ROBUST FACIAL RECOGNITION UNDER OCCLUSION

MAJOR-II REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE OF

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING

Submitted By
DEEPANSHU JAYANT
2K19/CSE/06

under the guidance of

Dr. ARUNA BHAT
(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042
September 2021

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Deepanshu Jayant, roll no. 2K19/CSE/06, student of M.Tech (Computer Science & Engineering), hereby declare that the Major-I report titled **Robust Facial Recognition under Occlusion** which is being submitted to Delhi Technological University, Delhi, in partial fulfilment for the requirements of the award of degree of Master of Technology in Computer Science and Engineering is a bonafide report of the work carried out by me. The material contained in this Report has not been submitted at any other University or Institution for the award of any degree.

Place: Delhi

Deepanshu Jayant

(2K19/CSE/06)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I, hereby certify that the Major-I report titled **Robust Facial Recognition under Occlusion** which is submitted by Deepanshu Jayant, roll no. 2K19/CSE/06, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology in Computer Science and Engineering is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree to this University or elsewhere.

Place: Delhi

Dr. Aruna Bhat
Associate Professor
Department of CSE, DTU

ACKNOWLEDGEMENT

I am extremely grateful to my project guide, **Dr. Aruna Bhat,** Associate Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for the extensive support and encouragement she provided.

I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout by giving new ideas, providing necessary information and pushing me forward to complete the work.

Deepanshu Jayant

2K19/CSE/06

ABSTRACT

The face detecting methodologies and frameworks are getting degraded due to outbreak of novel coronavirus, which resulted in the rise of face masks on all human. The presence of face masks on facial landmarks makes them more occluded and it gets difficult to detect the face properly as it is covering most of the important facial features such as nose and lips. Occlusion in face detection is defined by the angle of the face, image illumination, shadows, etc., but now face mask is a very crucial element to cause high level of occlusion. Frameworks sensitive to this occlusion problem fails to perform efficiently and produces inaccurate experimental results.

To overcome this problem an improved and efficient CNN framework is introduced which performs notably better under occluded conditions. The introduced framework processes by using three separate layers of CNN models such as P-net, R-net and O-net for face detection with low computational time and complexity. MTCNN processes by resizing the images to detect different sizes of faces and uses and feeds the scanned images between all the neural networks by performing non-maximum suppression to eliminate the false-positive cases at every step to increase the accuracy and speed. The testing and generation of results are produced using MaskedFace-Net [22] and FFHQ [20] datasets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Introduction

In today's world with the rise of novel coronavirus and rising difficulties with facial detection due to presence of the face mask requires an improved recognition and more robust detection technique to accurately find the face under the presence of occluded conditions in real-time situations. The face mask covers almost half of the face which hides the important face features which limits the face detecting framework from detecting the face properly and produces negative results.

The shape of facial objects and their related changes are arranged under geometric features. Face detection and alignment are essential to many face applications, such as face recognition and facial expression analysis. However, the large visual variations of faces caused by occlusion impose great challenges for these tasks in real world applications.

Face detecting techniques or frameworks which are sensitive to occlusion fails to deliver the desired output and also fails to complete the desired task of the face detection. Presently, face detection practices hold an important role in developing technologies and also security purposes. Due to which, the problem risen by the occluded images or real-time input needs to be eliminated.

Therefore, a new framework is proposed using the advantages of the CNNs to make the algorithm notably efficient and achieve high performance results. Multi-Task Cascaded Convolutional Neural Networks operates by dividing the face detection tasks into three stages. These three stages are defined as P-Net, R-Net and O-Net.
Previous frameworks for the face detection task such as HOG, LBP, RPCA, ODN and RAN performs very well but also has limitations towards occlusion

caused by face masks. While MTCNN detects the face from occluded images by refining the sample data via three layers and accurately calculates the facial features locations to detect the presence of face mask.

1.2 Research Questions and Objectives

1.2.1 Research Questions

- Can the performance of face detection be improved under occlusion due to face mask?
- Can we apply the MTCNN framework for face mask detection accurately?
- Can MTCNN provide better optimization and speed than other CNN models?

1.2.2 Research Objectives

- To perform robust face detection invariant to occlusion due to face mask.
- To perform face mask detection accurately using MTCNN
- To reduce the computational time of the face detection process taken by traditional CNN models.

1.3 Problem Statement

One of the most famous technique to detect facial object is Viola-Jones Algorithm, which works by utilizing AdaBoost and Haar-Like features to train cascaded classifiers, which results in good performance with real-time efficiency. However, with current scenarios in which face masks creates high occlusion, it's performance and accuracy may get degraded. Therefore, use of CNN is opted and it has produced very significant results but this technique is very time costly which limits it for the more robust recognition.

Most of the face detection methodologies provides effective results but also carries some kind of limitation with them, making it difficult to read the facial objects in the real-time input data or real-time surveillance under occlusion. The use of Convolution Neural Networks has been a great tool for the face detection with the help of deep learning techniques. Deep learning enables a framework to obtain high and accurate response in facial regions. However, CNNs possess a bit complex structure which makes the detection process a time costly operation. This deep CNN method is very much useful for training and research purposes but degrades the performance parameters in the case of real-time occluded subjects.

To overcome this limitation a new framework has been introduced using unified cascaded CNNs by multi-task learning. This new framework is called as Multi-Task Cascaded Convolutional Neural Networks. This technique creates a more powerful CNN by dividing the tasks into three different stages and produces very refined results with facial landmarks positions and the performance of the algorithm is comparatively very much efficient and accurate.

CHAPTER 2: BACKGROUND

2.1 Deep Learning

Deep learning is part of a larger family of techniques of machine learning, focusing on artificial neural networks with feature learning. Learning is categorized into three categories: supervised, unsupervised or semi-supervised. Supervised learning[Fig.1] requires a fully labelled dataset. And unsupervised learning does not require labelled data, but it performs poorly as compared to supervised learning. Semi-supervised learning utilizes the combination of some labelled and some unlabelled data.

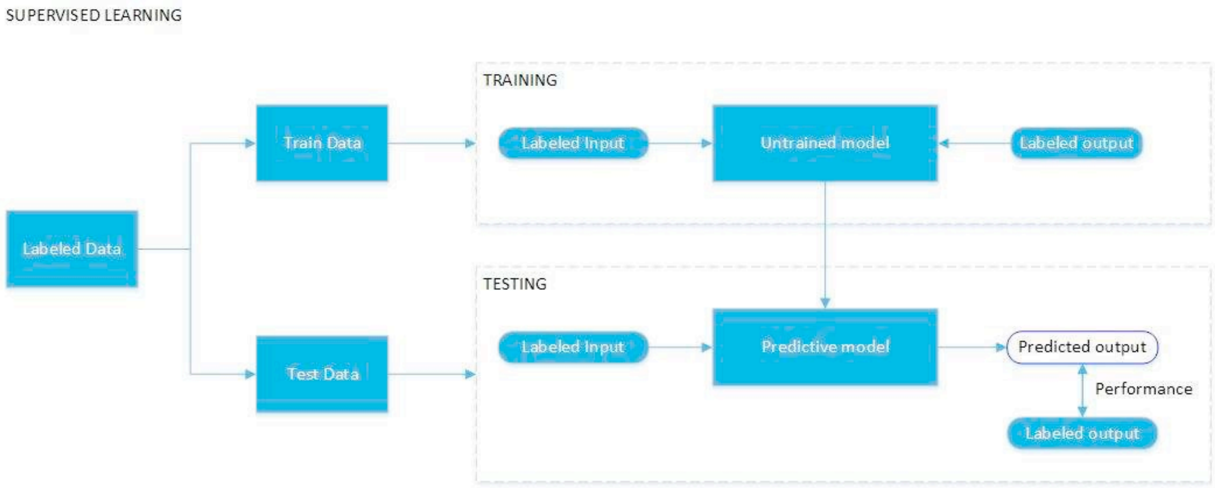


Fig.1: Supervised learning [11]

Various deep learning architectures like convolutional neural networks (CNN) and recurrent neural networks (RNN) have been implemented in areas like voice recognition, natural language processing, computer vision, medical image processing, face detection etc. and have generated competitive results.

