuracy than the Faster R-CNN. However, YOLO has the limitation that the frequency of faces or image size is directly proportional to the time taken detecting such faces, suggesting more computing time. An approach focused on features that rely on extracting important features to detect an entity. The concept of analyzing the pixels, edge detection, and greyscale in an image is part of the low-level analysis. Function analysis increases the low-level analysis outcome. Prominent attributes are calculated and help classify possible faces as a result. [7]Histogram of Oriented Gradients(HOG) features is good at describing an object's shape, so it is good at object detection but not good at object shading. Haar-like features are good at describes an object's shades. [8]Haar-like features yield better accuracy than HOG in terms of face detection. Viola-Jones procedure for face detection has an uncompetitive detection speed while detection accuracy is relatively high accuracy. Constructing a classifier cascade that decreases the calculation time while improving the precision of identification. It is an effective technique as it has a low false-positive rate yet has a long preparing time.[9]

3. Convolution Neural Networks (CNN)

Convolutional Neural Networks (CNN or ConvNets) are the most widely used neural networks in image detection and object recognition. Detection and recognizing faces in the real world are some of the areas of CNN. Keras and TensorFlow have been used to create a CNN model to classify the given input image set into a mask without a mask. Our two-step process is to train the model and apply the face mask detector.

3.1. Architecture overview

The proposed model Figure 1 taken the input from the camera or webcam. The initial step in this model is to detect the face in the video/images, which can also track down several items of different sizes at the same time. [10] After identifying the face, model goes into the mask detection part to find out whether the person wearing the mask or not in real-time.

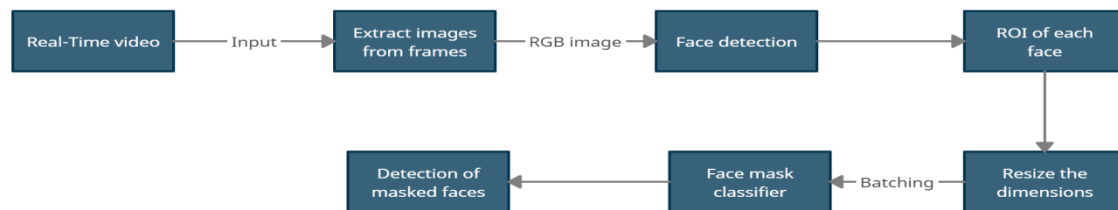


Figure 1. Architecture of CNN

First, extract the image from the frame of the given video. Then convert it into RGB images. Detect the varying faces even if there are multiple faces in a particular frame with the bounding boxes using an already pre-trained model. With this model train the CNN-based classifier. Crop the bounding images to extract ROI of faces for the detection of mask or not. [11] It comes under the face mask classifier process. In proposed system classification used is MobileNetV2 Architecture. When training the model, fine-tuning process Figure 2 takes place. Load this model and compare it with the extract images from the dataset to find out masked faces.

3.1.1. Dataset Collection

Have to train the model by using a lot of masked and without masked images. There is a lot of available, but mainly they are all artificially generated, and for real-world scenarios, it is not the best suit. Gathered an image dataset from the Kaggle's Mask Dataset, some dataset from GitHub, and some own dataset which are all not artificial. [12] Our dataset has about 4200 images of both masked faces and unmasked faces.

3.1.2. Pre-Processing

Outliers identification, missed value treatments, and the elimination of unnecessary or noisy data are the aspect of data pre-processing or data cleaning. The goal is to enhance data of image by changing certain aspects of the image or removing undesired distortion. At the least degree of abstraction, execute these operations on images. This process is because an unnoisy dataset will give better accuracy. Collect all the picture lists from our model's dataset directory and initialize the data list and class photos. It includes resizing to 224 to 224 pixels and converts it into array format by looping through the dataset and processing images. The pixel intensities on the given image scaled in the range $[-1, 1]$. Then, we save our training data in a format of an array (NumPy).

3.1.3. Splitting of data

When training the dataset, use partial original data. Separate the data into two groups: train and test. The image to be trained are found in the training collection. The images used to validate the train data are kept in the testing package. For image training and testing, split the dataset into 3/4 and 1/4.

