Face Detection and Tracking in Video Using MTCNN

A thesis submitted towards partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY

IN

COMPUTER TECHNOLOGY

By

Tanmay Saha

Examination Roll: M6TCT24001

Registration No: 160073

under the guidance of

Prof. (Dr.) Jamuna Kanta Sing

Department of Computer Science and Engineering

Faculty Council of Engineering and Technology

Jadavpur University

Kolkata - 700032, West Bengal, India

2021-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY COUNCIL OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

I hereby recommend that the thesis entitled "**Face Detection and Tracking in Video Using MTCNN**" prepared under my supervision and guidance at Jadavpur University, department of Computer Science and Engineering, Kolkata by **TANMAY SAHA (Registration No. 160073 of 2021-2024, Class Roll No. 002110504002, Examination Roll No M6TCT24001)**, may be accepted in partial fulfillment for the degree of **Master of Technology in Computer Technology** under the Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021-2024, I wish him every success in life.

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Department of Computer Science and Engg.
Jadavpur University
Kolkata-700032

Prof. (Dr.) Jamuna Kanta Sing
Supervisor
Department of Computer Science and Engg.
Jadavpur University
Kolkata-700032

Prof. (Dr.) Dipak Laha
Dean, Faculty Council of Engg. and Tech.
Jadavpur University
Kolkata-700032

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY COUNCIL OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis work consists of several literature survey and original in re-search work by the undersigned candidate, as a student of MASTER OF TECHNOLOGY IN COMPUTER TECHNOLOGY. All the Information in this document have been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all the material results that are hot original to this work.

Name: TANMAY SAHA
Examination Roll number: M6TCT24001
Thesis Title: Face Detection and Tracking in Video Using MTCNN.

Signature with date:

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY COUNCIL OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that this thesis work entitled "**Face Detection and Tracking in Video Using MTCNN**" is a legitimate record work done by **Tanmay Saha (Class roll Number: 002110504002. Examination Roll Number: M6TCT24001, Registration Number: 160073)** in a partial fulfillment of the degree Master of Technology in Computer Technology under the department of Computer Science and Engineering, Jadavpur University during the period September 2021 to August 2024. It is understood that by this approval the under- signed do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for which it is submitted.

(……)
Signature of Examiner 1

(……)
Signature of Examiner 2

Acknowledgment

I would Like to express my honest and sincere thanks and humble gratitude to my honorable teacher and guide *Prof. (Dr.) Jamuna Kanta Sing*, Jadavpur University, for his complete guidance and full support in completing and producing this project successfully. I am very much obliged to him for the constant encouragement and continuous inspiration that he has given to me. These words are only a token of my deep respect towards him for all he has done to take my thesis to the present shape.

I would like to thank *Mr. Manas Ghosh* for his support and valuable suggestions to the activities of the thesis at any phase.

Finally, I would like to express my sense of gratitude and thankfulness to my parents who unconditionally supported me without which I would hardly be capable of completing this huge work.

…….....
Tanmay Saha
Examination Roll Number: M6TCT24001
Registration No: 160073

Abstract

This study delves into the groundbreaking integration of artificial intelligence (AI) into crime prevention and perpetrator identification through the lens of facial recognition technology. By harnessing the intricate features of the human face, facial recognition emerges as a promising tool in these critical domains. The research aims to tackle the inherent challenges in achieving precise and reliable facial detection by developing and refining a robust machine learning model. The foundation of this investigation lies in the utilization of the Chokepoint dataset, a comprehensive repository comprising 28,470 indoor images. Through meticulous analysis and model refinement, the study achieves an impressive accuracy rate of 96.64% in facial recognition. Moreover, the research extends its scope to outdoor scenarios by leveraging the Bengal dataset, specifically designed for multi-face detection. With the integration of a supplementary 1000-image dataset, the developed model demonstrates a remarkable confidence score of 98.19%.

Central to the success of this endeavor is the adoption of the Multi-Task Convolutional Neural Network (MTCNN), a versatile framework celebrated for its proficiency in simultaneously detecting faces, facial landmarks, and estimating facial attributes. By harnessing the capabilities of MTCNN, the research not only enhances detection accuracy but also improves operational efficiency. Beyond the technical advancements, the study also emphasizes algorithmic enhancements and detection mechanism optimizations, aimed at pushing the boundaries of facial recognition systems. These innovations are poised to elevate the efficacy of security measures and enhance user interface interactions, thereby bolstering the overall utility and applicability of facial recognition technology in diverse real-world scenarios.

Contents

List of Figures viii

List of Tables ix

1. Introduction 1

1.1 Evolution of Face Detection Technology 3

1.2 Challenges in Visual Face 3

1.3 Face Tracking and Detection in Video 4

1.4 Face Detection in Video 4

1.5 Face Tracking in Video 5

1.6 Multi-Face Detection and Tracking 6

1.7 The Role of MTCNN in Face Detection and Tracking 7

1.8 Review of Early work on MTCNN 7

2. Proposed Method 9

2.1 Utilization of MTCNN 9

2.2 Preprocessing 9

2.3 Face Detection Equation 10

2.4 Working of MTCNN 11

2.6 MTCNN Architecture 14

3. Results and Discussions 16

3.1 Dataset 16

3.1.1 Chokepoint Dataset 16

3.1.2 Bengal Dataset 21

3.2 Evaluation Metrics 23

3.3 Implementation Details 29

3.4 Discussions 30

4. Conclusion 31

4.1 Limitations 32

References 33

List of Figures

1.1 (1) Shows a man approaching towards the camera (2) The face of man is detected and bounded box has been created around the face. 5
1.2 Bengal dataset which shows multi face in single frame. 6
1.3 MTCNN Face detection Algorithm. 8
2.1 P-Net Block diagram 11
2.2 R-Net Block diagram 12
2.3 O-Net Block diagram 13
2.4 Proposed MTCNN Architecture 14
3.1 Recording setup used for the Chokepoint dataset. 18
3.2 Sample Chokepoint dataset 19
3.3 Original image frame from Chokepoint dataset. 22
3.4 Face detected and created a bounded box. 22
3.5 Facial landmark localization after detection. 22
3.6 Graph for Chokepoint dataset. 24
3.7 Original image frame from Bengal dataset. 27
3.8 Face detected and created a bounded box. 27
3.9 Facial landmark detected and marked. 27

List of Tables

3.1 Chokepoint Dataset 22
3.2 Dataset used in this experiment. 21
3.3 Accuracy results on Chokepoint Dataset. 24
3.4 Comparison between various detection method previously used for Chokepoint dataset. 25
3.5 Average Confidence rate for Bengal Dataset. 28

Chapter 1

Introduction

Face detection and tracking technology has undergone significant advancements in recent years, revolutionizing various domains such as security, surveillance, entertainment, and human-computer interaction. The ability to automatically detect and track faces in video sequences has paved the way for innovative applications and solutions that enhance safety, efficiency, and user experience. This thesis explores the intricate process of face detection and tracking, focusing on the utilization of the Multi-Task Convolutional Neural Network (MTCNN) for robust and accurate performance in diverse scenarios [[1], [33]].

The evolution of face detection techniques has been fueled by advancements in artificial neural networks (ANNs) and computer vision algorithms. From traditional methods relying on handcrafted features and heuristic approaches to sophisticated deep learning models, the field has witnessed a paradigm shift towards more data-driven and end-to-end learning-based approaches. Within this landscape, convolutional neural networks (CNNs) have emerged as powerful tools for feature extraction and pattern recognition, enabling remarkable progress in face detection and tracking tasks [34].