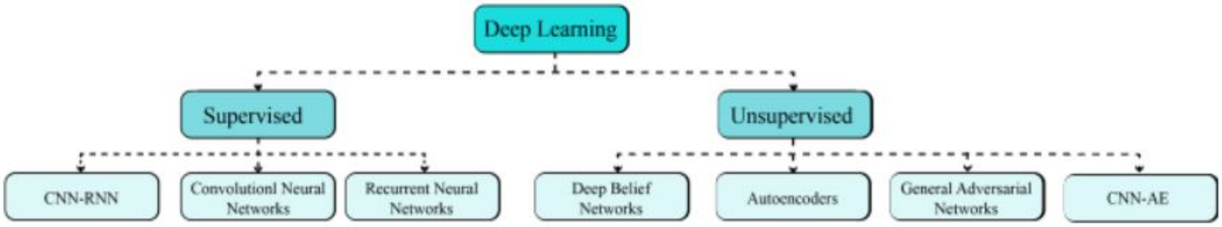

Deep Learning
Supervised
Unsupervised
CNN-RNN
Convolutionl Neural Networks
Recurrent Neural Networks
Deep Belief Networks
Autoencoders
General Adversarial Networks
CNN-AE

Fig.2: Various deep learning methods [11]

2.2 Convolutional Neural Network

Convolutional Neural Networks (CNN): A convolutional neural network is a branch of deep neural networks in deep learning, most widely applied to the analysis of image processing. The multiple layers of a convolutional neural network are specified as the input layer, the convolution layer, the pooling layer and the fully connected layer i.e. the dense layer. The input layer takes the image as an input. Convolution layer produces the output based on its kernel or filter values (i.e. feature extractors), which is fed as input to the layers that follow. The pooling layer is used for dimensionality reduction and to speed up the computation.

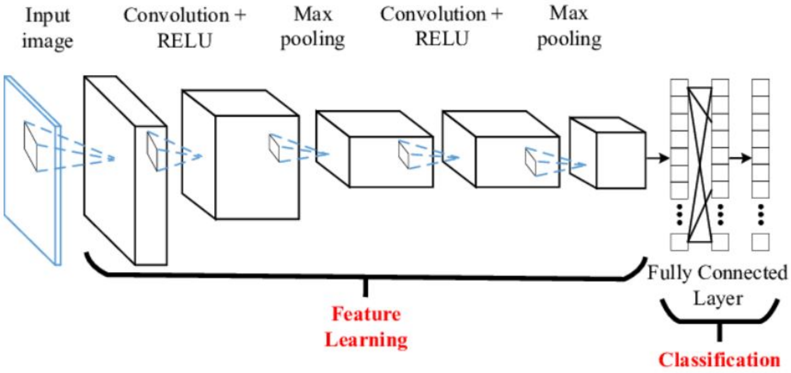


Fig.3: CNN [11]

Transfer learning: Transfer learning is a machine learning research problem that relies on information storage, acquired during solving one problem statement and applying it to a separate but similar issue.

2.3 Multi-task Cascaded Convolutional Neural Networks (MTCNN)

Mtcnn approach works is divided into three stages, which are P-Net, R-Net and O-Net. Where P-Net is defined as Personal Network, R-Net is defined as Refine Network and last stage O-Net is defined as Output Network.

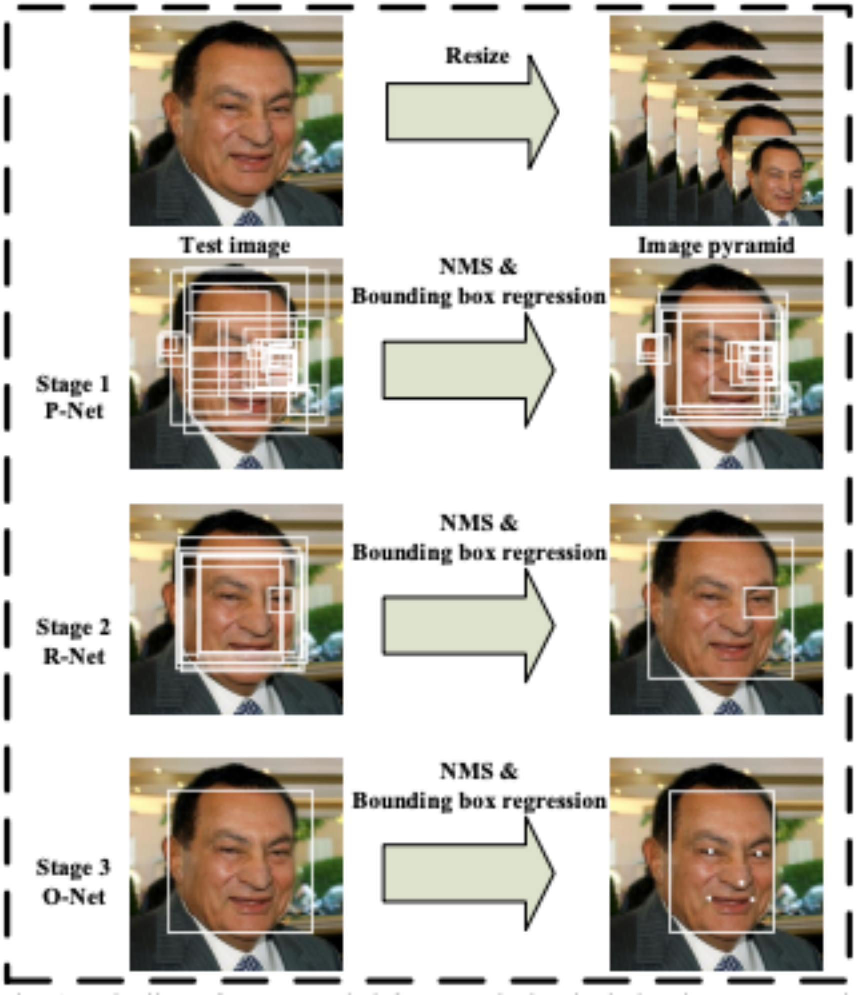


Fig.4 Working layers of MTCNN Framework [1]

2.3.1 Stage 1 (P-Net): It exploits a convolutional network, i.e., Personal Network, which stores the faces and face detecting bounding-box regression values and then the images are calibrated according to these obtained values by avoiding the images with low confidence and performs non-maximum suppression for all kernels to reshape the bounding-boxes into squares.

2.3.2 Stage 2 (R-Net): All the obtained and calibrated images are then pass to another CNN, i.e., Refine Network which refines the by padding out the out-of-bound boxes on a large number and then re-calibrates the images and resize them into 24x24 pixels and normalize the values between and 1 and -1, and conducts non-maximum suppression (NMS) to eliminate the overlapping of the bounding boxes. The result produced by the R-Net has the refined images which has more confidence.

2.3.3 Stage 3 (O-Net): This step while working similar to the R-Net, but strictly focuses on the facial landmarks under precise supervision provides the images with result of landmarks positions. In this step, the images are re-scaled to 48x48 pixels. The O-Net produces 3 outputs such as bounding box, facial landmarks and confidence level of each box and gets rid of boxes with low level of confidence.

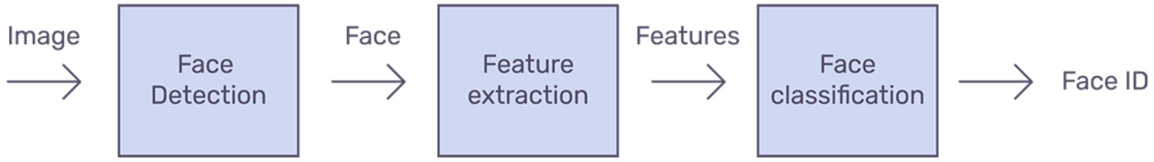


Fig.5. Face Detection Cascaded Workflow [1]

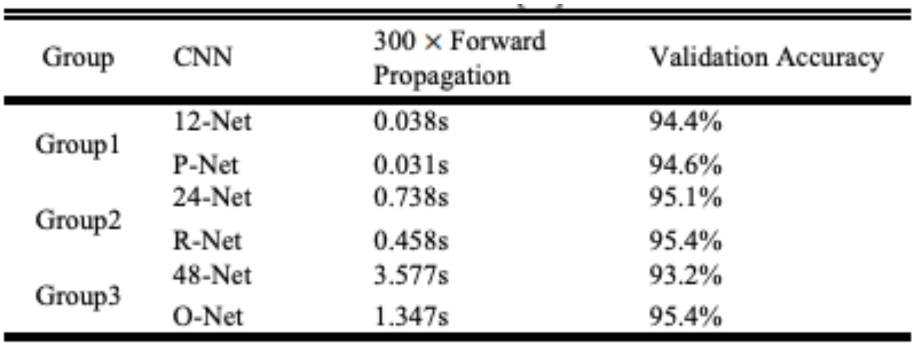
The MTCNN framework achieves high accuracy with less time as its low computation complexity results in faster run-time. The framework doesn't starts locating the facial landmarks from the initial stage rather it calibrates the image to reduce the detection area and then locates the positions of landmarks.

As it also uses deep neural network, it produces very high accuracy because of the 3 networks with multiple layers, which allows the framework to work under high precision. Also, this framework employs a large image pyramid to detect both small and large faces in the image.