3.1.4. Data augmentation

Data augmentation means increasing the diversity of images available for training models. Apply on-the-fly data augmentation on our dataset. This is Keras's "image data generator" class implements. This augmentation works at the training time. When training the model, it looks at our information in different variations at every stage of augmentation. Load a group of input images to an "Imagedata generator". It transforms the data in a random operations (rotation, zoom, shear, shift, and flip). Then this functionality return to the function call. The "ImageDataGenerator" does not return both the converted and original data. The class returns only random transform data results.

3.1.5. Fine-Tuning

The architecture of MobileNetV2 must then be prepared for fine-tuning.

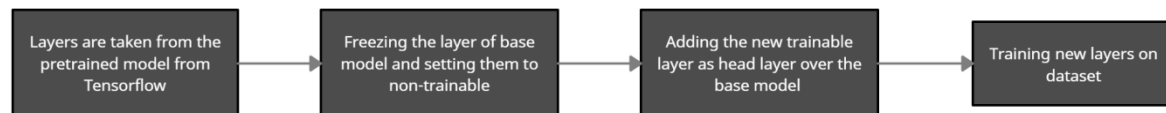


Figure 2. Process of Fine Tuning.

Load the MobileNetV2 with ImageNet's pre-trained weights. The FC layer's head is untouched. Create a model's head and place it on the base model's top using average pooling, Flatten, and dense layers Figure 2.

3.1.5.1. Average pooling

Pooling is the method of reducing the number of feature map dimensions. Pooling is because it reduces the computational power and processing time. There are several pooling techniques in the convolution process. We would be using Average Pooling here.

In the 2D Average pooling, The average value of each block will be taken. For example, the initial block looks like the 3x3 final size will be 2x2 Figure 3.

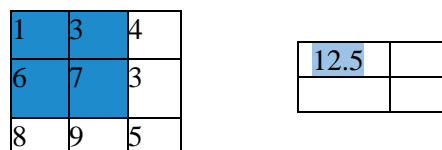


Figure 3. Calculation of Average Pooling 2D.

Formula to calculate the output size is

Output size = (input width/flat pooling factor) x (input tallness/vertical pooling factor) x (input channels).

3.1.5.2. Flattening and Fully Connected layers

Before going into the FC layer, we have to flatten the structure of the data. It is known as a dense layer, which is only a artificial neural network classifier. Furthermore, an ANN classifier needs singular features, much the same as some other classifier. It implies it needs an element vector. Therefore, for the artificial neural network, transform this portion of a convolutional layer into a 1D function vector. A dropout layer spontaneously loses 30 percent of the tensors after the flattening layer to prevent over-fitting. Flattens all its configuration to create a single long vector of features attached to the Fully Connected Layer(final classification model). This model of classification renders images categorized depend upon the characteristics derived from the former layer. And the value of the label of each class is output.

3.1.6. Face Mask Classifier

After Trained the data, load the data into the disk. Get the input data and convolution layers should be extracted without the bounding box parameters falling outside of the image. Extract the ROI from the taken image Figure 4.

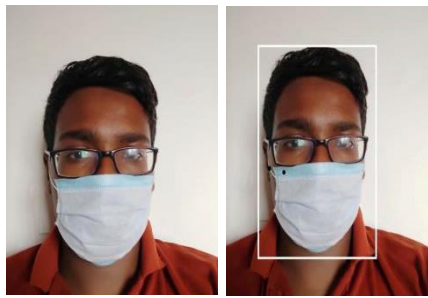


Figure 4. Applying ROI and extract face.

And finally, check the input data with the trained model to detect the masked face.

4. Viola-Jones

The Viola-Jones algorithm is extremely successful at detecting faces and objects in real-time, and its implementation has been exceptionally noticeable. This comprises many steps, namely: 1. Haar characteristics. 2. Integral image. 3. Ada-boost. 4. Cascading Figure 5.

4.1 Diagram of architecture

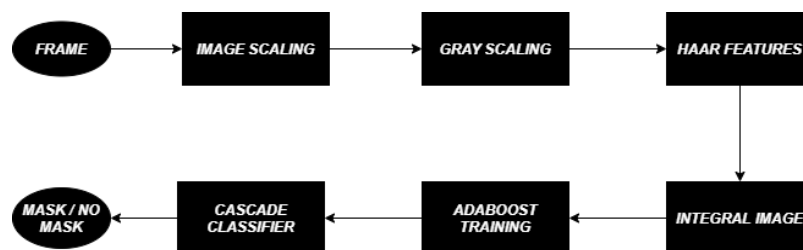


Figure 5. Viola Jones Architecture.