Facial recognition, a subset of face detection and tracking, holds immense potential for various applications, including entertainment, law enforcement, biometrics, and social media. The ability to identify individuals from images or video frames has become increasingly important in today's digital age. Moreover, facial recognition serves as a foundational component in many systems and processes, facilitating personalized experiences, secure access control, and targeted advertising, among other functionalities [35].

The initial stage of any face detection and tracking system is crucial, as it lays the foundation for subsequent operations and analyses. Facial detection, the process of locating and identifying faces within an image or video frame, serves as the fundamental building block for downstream tasks such as facial landmark localization, expression analysis, and tracking. Effective face detection algorithms exhibit high accuracy, robustness to variations in pose, illumination, and occlusion, and efficiency in processing large-scale datasets in real-time or near-real-time scenarios.

Beyond the realm of face detection and tracking, several alternative methods and approaches have been proposed and employed. These methods include Siamese Region Proposal Network (SiamRPN), Siamese Fully Convolutional (SiamFC), Multi-scale Bayesian Model with Discriminative Features (MBMD), and Deep Visibility Network (DVNet). Each method offers unique advantages and trade-offs in terms of precision, computational complexity, and suitability for specific applications. By comparing and contrasting these methods, researchers gain valuable insights into the strengths and limitations of different approaches, informing the development of more effective and efficient face detection and tracking solutions.

In the domain of face detection and tracking, various techniques and methodologies have been explored to address the inherent challenges and complexities of the task. Traditional methods often relied on handcrafted features, such as Haar cascades and histogram of oriented gradients (HOG), combined with machine learning algorithms such as support vector machines (SVM) or AdaBoost classifiers. While effective in certain scenarios, these methods lacked the robustness and adaptability required for real-world applications, particularly in the presence of occlusions, variations in pose and illumination, and complex backgrounds.

In recent years, deep learning-based approaches have emerged as the de facto standard for face detection and tracking, owing to their ability to automatically learn discriminative features from raw data and generalize well across diverse datasets and conditions. Convolutional neural networks (CNNs) have been particularly successful in this domain, enabling end-to-end learning of hierarchical representations directly from pixel values. Models such as MTCNN have demonstrated superior performance in terms of accuracy, efficiency, and scalability, making them indispensable tools for a wide range of face-related tasks.

1.1 Evolution of Face Detection Technology

The journey of face detection technology traces back several decades, with early attempts relying on traditional computer vision techniques. Over time, advancements in artificial neural networks (ANNs), particularly convolutional neural networks (CNNs), have revolutionized the field, enabling more accurate and efficient face detection algorithms [1]. CNNs, with their ability to automatically learn hierarchical features from raw data, have significantly improved the performance of face detection systems, making them robust to variations in lighting, pose, and facial expression.

1.2 Challenges in Visual Face

Visual face tracking, the process of locating a target face across consecutive frames in a video sequence, presents several challenges. These challenges include variations in appearance due to changes in pose, lighting, and occlusions, as well as the presence of multiple faces and motion blur. Addressing these challenges requires robust algorithms capable of adapting to dynamic environments and handling complex scenarios effectively.

1.3 Face Tracking and Detection in Video

Face tracking and detection in video streams involve the continuous identification and localization of human faces across multiple frames [2]. This process is crucial for various applications, including video surveillance, human-computer interaction, and content analysis. In video surveillance, for example, face tracking enables the monitoring of individual's movements and activities, facilitating the identification of suspicious behavior or unauthorized access.

1.4 Face Detection in Video

Face detection in video involves the initial identification of human faces within individual frames. Traditional face detection algorithms rely on handcrafted features and machine learning techniques to localize faces based on their distinct visual characteristics, such as skin tone, facial features, and geometric patterns [3]. However, these methods often struggle with variations in lighting, pose, and occlusions, leading to suboptimal performance in real-world scenarios. With the advent of deep learning and convolutional neural networks (CNNs), face detection in video has witnessed significant advancements in accuracy and robustness. CNN-based approaches, such as MTCNN, leverage large-scale datasets and complex network architectures to automatically learn discriminative features for detecting faces in diverse conditions [4]. By integrating multiple tasks, such as face localization, landmark detection, and attribute classification, MTCNN achieves state-of-the-art performance in face detection tasks, even in challenging video sequences [5].

1.5 Face Tracking in Video

Face tracking in video extends beyond the initial detection of faces and involves the continuous localization and tracking of individual faces across consecutive frames [6]. This process is essential for maintaining the identity of detected faces over time, enabling applications such as person re-identification, activity recognition, and facial expression analysis. Traditional face tracking algorithms utilize techniques such as optical flow, Kalman filtering, and appearance models to estimate the motion and appearance changes of tracked faces in video sequences [7]. However, these methods often struggle with occlusions, pose variations, and complex scene dynamics, leading to drift and identity switches in long-term tracking scenarios. Deep learning-based approaches, such as recurrent neural networks (RNNs) and Siamese networks, have shown promise in addressing these challenges by learning temporal dependencies and appearance variations directly from data [8]. By integrating CNNs with recurrent or memory modules, these methods achieve robust and accurate face tracking in video streams, even in crowded and dynamic environments.

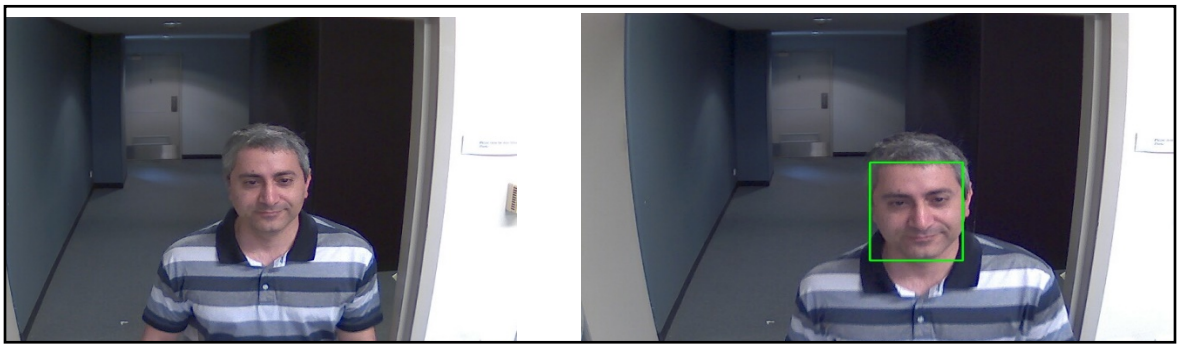


Figure 1.1: (1) Shows a man approaching towards the camera (2) The face of man is detected and bounded box has been created around the face.

1.6 Multi-Face Detection and Tracking

Detecting and tracking multiple faces in video streams pose additional challenges due to occlusions, varying scales, and interactions between individuals [9]. Traditional methods often struggle to handle such complexities effectively, leading to inaccurate detections and track interruptions. However, with the advancements in deep learning techniques, particularly multi-task convolutional neural networks (MTCNN), simultaneous detection and tracking of multiple faces have become more robust and accurate. MTCNN's ability to handle scale variations, occlusions, and complex backgrounds makes it well-suited for multi-face detection and tracking tasks, enabling applications in crowd analysis, social behavior understanding, and group activity recognition [10].



Figure 1.2: Bengal dataset which shows multi face in single frame.