CHAPTER 3: RELATED WORK

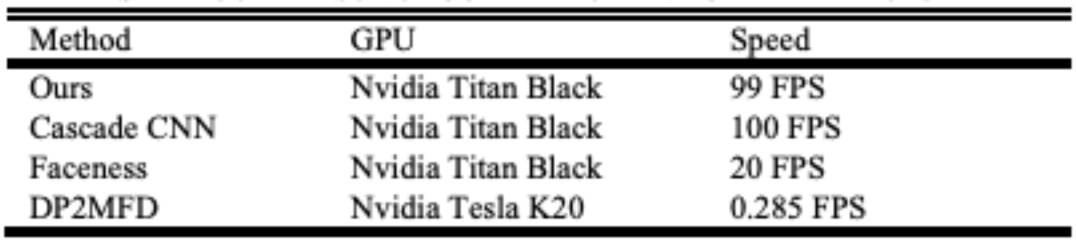
[1] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”

- This research article focused on the accurate detection of face and face alignment under various constraints such as illumination, pose, angle, etc. The paper proposes a new framework by exploiting the inherent correlation between detection and alignment. The study solely focuses on the deep learning technologies to achieve a certain boost and accuracy in the performance of the face detecting practices. The use CNN architecture is given priority here by diving the whole procedure via three different layers of model to get refined and absolute results.
- The article uses two datasets, named, WIDER FACE and FDDB benchmarks for face detection and AFLW benchmarks for face alignment and the method observes superior accuracy against other face detecting technologies.
- Speed comparisons of CNN models and results are also shown by the authors against other present models by two tables shown below:



Group	CNN	300 × Forward Propagation	Validation Accuracy
Group1	12-Net	0.038s	94.4%
	P-Net	0.031s	94.6%
Group2	24-Net	0.738s	95.1%
	R-Net	0.458s	95.4%
Group3	48-Net	3.577s	93.2%
	O-Net	1.347s	95.4%

Table.1 CNN Performance Comparison [1]



Method	GPU	Speed
Ours	Nvidia Titan Black	99 FPS
Cascade CNN	Nvidia Titan Black	100 FPS
Faceness	Nvidia Titan Black	20 FPS
DP2MFD	Nvidia Tesla K20	0.285 FPS

Table.2 GPU Performance Comparison [1]

[2] Xavier P. Burgos-Artizzu and Pietro Perona. “Robust face landmark estimation under occlusion.”

- The research paper proposes a novel method, named, Robust Cascaded Pose Regression (RCPR) to overcome the limitation caused by the robustness of real-time face detection due to different face shape, pose, expression, etc. and locate the facial landmarks positions accurately regardless of the robust conditions. The study uses popular datasets such as LFPW, LFW and HELEN and observes the reduction in failure cases by 50% on all used face images datasets.
- This study is solely focused on improving the performance of Cascaded Pose Regression (CPR) by the novel approach which improves the method by increasing its capability towards robustness caused by occlusions and large shape varaitions.

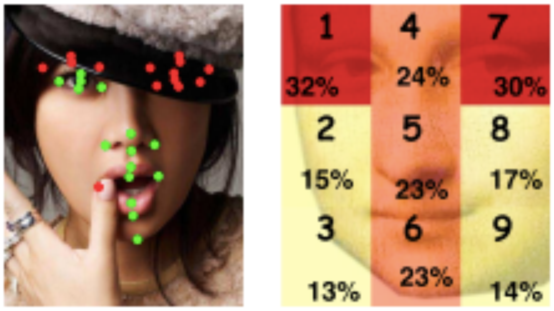


Fig.6 Landmark Estimation [2]

- The paper produces very accurate and refined results of the tests carried out by the proposed method against traditional Cascaded Pose Regression.

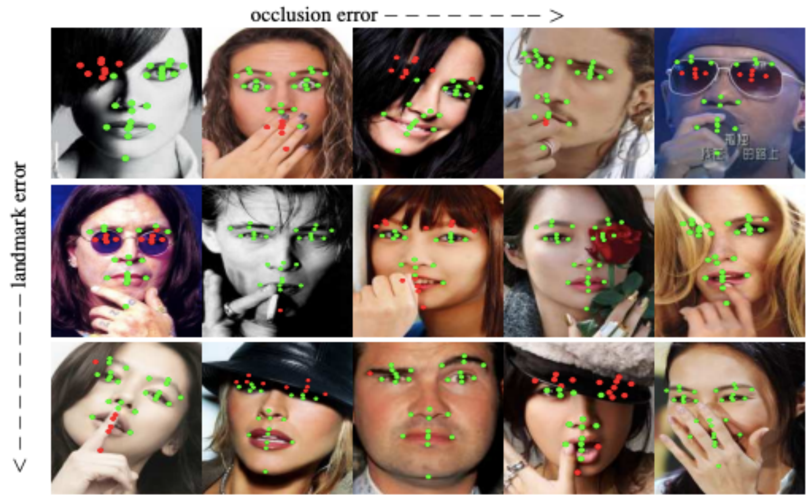

occlusion error - - - - - - - - >
landmark error - - - - - - >

Fig.7 Grouped Estimation [2]

[3] X. Zhu, and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild"

- The study proposes a unified model for face detection and poses , landmarks estimation in the real-world scenarios. The proposes model renders the face landmarks using a tree-based SVM. The study shows a brief comparison between their and AAM model as shown,

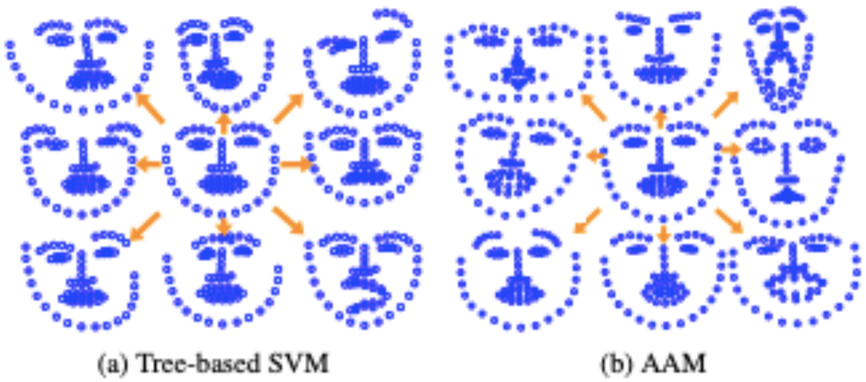

(a) Tree-based SVM
(b) AAM

Fig.8 Landmark Localization [4]

- This study uses face images datasets like CMU-MultiPIE and AFW. The model does the face detection, pose estimation and landmark localization on all available images and generates the results accordingly. The results of this model in terms of precision and accuracy against other compared models is notably better and efficient.

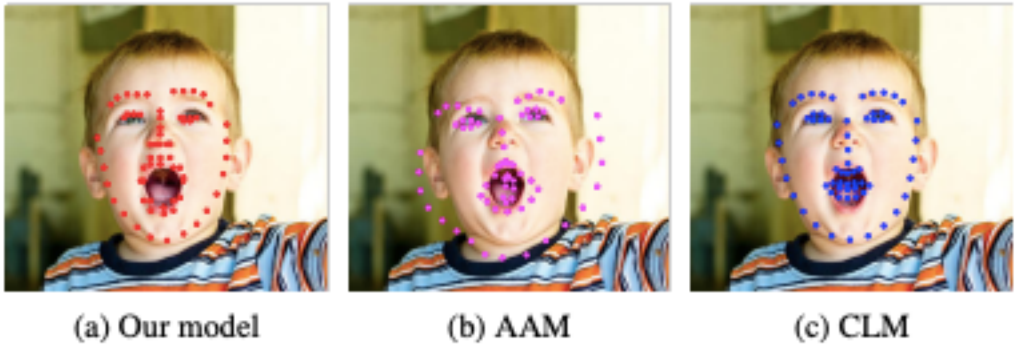
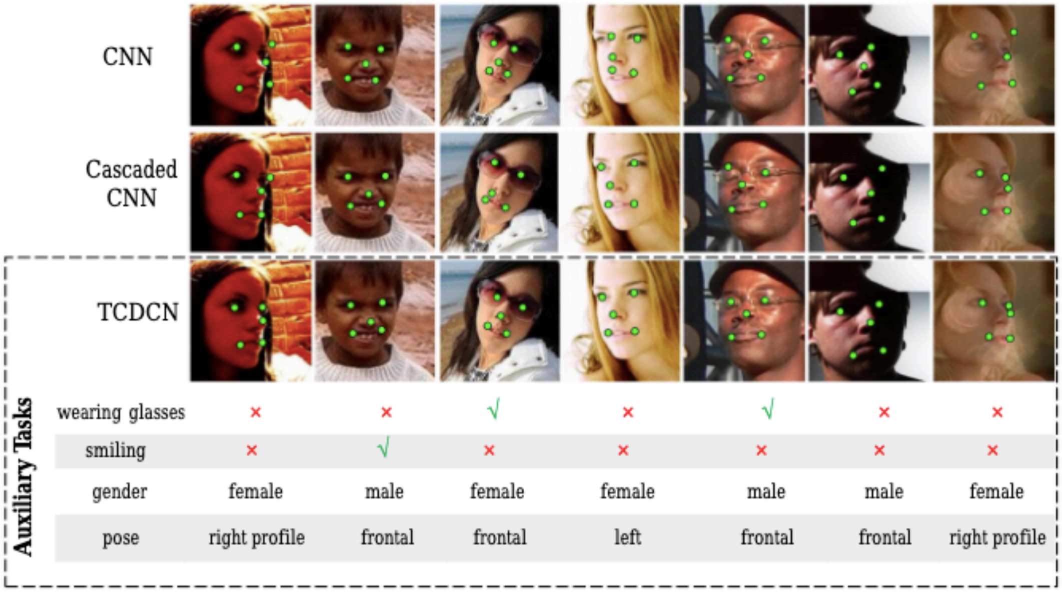