4.1.1 Haar characteristics

There are many benefits of using features instead of pixels, including the fact that a feature-based interface(Haar-features) is considerably faster than a pixel-based system. This function displays a light-side and a dark-side box, which is how the computer decides what the function is Figure 6.

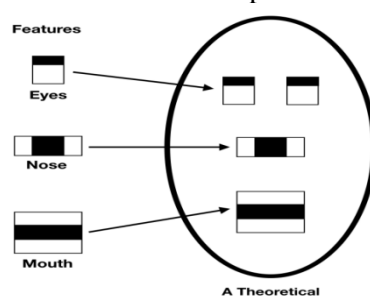


Figure 6. Kinds of features in a Face.

Often, like the tip of an eyebrow, only one side would be lighter than the other. Often the middle section may be shinier than the surrounding boxes, which can be perceived as a mouth. Three sorts of Haar-like features were identified by Viola and Jones:

- Edge-feature
- Line feature
- Four- sided feature

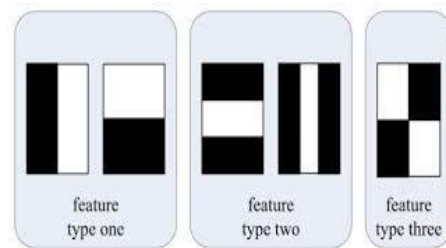


Figure 7. Types of Haar-Characteristics.

When the images are analyzed, each function has its purpose. The Edge feature evaluates the difference in the number of pixels between two rectangular areas. In a center rectangle between two outer rectangles, a line-feature calculates the number subtracted from the sum. Finally, a four-rectangle characteristic measures the contrast between diagonal couples of rectangles. With the assistance of certain characteristics, the significance of the given function is determined Figure 7.

4.1.2 Integral images

The integral value of any particular image point is equivalent to the aggregate of all the image pixels on the left and above a specific point. The value of all the pixels in an area can be calculated with a single traversal, as shown in figure 8, greatly improving computational efficiency.

Let $DAT(x,y)$ be the point value (x,y) and $I(x',y')$ be the grey value of any pixel in the integral image (x',y') , then:

$$DAT(x,y) = I(x',y') \quad x'=x, y'=y$$

$$DAT(x,y) = \sum_{(x' \leq x, y' \leq y)} I(x', y')$$

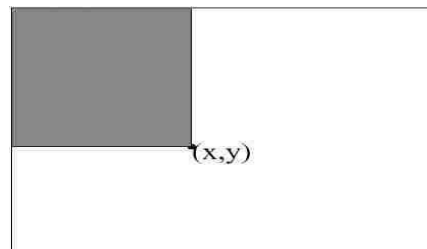


Figure 8. Value of a pixel in an integral image.

One can obtain the following recursion formula traversals from floor to ceiling and from leftmost to rightmost : $DAT(x,y-1) + DAT(x-1,y) + I(x,y) - DAT(x,y-1) + DAT(x-1,y) + DAT(x-1,y-1)$

Likewise, it is possible to calculate the sum of the pixels of any rectangular region in the image, as shown in Figure 9 below. Let the rectangle coordinates be x, y , and the height and width be h, w which is then denoted as a rectangle (x, y, w, h) .

The formula to obtain an integral image:

$$Sum(x, y, w, h) = DAT(x, y) + DAT(x + w, y + h) + I(x, y) - DAT(x, y+h) - DAT(x+w, y) / (x, y).$$

Integral image

A	B	
	1	2
C	3	D
		4

$$\begin{aligned} \text{Sum of all pixels in} \\ D &= 1+4-(2+3) \\ &= A+(A+B+C+D)-(A+C+A+B) \\ &= D \end{aligned}$$


Figure 9. Calculation of the cost of a pixel.

With four array references with integral image, any rectangular sum can be determined. For edge

features, six references are needed. Similarly, for line-features, eight references are required and nine references for four-sided attributes are required Figure 9.