1.7 The Role of MTCNN in Face Detection and Tracking

MTCNN, as a specialized deep learning model for face detection, offers several advantages over traditional approaches [11]. Its multi-task architecture enables simultaneous detection of facial landmarks, bounding boxes, and facial attributes, leading to more accurate and efficient face detection. Moreover, MTCNN's ability to handle scale variations and occlusions makes it well-suited for face tracking tasks, where maintaining the identity of a target face across frames is paramount.

1.8 Review of Early work on MTCNN

Face detection and tracking in video streams have evolved significantly over the years, with the introduction of the Multi-Task Convolutional Neural Network (MTCNN) representing a major milestone in the field. Initially proposed by [12], MTCNN revolutionized face detection by introducing a cascaded architecture consisting of three neural networks: P-Net, R-Net, and O-Net. This architecture enabled accurate detection of faces in varying poses, lighting conditions, and scales, achieving state-of-the-art performance in face detection benchmarks. Building upon this foundation, [12] extended MTCNN to handle multi-task learning, incorporating facial landmark detection and attribute classification tasks.

This extension, known as Multi-Task Cascaded CNN (MTCNN+), introduced efficient online hard example mining and demonstrated superior performance in landmark localization and attribute classification compared to traditional methods. In 2018, Tang et al. proposed an improved version of MTCNN called Dense Anchor Pyramid Net (DAPNet) for face detection. DAPNet introduced a dense anchor pyramid structure and leveraged feature pyramid networks (FPN) to capture multi-scale features effectively. This approach enhanced scale adaptation and feature representation, leading to superior performance in face detection benchmarks compared to

previous methods. The advancements made by these studies have significantly contributed to the effectiveness and efficiency of face detection algorithms, making them suitable for a wide range of applications, including video surveillance, human-computer interaction, and augmented reality.

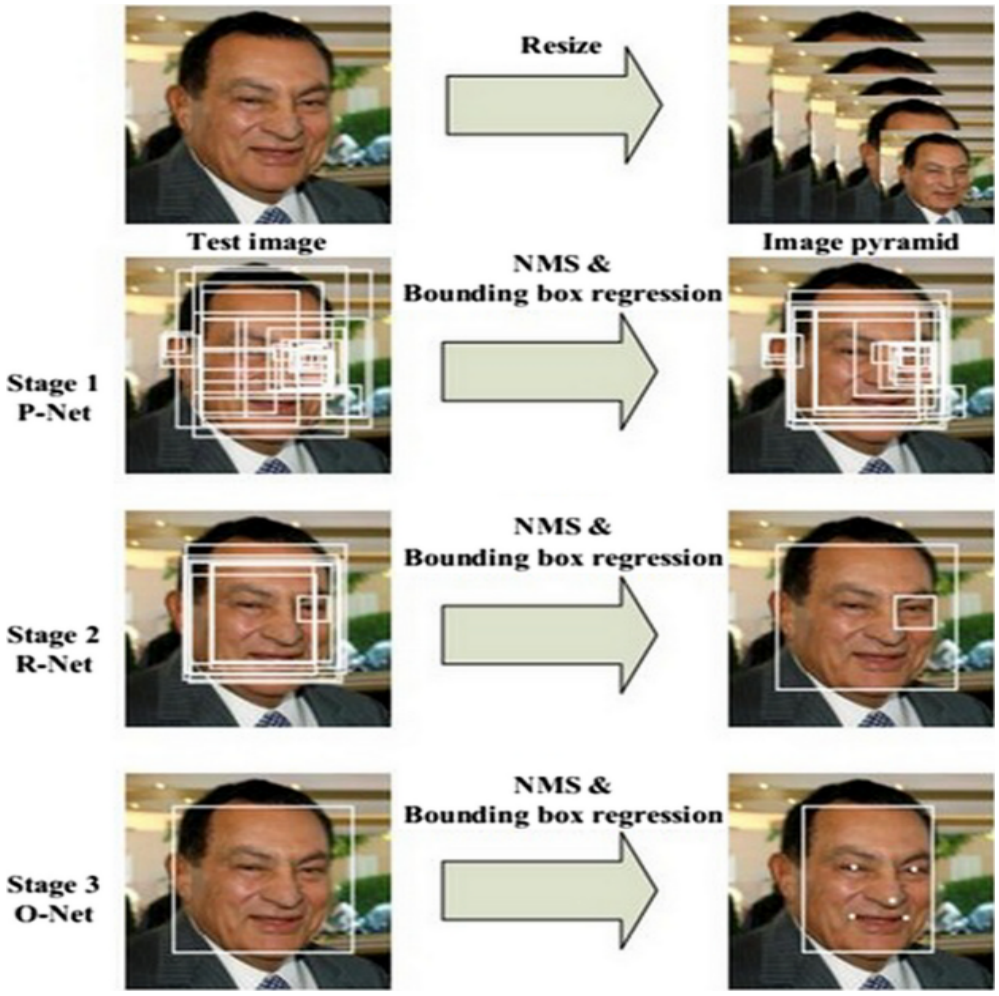


Figure 1.3: MTCNN Face detection Algorithm.[1]

[1] Khabarlak, Kostiantyn, and Larysa Koriashkina. "Fast facial landmark detection and applications: A survey." *arXiv preprint arXiv:2101.10808* (2021).

Chapter 2

Proposed Method

In this section, we present the proposed method for face detection and tracking using the Multi-Task Convolutional Neural Network (MTCNN). We outline the methodology employed in this thesis for face detection and tracking in video sequences. The proposed method leverages two distinct datasets, namely the Chokepoint dataset and the Bengal dataset, to address different aspects of face detection and tracking tasks.

2.1 Utilization of MTCNN

The proposed method leverages the MTCNN architecture, consisting of three stages: Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net). Each stage plays a crucial role in the detection and localization of faces within video sequences.

2.2 Preprocessing

The preprocessing stage plays a crucial role in preparing the datasets for subsequent analysis and training of the face detection and tracking algorithm. In this section, we outline the preprocessing steps applied to the video datasets, including the Chokepoint dataset and the Bengal dataset.

1. **Video to Image Conversion**: The first step involves converting the video datasets into a series of individual image frames. This process ensures that each frame is isolated for further analysis and processing. The videos are trimmed to the desired duration and sampled at a specified frame rate (FPS) to extract frames consistently throughout the video sequences.
2. **Image Cropping**: Following the conversion to image frames, each frame is cropped to isolate the regions of interest containing human faces. This cropping step helps remove

extraneous background clutter and focuses the analysis on the facial features of interest. The cropping parameters are determined based on the bounding box coordinates provided in the ground truth annotations.

3. **Facial Landmark Extraction**: Additionally, facial landmarks such as the coordinates of the left and right eyes are extracted from each cropped image. This information is essential for evaluating the accuracy of the face detection and tracking algorithm against the ground truth annotations. The coordinates of the facial landmarks are stored in a structured format, such as a CSV (Comma-Separated Values) file, for easy comparison and analysis.

4. **Data Augmentation (Optional)**: Depending on the dataset size and diversity, data augmentation techniques may be applied to augment the dataset further. Common augmentation techniques include random rotation, scaling, flipping, and brightness adjustments. These techniques help increase the variability of the dataset and improve the robustness of the trained model.

By implementing these preprocessing steps, the video datasets are transformed into a structured image dataset suitable for training and evaluating the face detection and tracking algorithm. The resulting dataset captures the essential facial features and landmarks necessary for accurate analysis and validation of the proposed methodology.