Fig.9 Accuracy Comparison [4]

[4] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning"

- The research study focuses on the facial landmark localization over various set of face images. The study uses the dataset, known as Multi-Task Facial Landmark (MTFL) dataset. It introduces a new model, named as, Tasks-Constrained Deep Convolutional Network (TCDCN). It also shows that the proposed method notably outperforms other methods in the case of occlusion and also offers low computational complexity.
- The method focuses on classifying four major properties of the face images such as, pose, gender, expression and occlusion. The performance of the method is shown below with the comparison of other CNN models



wearing glasses	×	×	√	×	√	×	×
smiling	×	√	×	×	×	×	×
gender	female	male	female	female	male	male	female
pose	right profile	frontal	frontal	left	frontal	frontal	right profile

Fig.10 TCDCN Results [7]

- The study generates the final results by comparing the achieved result values by present CNN models and it is observed that the proposed model has performed better and accurately than other models. Also it proves that the current model has less computational time , hence making it a light-weight CNN model.

[5] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution"

- The research paper performs the face recognition on the real-world crowd images which are often not available in desired constraints. The proposed method eliminates the robustness caused by the noise data collected via crowd images. The study demonstrated the use of Deep Convolutional Neural Networks (DCNN) to recognise the facial expressions. The main focus of the study is to compare different approaches to generate the best results from the noisy labels. The dataset used in this study is known as FER+ dataset.
- The model uses a custom VGG13 network to generate accurate results from four different schemes for a detailed comparison.

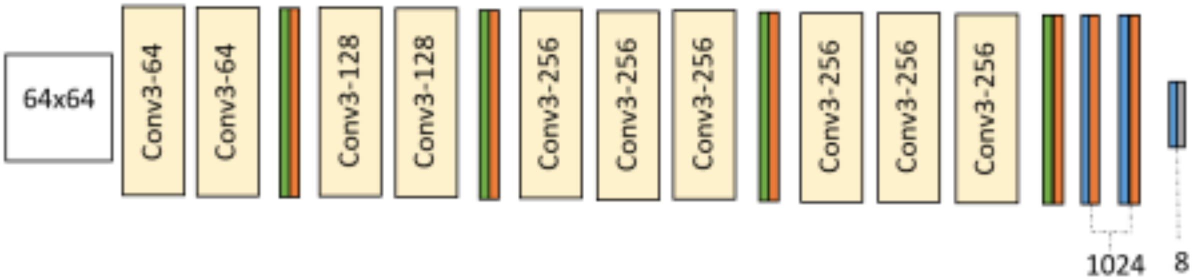
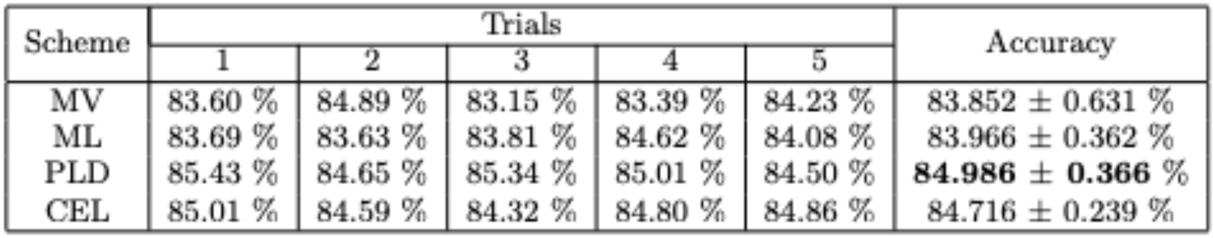


Fig.11 VGG13 Network [9]

- The research article generates a comparative result by comparing the data generated by each different scheme used for the training and testing purposes. The different schemes are named as, (i) Majority Voting, (ii) Multi-Label Learning, (iii) Probabilistic Label Drawing & (iv) Cross-Entropy Loss



Scheme	Trials					Accuracy
	1	2	3	4	5	
MV	83.60 %	84.89 %	83.15 %	83.39 %	84.23 %	83.852 ± 0.631 %
ML	83.69 %	83.63 %	83.81 %	84.62 %	84.08 %	83.966 ± 0.362 %
PLD	85.43 %	84.65 %	85.34 %	85.01 %	84.50 %	**84.986 ± 0.366** %
CEL	85.01 %	84.59 %	84.32 %	84.80 %	84.86 %	84.716 ± 0.239 %

Table.4 Scheme Comparison [9]

CHAPTER 4: METHODOLOGY

4.1 Datasets

The main datasets used for the training and testing purposes for the collection of results produced by the introduced framework, i.e., Multi-task Cascaded Convolutional Neural Networks are:

1. **MASKEDFACE-NET [22]** – This particular dataset provides the multiple types of face detection datasets such as CMFD and IMFD.
2. **FFHQ [20]** – It consists of more than 50,000 png face images and contains variation in terms of several factor such as age, image background, illumination, etc.

4.2 Pre-trained Models

- **VGGFace Model**
    - This model is used by integrating deep CNNs for the tasks of face recognition such as identification and verification. This model is trained with a very large face images dataset. It is also evaluated on the benchmark face datasets. Resulting in efficient and generalized face features generation.

- **VGGFace2 Model**
    - VGGFace2 model is the more efficient and effective than the VGGFace model as it is trained with much larger datasets than the other to evaluate more effective face recognition models.
    - The VGGFace2 model is trained with models like ResNet-50 and SqueezeNet-ResNet-50 models. While SqueezeNet based model offers a very great performance for the training of the VGGFace2 model.

- **CASIA-Webface**
  - The database consists of over 10,000 face images collected by the Institute of Automation (CASIA). It is used for scientific research.
  - The webface dataset is used for various face verification and identification tasks. It contains over 500,000 face images.

4.3 Training

To train the CNN detectors, the three main tasks can be classified as face/non-face detection, bounding-box evaluation and facial landmarks positioning.

4.3.1 Face Classification:

$$L_i^{det} = -(y_i^{det}\log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \qquad (1)$$

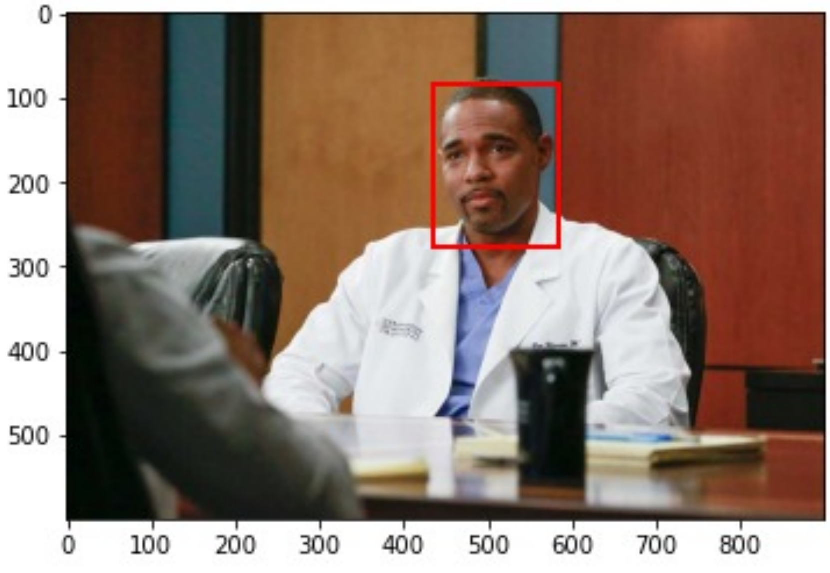


Fig.12 Face Classification

The formula is generated as a two-class classification problem, where p is probability of sample being a face and y denotes ground-truth level.