4.1.3 Ada-boosting

We're training the machine to identify these features. The classifier reduces the image to 24 x 24 dimensions and looks for the trained features while determining if something can be classified as a feature. It requires a lot of images of masked and unmasked faces to find out features in various forms. A version of Ada-Boost (Gentle Ada-Boost-GAB) is used in our system to pick a limited collection of characteristics and train the classifier. The Ada-Boost algorithm improves the classification efficiency of a simple algorithm. The loss rate of an effective classifier achieves zero statistically.



$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \dots$$

Figure 10. Equation of Ada-Boosting.

From figure 10 above, f_1 , f_2 , and f_3 are the features and α_1 , α_2 , α_3 are the respective weights of the features. Each of these features is referred to as weak classifier and $F(x)$ is called an effective classifier. We get a effective classifier when we combine two or more weak classifiers. As it continues to be incorporated, it becomes stronger and stronger. Once the algorithm is optimized and can accurately quantify both positives and negatives, it ensures that important front characteristics are not overlooked.

4.1.4 Cascading classifiers

A degenerate decision tree, we call a "cascade," is the overall detection mechanism. We have a 24X24 window over the input image, and we need to find out if the characteristics are included in any of these regions. The task is to discard non-masked faces early and reduce wasting valuable time and calculations. Thus, achieving the speed required for real-time face mask detection. A cascaded system in which the process of identifying faces with a mask is split into several steps.

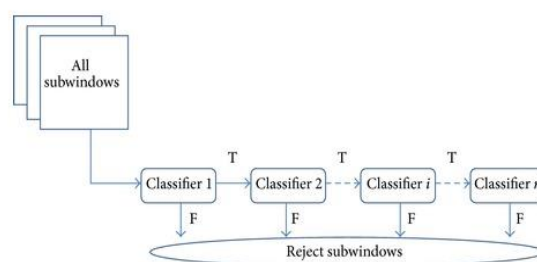


Figure 11. Multi stages of Cascading Classifiers.

The sub-region passes through the finest characteristics. The sub-region that approaches the cascade is assessed in the first phase. If that stage evaluates the sub-region as positive, which means that it thinks it's a masked face, the phase's output is 'possibly' a masked face. It is sent to the succeeding stages of cascade when a sub-region gets a 'possibly', and the loop continues as such until we hit the last stage. If the image is approved by all classifiers, it is labeled as a face with the mask and is shown as an identification to the viewer. This allows us to improve the speed, and if so how? First of all, if a negative verdict is made in the initial phase, then the picture is automatically discarded as not having a masked face. It is also discarded if it effectively passes the initial step but fails the second Figure 11.

5. Result and analysis

A dataset that contains about four-thousand images of Faces with mask (Positive) and without a mask (Negative) as the same dataset for the analysis of both the CNN and Viola Jones approaches. The fed dataset has a good resolution of images but does not have a plane or similar background.

Comparison between CNN and Viola-Jones:

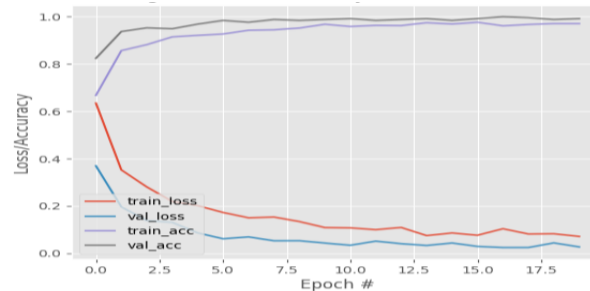


Figure 12. Result after training - CNN

From figure 12, The precision and loss curves of the model file training show the high accuracy and minor signs of overfitting on the final results. The qualified model achieved an accuracy of 98%. CNN uses this model file, and Viola jones's algorithm uses the Haar-cascade features file.