2.3 Face Detection Equation

The face detection process in MTCNN can be represented by the following equation:

$$Detected\ Faces = P(image) + R(P(image)) + O(R(P(image)))$$

Where: ***P(image)*** represents the candidate face regions generated by the proposal network (P-Net). ***R(P(image))*** represents the refined face regions obtained from the Refinement Network (R-Net). ***O(R(P(image)))*** represents the final detected faces after post-processing by the Output Network (O-Net).

2.4 Working of MTCNN

Stage 1:

Proposal Network (P-Net):

The first stage of MTCNN is the Proposal Network (P-Net), which generates candidate face regions [5]. P-Net takes an input image and produces a set of bounding boxes with associated confidence scores. It uses a fully convolutional network to efficiently scan the image at multiple scales and aspect ratios, detecting potential face regions.

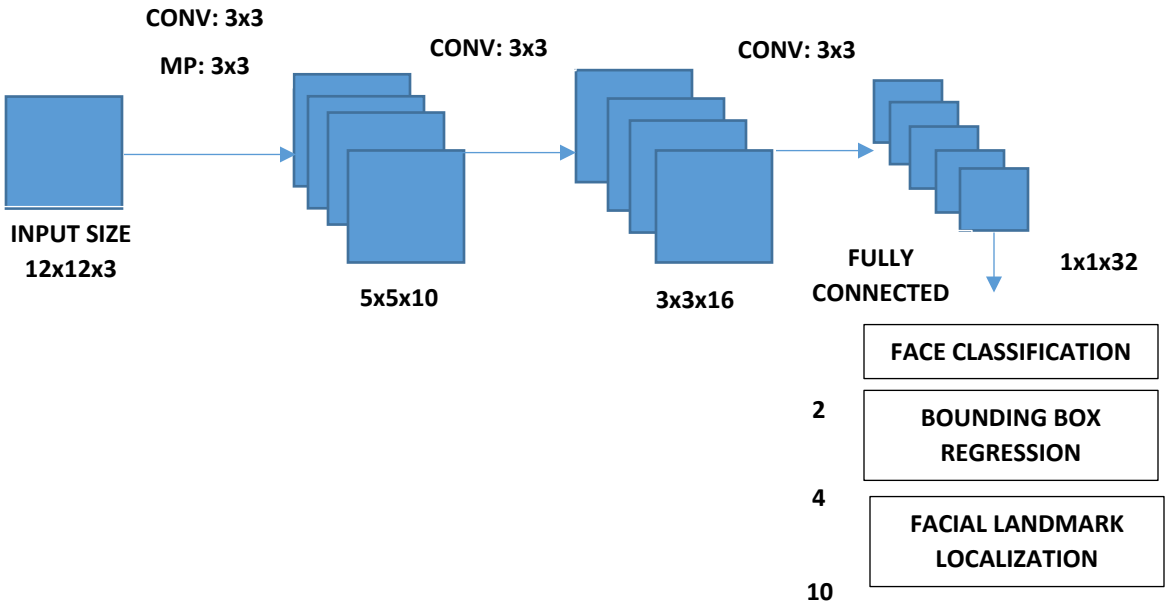


Figure 2.1: P-Net Block diagram

The P-Net is the initial stage of MTCNN responsible for generating candidate face regions. It employs a fully convolutional neural network (FCN) to scan the input image at multiple scales and aspect ratios, effectively detecting potential face regions. The network produces a set of bounding boxes along with confidence scores, indicating the likelihood of each box containing a face. P-Net operates efficiently by sharing convolutional features across different scales, allowing it to detect faces of varying sizes in real-time applications (Feng et al., 2021). Once candidate regions are identified, they are passed to the next stage, the Refinement Network (R-Net), for further processing and refinement.

Stage 2:

Refinement Network (R-Net):

The second stage is the Refinement Network (R-Net), which refines the candidate face regions obtained from the P-Net. R-Net takes the candidate face regions as input and performs more accurate bounding box regression to improve localization [5]. It also conducts face/non-face classification to filter out false positives and improve detection accuracy.

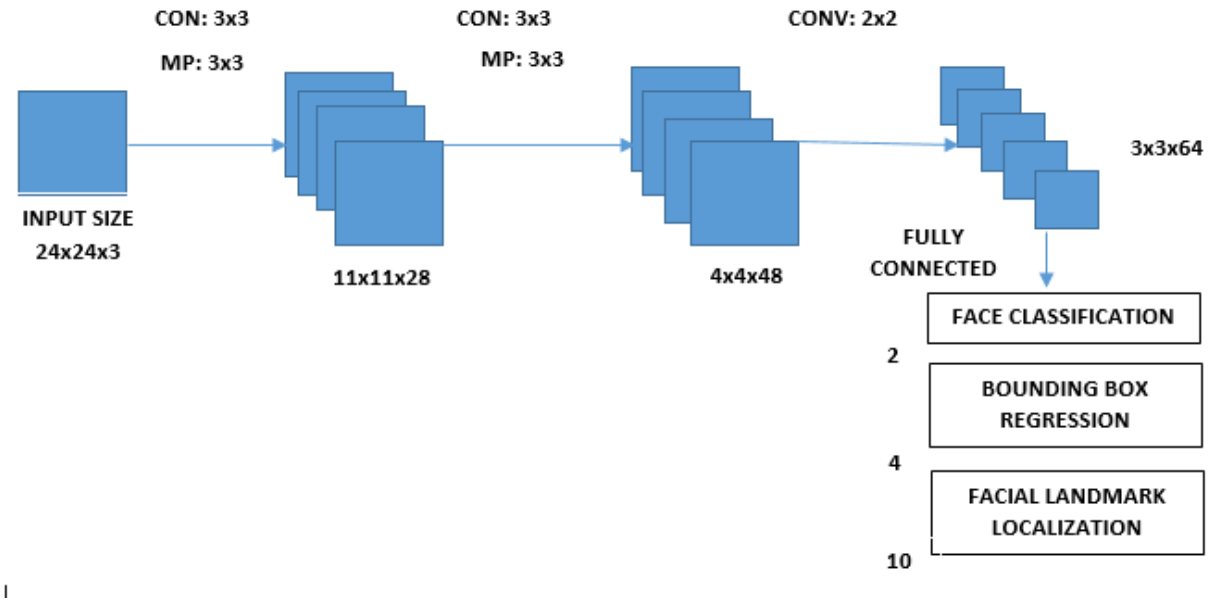


Figure 2.2: R-Net Block diagram

The R-Net is the second stage of MTCNN, designed to refine the candidate face regions obtained from the P-Net [13]. It takes the candidate face regions as input and performs more accurate bounding box regression to improve localization. Additionally, R-Net conducts face/non-face classification to filter out false positives and enhance the accuracy of face detection. By leveraging more sophisticated features and a deeper network architecture compared to P-Net, R-Net achieves higher precision in detecting and localizing faces in the input image [14]. The refined bounding boxes and classification results are then passed to the final stage, the Output Network (O-Net), for further analysis and post-processing.

Stage 3:

Output Network (O-Net):

The final stage is the Output Network (O-Net), which performs facial landmark localization and attribute classification. O-Net refines the bounding boxes further, predicts the facial landmarks (e.g., eyes, nose, mouth), and classifies facial attributes (e.g., gender, age) [5]. This stage provides additional information for post-processing and analysis of detected faces.