4.3.2 Bounding-box Evaluation: For every sample images processed by the network, a bounding box is generated with the coordinates values and the level of confidence.

$$L_i^{box} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2 \tag{2}$$

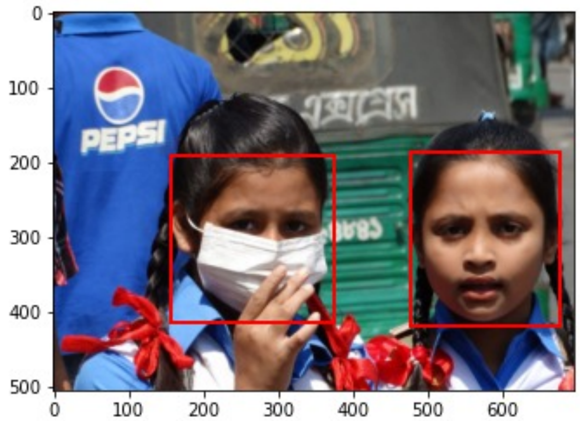


Fig.13. Bounding Box Generation

This formula is generated as a regression problem where first parameter is the value of regression target obtained, i.e., bounding box and second parameter is the ground-truth coordinate which calculates the confidence level.

4.3.3 Facial Landmarks Positioning: After obtaining the non-overlapped and minimized bounding-box along with confidence level values. The formula for this task is also denotes as regression problem along with the minimization of the Euclidean Loss.

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2 \tag{3}$$

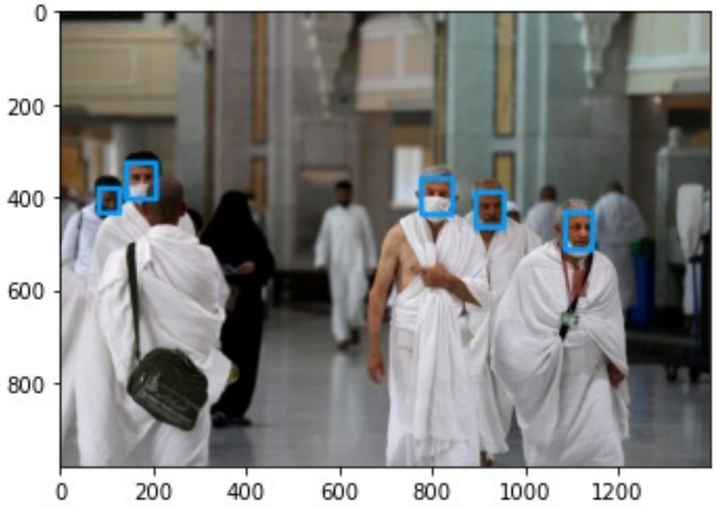


Fig.14. Robust Face Detection

where the first parameter is the landmarks positions obtained by the network and second is confidence level values.

4.4 Brief comparison between Viola-Jones Algorithm and MTCNN.

4.4.1 Viola-Jones Algorithm

This algorithm is a very popular and one of oldest method to practice face detection on sample images. This framework works by scanning the images in grayscale tone because it reads the images like a haar-features collection. It mainly depicts that the framework detects the face using lighter and darker boxed around the face outline. A technique, named, integral images is used to compute the sample images at a very fast rate with less computational time and complexity.

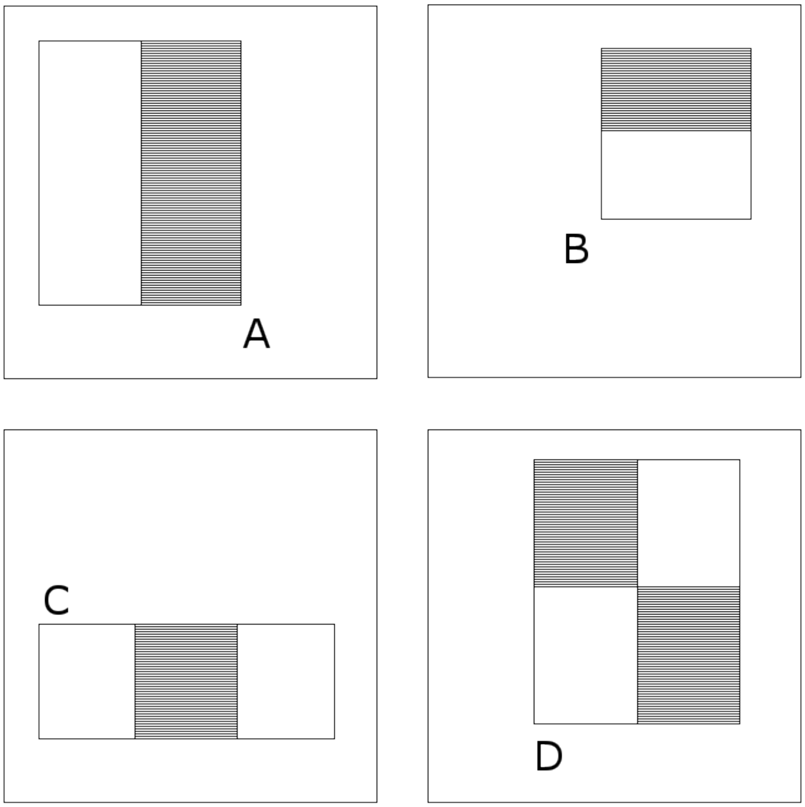


Fig.15. Haar features [10]

Viola-jones collects the processed haar-features and pass them to a cascade of classifiers of AdaBoost. These classifiers checks and scans the images to check the presence of a facial object. The process of face detection is done via multiple classifiers. If the image is rejected at the first classifiers, i.e., does not detects the face, the image gets rejected instantly and is not scanned again. While, the accepted images are then fed to next classifiers and those classifiers scans the images under more precise environment. Therefore, it results in faster filtering of the non-faces images as it follows a cascading approach.

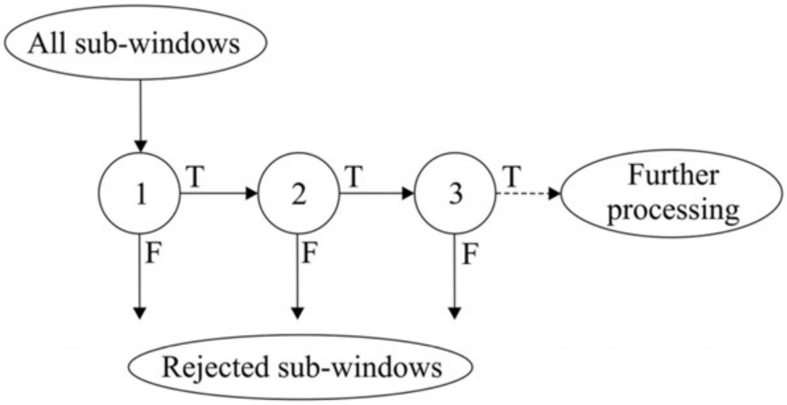
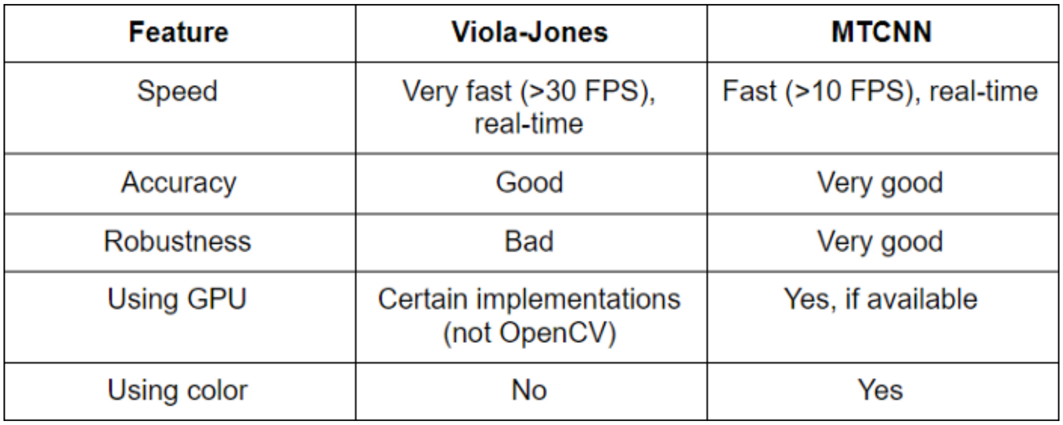


Fig.16 Viola-Jones Cascade [10]

Viola-jones while having a less computational time lacks in the robust conditions, such as if the images have the faces of different sizes. The framework resizes the images to scan the images properly and re-assigns the classifiers on them, which increase the computational process and complexity. The accuracy of the framework in these cases get degraded as the Viola-jones predicts or assumes that the face images are looking directly towards the camera with ideal amount of light and face profile. These conditions are not often met in real-time processing.

4.4.2 MTCNN over Viola-Jones



Feature	Viola-Jones	MTCNN
Speed	Very fast (>30 FPS), real-time	Fast (>10 FPS), real-time
Accuracy	Good	Very good
Robustness	Bad	Very good
Using GPU	Certain implementations (not OpenCV)	Yes, if available
Using color	No	Yes

Table.1. Comparison of Viola-Jones and MTCNN detectors [10]

As it is already discussed that MTCNN also follows a cascaded approach to detect the faces just like Viola-jones, but it performs better in the robust conditions and also produces better and precise face detection than Viola-jones algorithm.