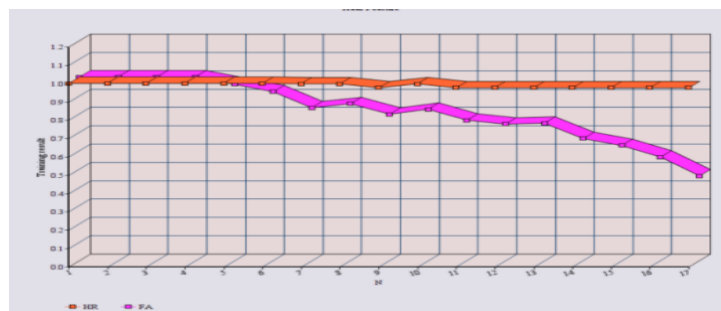


Figure 13. Result for Haar-Cascade Training

From figure 13, Every feature got a Hit rate of 98 % and a False positive rate of only 3 to 5%. In terms of training, both approaches have approximately similar accuracy. In terms of recognizing the unmasked faces (Negative set), both find out the face with the highest precision of 98%. The Viola-Jones approach detects a little faster than that of CNN, but the insignificant difference is seen. In terms of recognizing the masked faces (Positive set), both have the highest precision of 97%. But a significant difference in the speed of detection is observed under different circumstances. So Viola-Jones is ahead of CNN in terms of faster detection Figure 14.

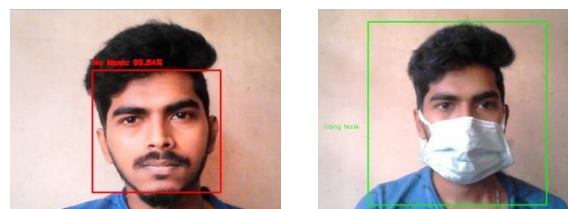


Figure 14. Detection of masks in a real-time.

6. Conclusion

The dataset is of binary categorization which only contains masked and non-masked faces. Training a dataset that has many classifications is quite difficult in Haar-Cascade since it takes extremely more time than CNN. Both have a drawback of identifying the faces under varying illumination. Very

abrupt changes caused non-detection of faces in CNN. Viola-Jones continues to find out them but with an increase in false positives. Haar-Classifiers have the advantage of doing both identification and classification at the same time, obviating the need for a separate detection algorithm that extracts the requisite information from the image and feeds them into a classifier. This decreases computational time, but similarities between the entities and the environment, losses in detail can occur. Despite the training time, Viola-Jones has precisely good accuracy, less computational time, faster classification of images, and easier when it comes implementation.

References

- [1] Paul Viola, Michael Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, Accepted Conference on computer vision and pattern recognition, 2001.
- [2] M. Coşkun, A. Uçar, Ö. Yildirim and Y. Demir, Face recognition based on convolutional neural network, 2017 International Conference on Modern Electrical and Energy Systems (MEES), Kremenchuk, Ukraine, 2017.
- [3] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. International, Conference on Computer Vision & Pattern Recognition (CVPR '05), Jun 2005.
- [4] Redmon, Joseph, You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. arXiv:1506.02640. Bibcode:2015arXiv150602640R, 2016.
- [5] Wei Liu, Dragomir Anguelov, Dumitru Erhan. SSD: Single shot multibox detector. Computer Vision – ECCV 2016. European Conference on Computer Vision, 2016.
- [6] M. Suganya and H. Anandakumar, Handover based spectrum allocation in cognitive radio networks, 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), Dec. 2013.doi:10.1109/icgce.2013.6823431. doi:10.4018/978-1-5225-5246-8.ch012
- [7] Haldorai and A. Ramu, An Intelligent-Based Wavelet Classifier for Accurate Prediction of Breast Cancer, Intelligent Multidimensional Data and Image Processing, pp. 306–319.
- [8] C Rahmad, R A Asmara, D R H Putra, I Dharma, H Darmono, I Muhiqqin, Comparison of Viola-Jones Haar Cascade Classifier and Histogram of Oriented Gradients (HOG) for face detection, IOP Conference Series Materials Science and Engineering, 2020.
- [9] Ssvr Kumar Addagarla, G Kalyan Chakravarthi, P Anitha, Real Time Multi-Scale Facial Mask Detection and Classification Using Deep Transfer Learning Techniques, International Journal of Advanced Trends in Computer Science and Engineering 9(4):4402-4408, September 2020.
- [10] Mr.Jaspreet Kaur , mr. Anand Sharma - Performance Analysis of Face Detection by using Viola-Jones algorithm, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, 2017.
- [11] H. Deshpande, A. Singh, H. Herunde - Comparative Analysis on YOLO Object Detection with OpenCV, International Journal of Research in Industrial Engineering, 2020.
- [12] M. R. Bhuiyan, S. A. Khushbu and M. S. Islam, A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3, 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020.