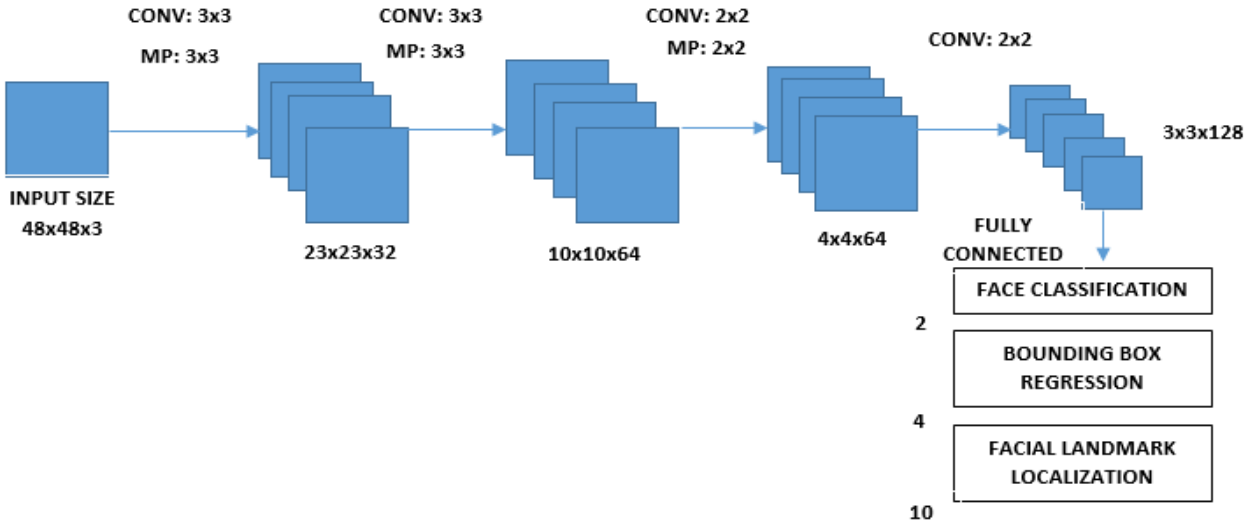


Figure 2.3: O-Net Block diagram

The O-Net is the last stage of MTCNN, responsible for performing facial landmark localization and attribute classification [15]. It further refines the bounding boxes obtained from the previous stages and predicts the positions of facial landmarks such as eyes, nose, and mouth. Additionally, O-Net conducts attribute classification tasks such as gender, age, and facial expression recognition, providing valuable information for downstream applications [16]. By integrating both localization and classification tasks, O-Net enhances the overall robustness and versatility of MTCNN, making it suitable for a wide range of face analysis tasks in real-world scenarios. The output of O-Net, including refined bounding boxes and predicted facial landmarks, can be used for various applications such as facial recognition, emotion detection, and facial expression analysis.

2.5 MTCNN Architecture

In this section, we provide a detailed description of the architecture employed in our proposed method for face detection and tracking. The cornerstone of our architecture is the Multi-Task Convolutional Neural Network (MTCNN), a deep learning model specifically designed for face detection and facial landmark localization tasks. MTCNN consists of three interconnected stages: Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net), each serving a distinct purpose in the face detection pipeline.

The MTCNN serves as the backbone of our architecture, facilitating accurate and efficient face detection in video sequences.

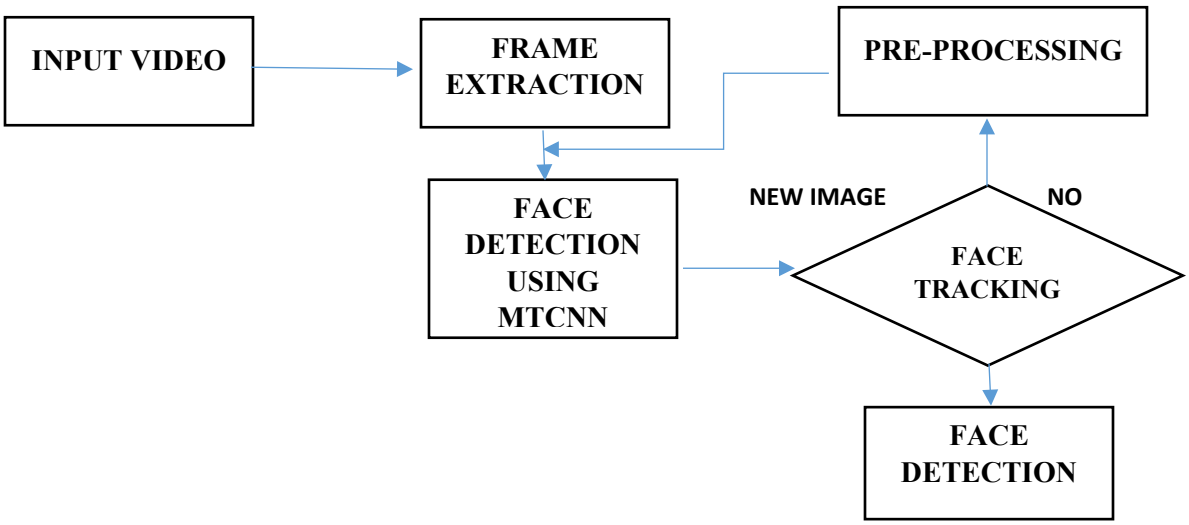


Figure 2.4: Proposed MTCNN Architecture.

It comprises three stages:

Proposal Network (P-Net): The initial stage of MTCNN responsible for generating candidate face regions within the input image. It efficiently scans the image at multiple scales and aspect ratios, producing bounding boxes along with confidence scores for potential face regions.

Refinement Network (R-Net): The intermediate stage that refines the candidate face regions obtained from P-Net, improving localization accuracy and filtering out false positives. It employs a deeper network architecture compared to P-Net, allowing for more sophisticated feature extraction and fine-tuning of bounding box coordinates.

Output Network (O-Net): The final stage responsible for post-processing the refined bounding boxes and predicting facial landmarks. It further refines the bounding boxes and estimates the positions of key facial landmarks such as eyes, nose, and mouth, providing additional information for facial analysis and tracking tasks. This MTCNN architecture forms the foundation of our proposed method for face detection and tracking, offering robustness, accuracy, and versatility in various real-world scenarios.

Chapter 3
Results and Discussions

3.1 Dataset

The datasets utilized in this study comprise video sequences capturing real-world scenarios. These datasets encompass a variety of environments, including both indoor and outdoor settings, reflecting diverse contexts for face detection and tracking. The videos are processed to extract individual frames, which are then cropped to isolate the regions of interest containing faces. Through this approach, the datasets provide a rich source of data for training and evaluating face detection and tracking algorithms, enabling comprehensive analysis and validation of the proposed methodology.

The Chokepoint dataset captures indoor scenes, depicting scenarios such as corridors, hallways, and entryways, where individuals traverse through controlled environments. In contrast, the Bengal dataset encompasses outdoor scenes, showcasing dynamic contexts such as streets, parks, and public spaces, where individuals interact amidst varying lighting conditions and background clutter. By encompassing both indoor and outdoor environments, these datasets offer a diverse range of challenges and complexities, making them ideal for evaluating the robustness and adaptability of face detection and tracking algorithms across different scenarios.

3.1.1 Chokepoint Dataset

The Chokepoint dataset serves as the primary dataset for evaluating face detection and tracking algorithms in controlled environments [17]. This dataset consists of video sequences captured from surveillance cameras at various checkpoints, focusing on single individuals passing through a designated area. The proposed method utilizes the Chokepoint dataset to achieve the following objectives:

Face Detection: The first step involves detecting faces within each frame of the video sequence using the Multi-Task Convolutional Neural Network (MTCNN). The MTCNN algorithm is applied to localize faces accurately, providing bounding box coordinates for each detected face [18].