The comparison between the performance of both frameworks are shown by images below:

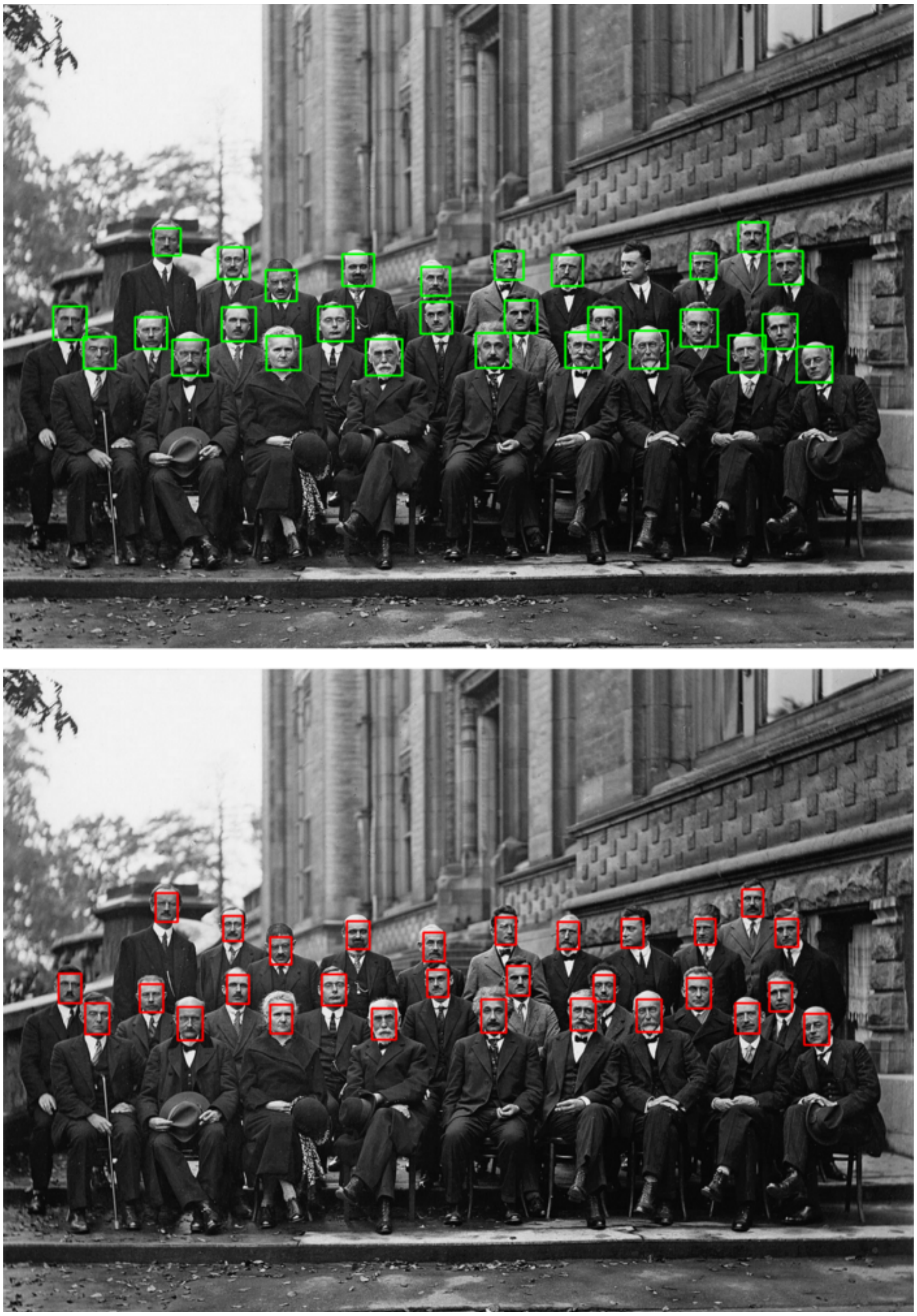


Fig.17 Upper: Viola-Jones, lower: MTCNN – 1 [10]

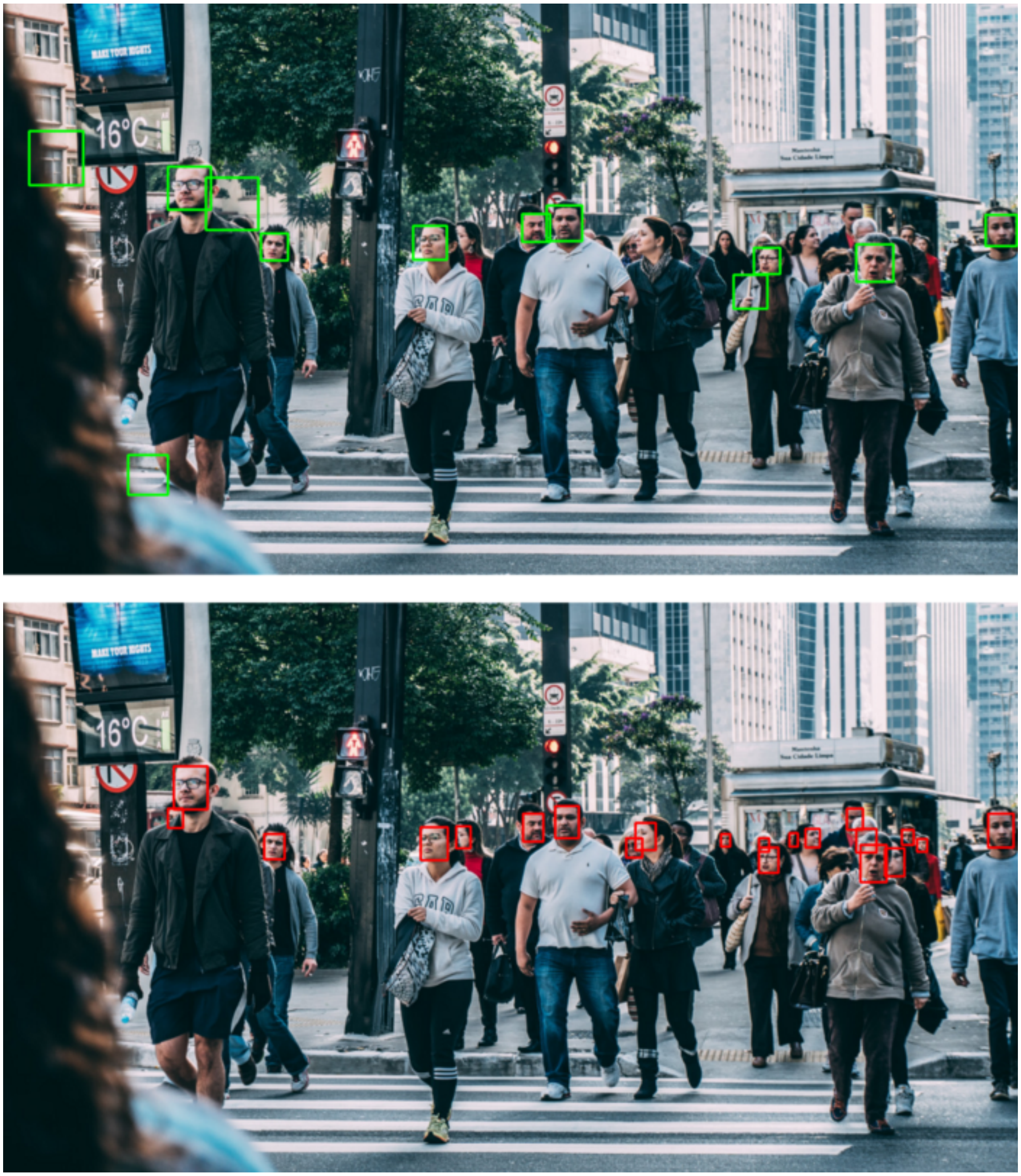


Fig.18 Upper: Viola-Jones, lower: MTCNN – 2 [10]

From both the images shown above, it can be observed that MTCNN performs better than the Viola-jones as the bounding box generated by the MTCNN framework are more precise and closer to the facial landmarks than Viola-jones algorithm.

MTCNN easily detects the faces of variable sizes and also catches the faces which are not looking towards the camera as seen in Fig.8, while in Fig.9, it is clearly observed that Viola-jones only detected the faces with similar sizes, the MTCNN detected almost all faces regardless of their size and distance to the camera.

4.5 Limitations Observed

The current methodology however performs very effectively and provide accurate mask localization from the given face samples. The MTCNN framework also has certain limitations. MTCNN framework was basically used to detect faces only at the start by localizing the facial features such as lips, nose, eyes, etc. But in the given scenarios like face mask occlusion, MTCNN is used to detect the face mask by using the facial features as a crucial factor such as detection of nose and lips.

Current framework sometimes fails to differentiate between a mask and thin cloth item such as scarfs or handkerchief as it treats both the occluding objects as a same entity, resulting in classifying mask and other object in same class which can be crucial in result accuracy.

Another limitation observed with the used methodology is that at very rare edge cases where nose is not visible due to the acute face angle or face profile, the MTCNN fails to generate the bounding box around the detected face which can alter the accuracy and efficiency of the framework for future results, while MTCNN performs drastically better than Viola-Jones algorithm in the cases of in-consistent face size, shape, etc., MTCNN does possess few limitations with it.

CHAPTER 5: EXPERIMENTAL RESULTS

5.1 Results of Experimentation

MTCNN produces the result by classifying the unique images, which are passed into the layers of CNN model from the selected datasets into two classes, named as, **“face_with_mask”** and **“face_no_mask”**. The framework calculates the facial landmarks using the bounding boxes.

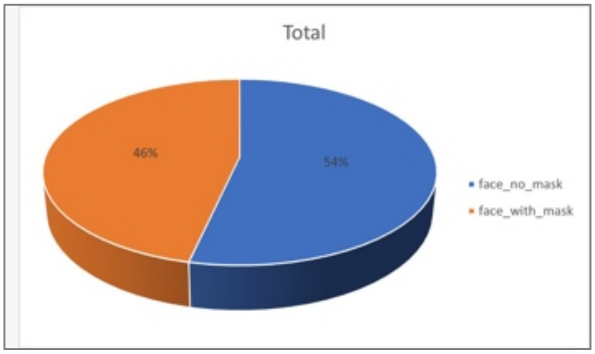
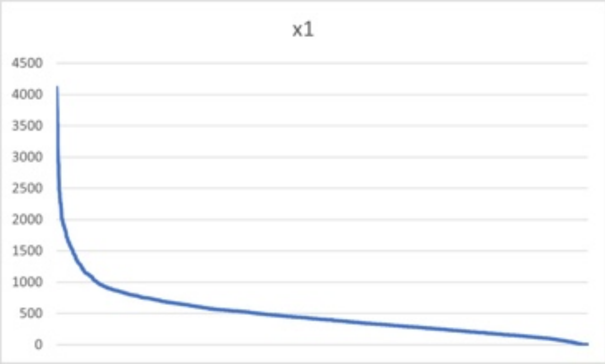
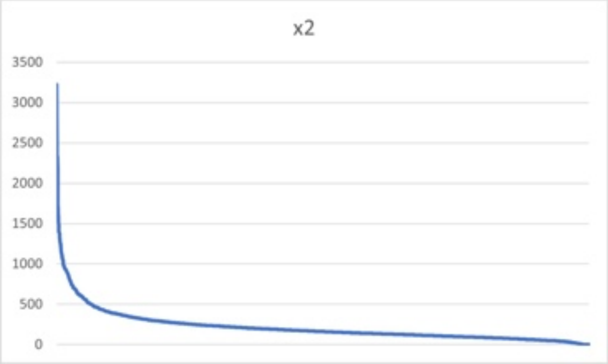


Fig.19. Classification of test images

The produced result is generated by using 4326 unique training images to train the model and 3165 unique testing images out of 6024 face images to collect the results of performance. Also, the framework calculates the confidence level of the scanned images using the coordinate values of the bounding box which are detecting the facial landmarks positions.





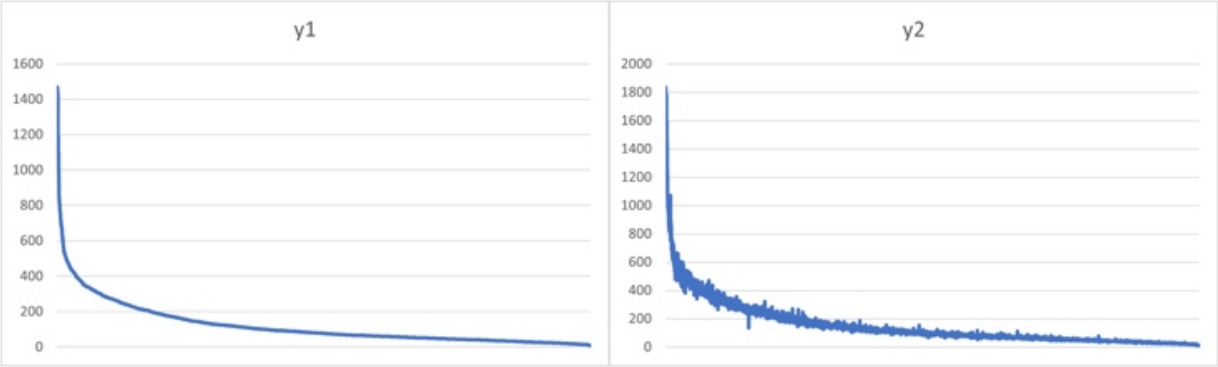


Fig.20 Bounding Box Localization

All of the graphs show the localization of the bounding boxes coordinates generated by the MTCNN framework around the detected faces from the dataset images passed in the neural network.

The deep curve of the graph shows the accurate values of facial features detected by the framework hence the big arc represents the mass classification of those bounding boxes. The coordinates of the bounding boxes are denoted by the values X1, X2, Y1 and Y2.

MTCNN localizes the critical facial features such as nose, eyes and lips to an absolute point with the generation of bounding boxes around the faces. Some of the results are shown below

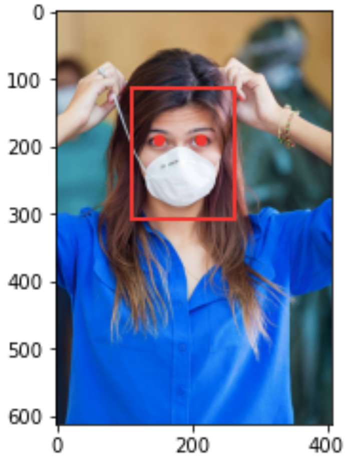


Fig.21 Facial Features Detection with Face-mask (1)

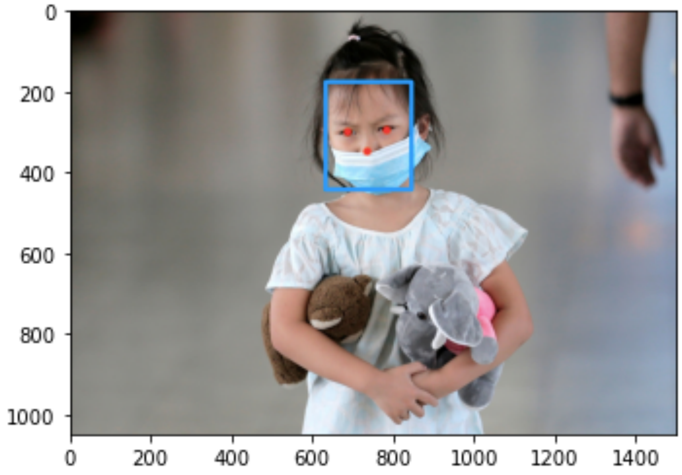


Fig.22 Facial Features Detection with Face-mask (2)

As it is observed that, MTCNN only detects the eyes of the face as other features are hidden by the face mask, Fig.21, which then classified under “face_with_mask”.

While the other image, Fig.22, has detected 3 features positions where nose is partially covered by the face mask, resulting in MTCNN detecting the nose at very diminished point, but as it is observed that not all the features are being detected by the framework. The result, Fig.22, also gets classified as “face_with_mask”.

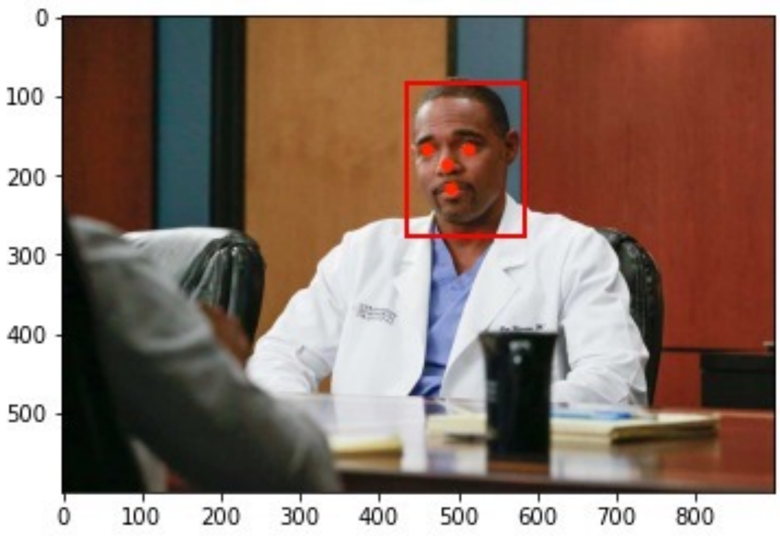


Fig.23 Facial features detection without face-mask

Here, it can be observed that the framework successfully detects all four absolute position of the face features, i.e., eyes, nose and centre point of lips of the scanned face. As, no occlusion is present in the sample image, it is classified as "face_no_mask" by the MTCNN framework.