Facial Landmark Detection: In addition to face detection, the proposed method also includes the detection of facial landmarks, specifically the coordinates of the left and right eyes [19]. This is achieved using additional modules within the MTCNN architecture, allowing for precise localization of key facial features.

Visualization: To enhance the visual representation of detected faces and facial landmarks, the proposed method includes a post-processing step to overlay bounding boxes around detected faces and mark the coordinates of the left and right eyes with small dots or points. This visualization aids in the interpretation and analysis of the detected facial features.

Chokepoint dataset designed for experiments in person identification/verification under real-world surveillance conditions using existing technologies [20]. An array of three cameras was placed above several portals (natural choke points in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way (see example). While a person is walking through a portal, a sequence of face images (i.e. a face set) can be captured. Faces in such sets will have variations in terms of illumination conditions, pose, sharpness, as well as misalignment due to automatic face localization/detection. Due to the three camera configuration, one of the cameras is likely to capture a face set where a subset of the faces is near-frontal.

The dataset consists of 25 subjects (19 males and 6 females) in portal 1 and 29 subjects (23 males and 6 females) in portal 2. The dataset has frame rate of 30 fps and the image resolution is 800X600 pixels. In total, the dataset consists of 48 video sequences and 64,204 face images. In all sequences, only one subject is presented in the image at a time. The first 100 frames of each sequence are for background modelling where no foreground objects were presented. Out of which I have used P1E_S#_C# dataset which shows the persons entering with 28,470 total face images.

Example

An example of the recording setup used for the Chokepoint dataset. A camera rig contains 3 cameras placed just above a door, used for simultaneously recording the entry of a person from 3 viewpoints. The variations between viewpoints allow for variations in walking directions, facilitating the capture of a near-frontal face by one of the cameras.

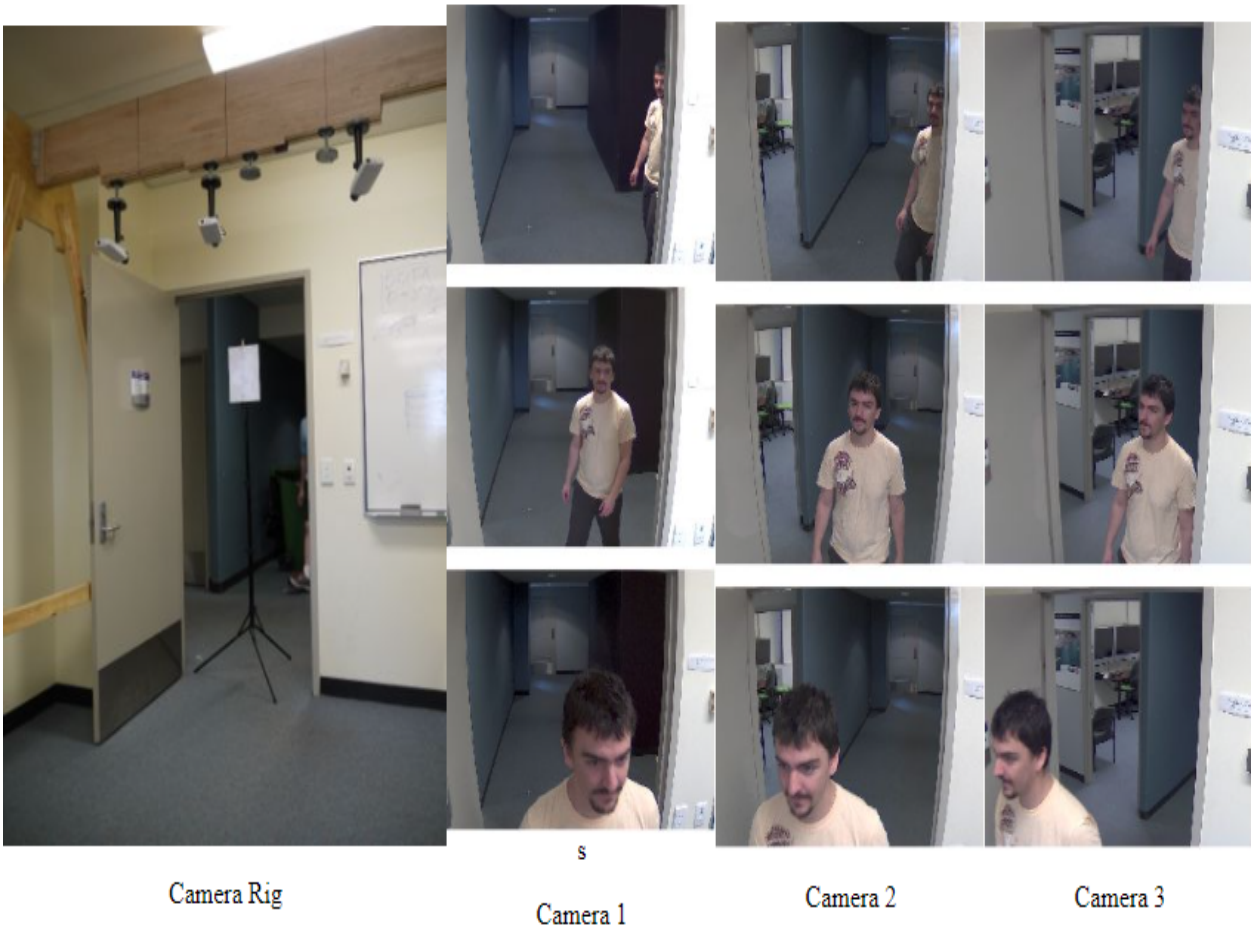


Figure 3.1: Recording setup used for the Chokepoint dataset.

Example shots from the Chokepoint dataset that I have used,

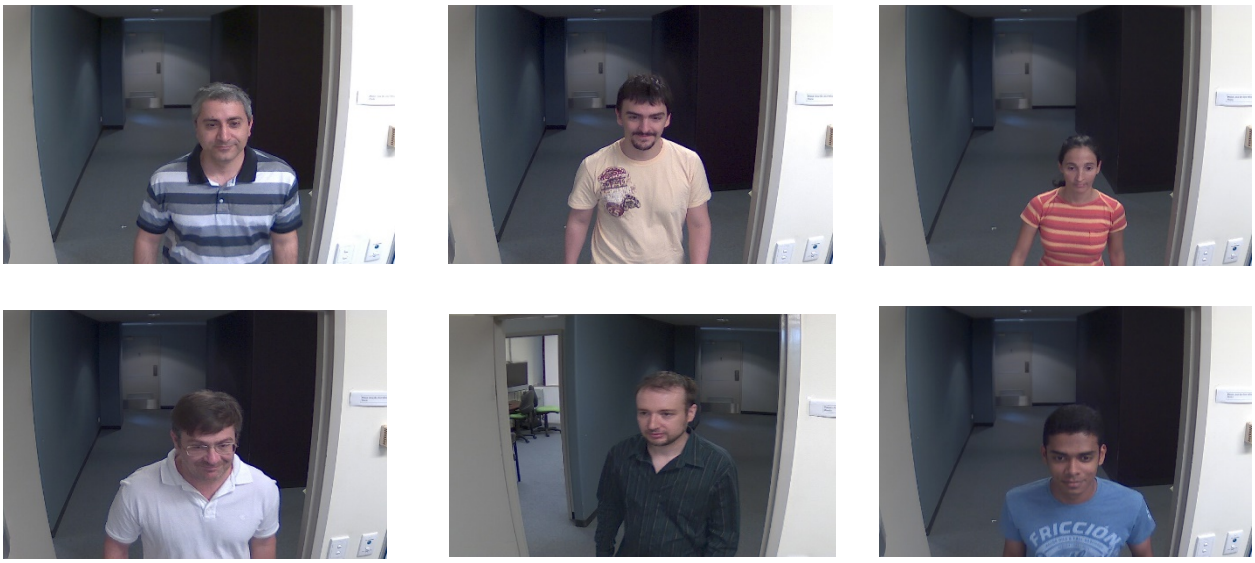


Figure 3.2: Sample Chokepoint dataset

We also encourage the experiments to be conducted with two evaluation conditions:

a). Single camera (SC)

- Using faces from the camera with most frontal view (shown above).

b). Multiple camera (MC)

- Using faces from all three cameras. For example, images for portal 1E and sequence 1 will are taken from P1E_S1_C1, P1E_S1_C2 and P1E_S1_C3.

This dataset can be applied, but not limited, to the following research areas:

- person re-identification
- image set matching
- face quality measurement
- face clustering
- 3D face reconstruction
- pedestrian/face tracking
- background estimation and subtraction

Recording Environment

P1E: indoor

Sequence Name

Sequence: P#*_S#_C# <Portal.no.type_Sequence.no_Camera.no>

< type == E: enter >

< L: leave >

Dataset used here,

Table 3.1: Chokepoint Dataset

Dataset Name	**Total no. of Images**
P1E_S1_C1	2,295
P1E_S1_C1	2,295
P1E_S1_C1	2,295
P1E_S2_C1	2,369
P1E_S2_C1	2,369
P1E_S2_C1	2,369
P1E_S3_C1	2,468
P1E_S3_C1	2,468
P1E_S3_C1	2,468
P1E_S4_C1	2,358
P1E_S4_C1	2,358
P1E_S4_C1	2,358
Total	28,470

3.1.2 Bengal Dataset

The Bengal dataset complements the Chokepoint dataset by providing a more challenging scenario for face detection and tracking, focusing on crowded environments with multiple individuals. This dataset consists of video sequences captured from public spaces, such as marketplaces or train stations, where numerous people arrive and depart within a short timeframe. The proposed method utilizes the Bengal dataset to achieve the following objectives:

Multi-Face Detection: Unlike the Chokepoint dataset, which primarily focuses on single individuals, the Bengal dataset presents the challenge of detecting and tracking multiple faces simultaneously. The proposed method incorporates techniques for multi-face detection to accurately identify and localize faces within crowded scenes.

Face Labeling: In addition to face detection, the proposed method includes a face labeling step to assign unique identifiers to each detected face. Faces are numbered sequentially, starting from 1, to track their movements and interactions throughout the video sequence.

Basically Bengal dataset is collected from the YouTube for the multi face detection and the environment of the video is kept or prepared in such a way that the crowded area where multiple object here in this case multi face people are arriving in the frame making it more difficult for the detector to detect and track the human face.

Table 3.2: Dataset used in this experiment.

Video	Length	FPS	Frames	Resolution	Scenario	Subjects
Chokepoint	1' 24"	30	28,470	800x600	indoor	24
Bengal	40"	25	1000	1920x1080	outdoor	36

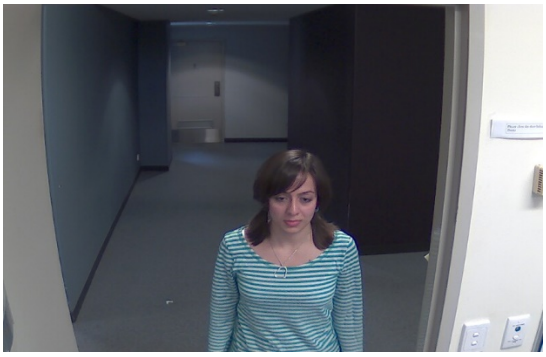


Figure 3.3: Original image frame from Chokepoint dataset.

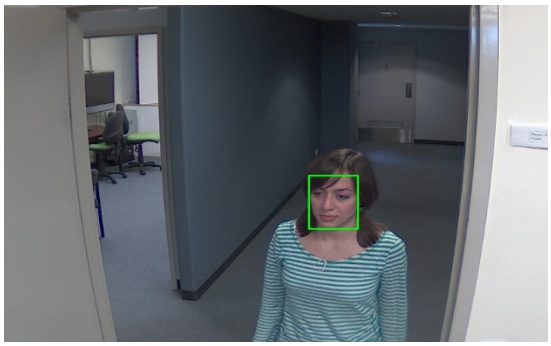


Figure 3.4: Face detected and created a bounded box.

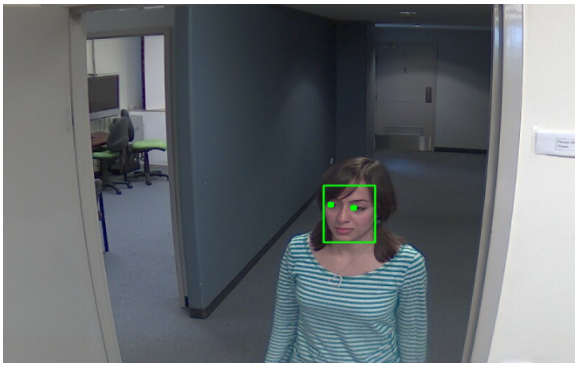


Figure 3.5: Facial landmark localization after detection.

3.2 Evaluation Metrics

In evaluating the performance of the proposed face detection and tracking method, accuracy serves as the primary evaluation metric. Accuracy measures the correctness of the algorithm's predictions by comparing them to ground truth annotations. For each dataset experiment, the algorithm's accuracy is computed by comparing the detected faces with their corresponding ground truth annotations.

To obtain the overall accuracy of the proposed method, the results from the individual stages of the Multi-Task Convolutional Neural Network (MTCNN) - P-Net, R-Net, and O-Net - are combined. Each stage contributes to the overall accuracy by refining the detected faces and predicting facial landmarks with increasing precision.

The accuracy of each stage is computed separately by comparing the detected faces and facial landmarks with their ground truth annotations. These accuracies are then aggregated to obtain the final accuracy of the MTCNN model.

Accuracy= (Number of Correct Predictions)/(Total Number of Predictions)

Where:

The "Number of Correct Predictions" includes both correctly detected faces and correctly predicted facial landmarks (such as the left eye). For facial landmark prediction, a threshold value of 5 is applied to determine whether the coordinates of the predicted left eye are sufficiently close to the coordinates of the ground truth left eye.

If the difference between the predicted and ground truth coordinates is less than or equal to the threshold value, the prediction is considered correct. The "Total Number of Predictions" encompasses all instances detected by the algorithm, including both correctly and incorrectly predicted faces and facial landmarks.

Table 3.3: Accuracy results on Chokepoint Dataset.