5.2 Confusion Matrices

5.2.1 VGGFace Model

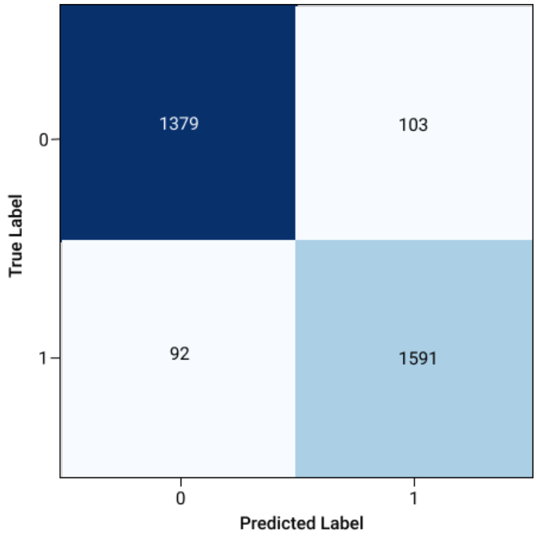


Fig.24 Confusion-Matrix(VGGFACE)

5.2.2 VGGFace2 Model

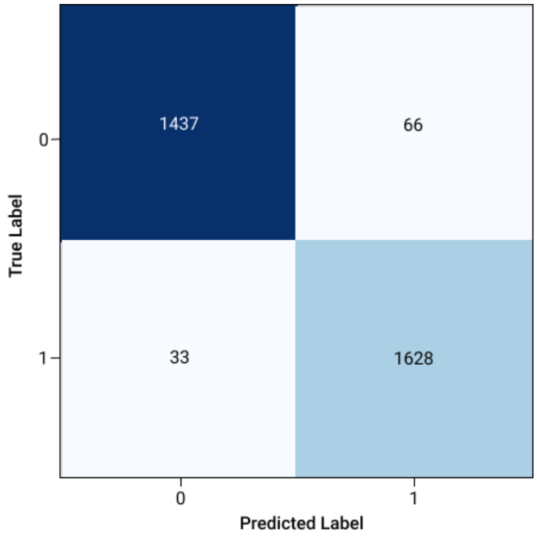


Fig.25 Confusion-Matrix(VGGFACE-2)

5.2.3 CASIA-Webface Model

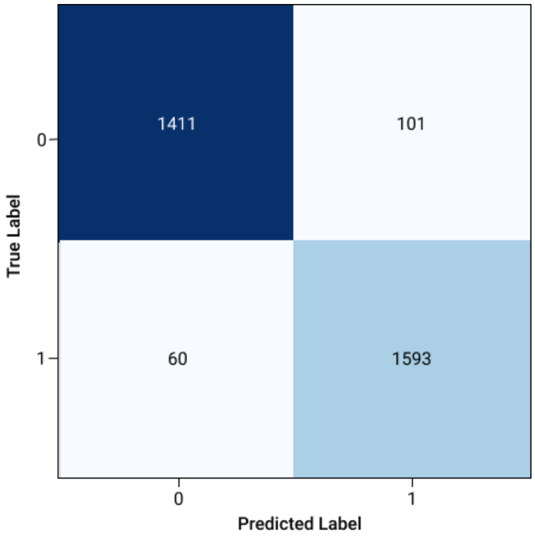
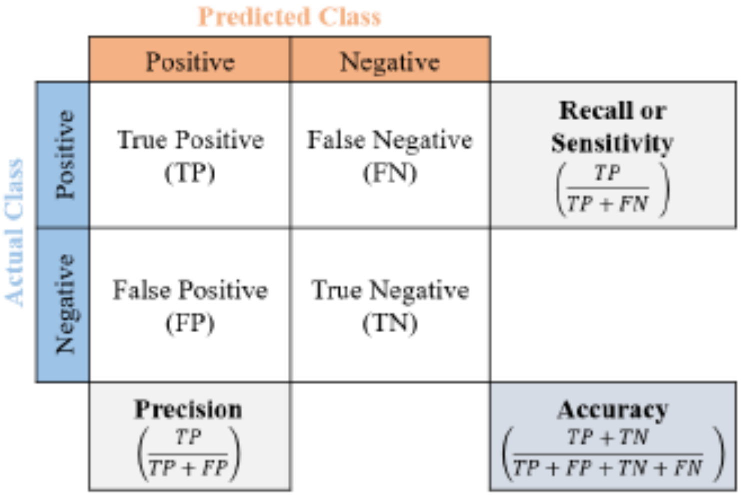


Fig.26 Confusion-Matrix(CASIA)

The results of the three models, VGGFACE, VGGFACE2 and CASIA-Webface are generated on the basis of their precision, accuracy and recall. These metrics are used to observe the best performing model.



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN)	Recall or Sensitivity $\left(\frac{TP}{TP+FN}\right)$
	Negative	False Positive (FP)	True Negative (TN)	
		Precision $\left(\frac{TP}{TP+FP}\right)$		Accuracy $\left(\frac{TP+TN}{TP+FP+TN+FN}\right)$

Fig.27 Confusion matrix with metrics calculations

5.3 Performance comparison

Model	Precision (%)	Accuracy (%)	Recall (%)
VGGFACE	93.05	93.84	93.75
VGGFACE-2	95.61	96.87	97.76
CASIA-Webface	93.32	94.91	95.92

Table 5. Model Performance Comparison

Each of the model is trained with total of 3165 sample images which included both masked and non-masked images. From table.5, it is observed that model VGGFACE-2 performs better than other models.

CONCLUSION & FUTURE SCOPE

This research article introduced an efficient framework with notably better and improved results than other working frameworks present. The MTCNN frameworks successfully classifies the difference between both masked and non-masked face images taken by the MaskedFace-Net, FFHQ datasets. This practice of detecting facial features under robust conditions like angle, illumination and importantly face masks defines the benchmark results and performance level against other CNN models and face detecting frameworks. Hence, the three cascaded layers present in the introduced framework provides less computational processing and proves the algorithm efficient in real-time robust scenarios.

One of the crucial issue created by the occlusion caused by face masks is that it is now very difficult for automated technologies to verify a person or identify a person's face as it is covered by the face mask. One of the main future scope for the MTCNN framework is to detect the face-masks accurately and achieve full identification of a person regardless of the occlusions. This model can be improved by introducing it to more advanced models which uses more effective neural networks and is able to eliminate the occlusion present in face images from the real-world scenarios. Also, MTCNN can also be improved for certain edge cases such as when the framework is unable to detect crucial facial features such as nose and it does not generate the bounding-box around the face.

LIST OF PUBLICATIONS

[1]. D. Jayant and A. Bhat, "Study of robust facial recognition under occlusion using different techniques," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1230-1235, doi: 10.1109/ICICCS51141.2021.9432300.

[2]. Communicated Paper: D. Jayant and A. Bhat, “Robust face mask detection using multi-task cascaded convolutional neural networks”, *3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–21).*

REFERENCES

[1]. K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016

[2]. Xavier P. Burgos-Artizzu and Pietro Perona. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, pages 1513–1520, 2013.

[3]. Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Largescale, Real-world Database for Facial Landmark Localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, 2011.

[4]. X. Zhu, and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879-2886.

[5]. P. Viola and M. J. Jones, "Robust real-time face detection. International journal of computer vision," vol. 57, no. 2, pp. 137-154, 2004

[6]. Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2014.

[7]. Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in European Conference on Computer Vision, 2014, pp. 94-108.

[8]. Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.

[9]. E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, 2016.

[10]. MTCNN: "*https://towardsdatascience.com/robust-face-detection-with-mtcnn-400fa81adc2e*"

[11]. Deep learning. https://en.wikipedia.org/wiki/Deep_learning

[12]. A.Dhall, R.Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011.

[13]. S. Li and W. Deng, "Reliable crowdsourcing and deep localitypreserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[14]. A.Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[15]. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, 2010, pp. 94-101.

[16]. Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti PietikInen. 2011. Facial expression recognition from near-infrared videos. Image Vision Comput. 29, 9 (August 2011), 607-619.

[17]. Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba.

[18]. B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in IEEE International Conference on Computer Vision, 2015, pp. 82-90.

[19]. Cands, Emmanuel J., et al."Robust principal component analysis?" Journal of the ACM (JACM) 58.3 (2011): 11. New York: Academic, 1963, pp. 271–350.

[20]. Karras, Tero & Laine, Samuli & Aila, Timo. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. 4396-

4405. 10.1109/CVPR.2019.00453. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Con- ference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[21]. W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local gabor features and classifier synthesis," Pattern Recognition, vol. 45, no. 1, pp. 80–91, 2012.

[22]. Cabani, Adnane & Hammoudi, Karim & Benhabiles, Halim & Melkemi, Mahmoud. (2020). MaskedFace-Net – A dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health. 19. 10.1016/j.smhl.2020.100144.