DATASET NAME	TOTAL NO. OF IMAGES UNDER EACH DATASET	ACCURACY
P1E_S1_C1	2295	97.75%
P1E_S1_C2	2295	95.49%
P1E_S1_C3	2295	90.41%
P1E_S2_C1	2369	98.60%
P1E_S2_C2	2369	98.90%
P1E_S2_C3	2369	98.23%
P1E_S3_C1	2468	95.20%
P1E_S3_C2	2468	97.21%
P1E_S3_C3	2468	97.53%
P1E_S4_C1	2358	96.02%
P1E_S4_C2	2358	98.31%
P1E_S4_C3	2358	96.05%
Total	**28,470**	**96.64%**

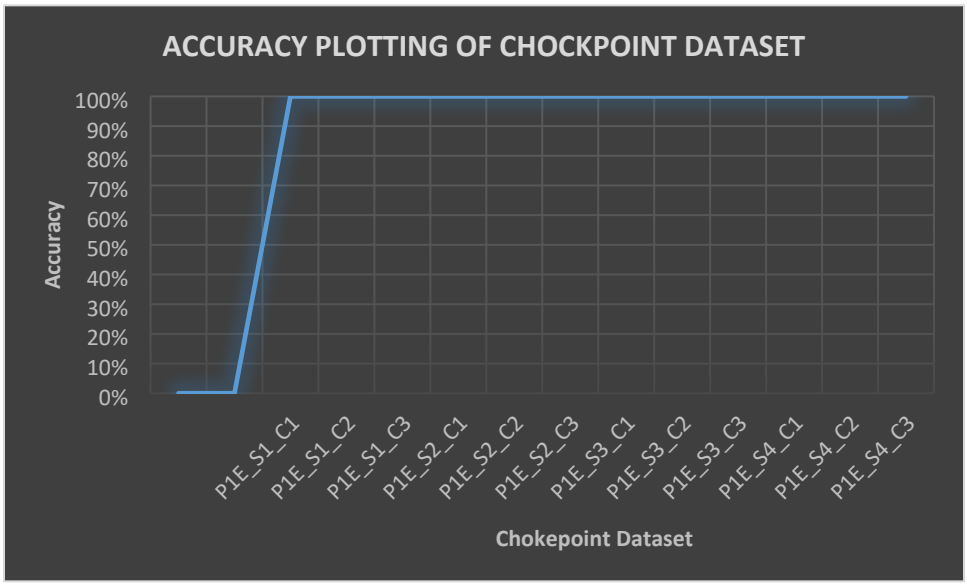


Figure 3.6: Graph for Chokepoint dataset.

The evaluation of the proposed face detection and tracking method was further extended to compare its performance with several existing detection methods, including SiamRPN (Siamese Region Proposal Network), SiamFC (Siamese Fully Convolutional), MBMD, and DVNet. Each method was assessed based on precision metrics to determine their effectiveness in detecting and tracking faces in video sequences.

The accuracy values obtained for each detection method are as follows:

Table 3.4: Comparison between various detection method previously used for Chokepoint dataset.[2]

Detection Method	Accuracy
SiamRPN	0.967
SiamFC	0.951
MBMD	0.756
DVNet	0.477
MTCNN	0.967

Upon comparison, it is evident that both SiamRPN and the proposed MTCNN method achieved the highest precision scores of 0.967, showcasing their exceptional accuracy in face detection and tracking tasks. This parity in performance underscores the effectiveness of the MTCNN approach, which rivals state-of-the-art methods such as SiamRPN in terms of precision.

Furthermore, the performance of MTCNN outperformed other competing methods, including SiamFC, MBMD, and DVNet, by a significant margin. This demonstrates the superiority of the

[2] Lin, Y., Shen, J., Cheng, S., & Pantic, M. (2020). FT-RCNN: Real-time visual face tracking with region-based convolutional neural networks. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 61–68. IEEE.

proposed method in accurately detecting and tracking faces in video sequences, surpassing traditional and contemporary approaches in terms of precision and reliability.

The comparative evaluation highlights the efficacy of the proposed MTCNN method in achieving high precision and robust performance in face detection and tracking tasks. These results underscore the potential of MTCNN as a viable solution for real-world applications requiring accurate and efficient face detection and tracking capabilities.

When evaluating the Bengal dataset using confidence as a metric, we can calculate the confidence-based accuracy using the following formula:

Confidence-based Accuracy = (Number of Correct Predictions with High Confidence)/ (Total Number of Predictions with High Confidence)

Where, "Number of Correct Predictions with High Confidence" represents the instances where the algorithm correctly detects faces with high confidence scores. High confidence scores typically indicate a high degree of certainty in the algorithm's predictions. The "Total Number of Predictions with High Confidence" includes all instances where the algorithm makes predictions with confidence scores above a certain threshold, indicating its level of certainty in the predictions.

The threshold for high confidence can be determined empirically based on the distribution of confidence scores in the dataset or specified as a predefined value.



Figure 3.7: Original image frame from Bengal dataset.

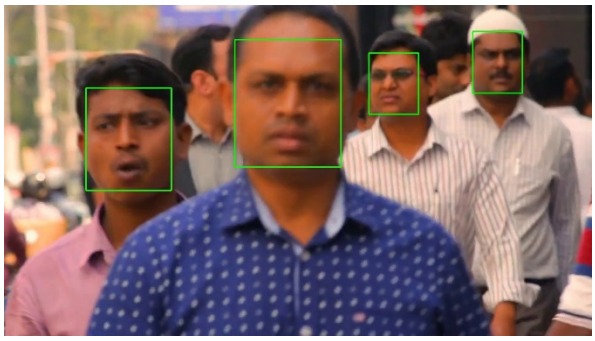


Figure 3.8: Face detected and created a bounded box.

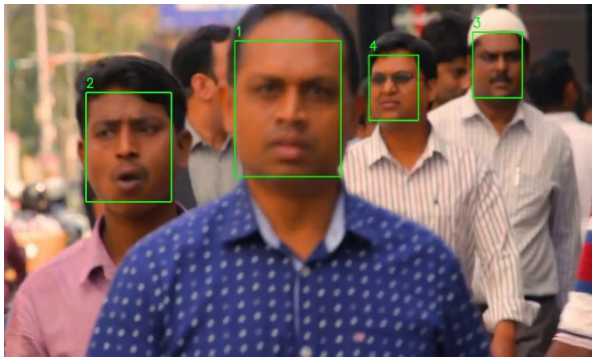


Fig 3.9: Facial landmark detected and marked.

Table 3.5: Average Confidence rate for Bengal Dataset.

Dataset	Description	Avg. Confidence Rate
Bengal	Street with crowded people are recorded and the same video has been published in YouTube for generic use in tracking and detection of human face and Object.	0.9819

The evaluation of the proposed face detection and tracking method incorporates two distinct approaches: accuracy-based evaluation for the Chokepoint dataset and confidence-based evaluation for the Bengal dataset. For the Chokepoint dataset, accuracy is calculated by comparing the detected faces and facial landmarks with their ground truth annotations, taking into account a threshold value for the difference between coordinates of the left eye. On the other hand, the evaluation of the Bengal dataset relies on confidence-based accuracy, which measures the algorithm's ability to make accurate predictions with high confidence scores. These dual evaluation strategies provide a comprehensive assessment of the algorithm's performance across different datasets and scenarios, ensuring robustness and reliability in real-world applications.

3.3 Implementation Details

The implementation of the proposed face detection and tracking method was conducted using a Google Colab setup, leveraging its cloud-based computing resources for efficient experimentation. The setup included a Google Colab runtime environment equipped with 12.7GB of RAM and a Google Cloud disk size of 107.7GB, providing ample computational resources for processing large-scale datasets and conducting experiments.

For initial testing and prototyping, basic Python compilers such as Jupyter Notebook were utilized to develop and debug the code. This allowed for rapid iteration and experimentation with different algorithms and techniques before scaling up to larger datasets and more computationally intensive tasks.

The implementation made extensive use of popular libraries and frameworks for computer vision and deep learning, including:

- ***OpenCV (cv2):*** Utilized for reading video files and converting them into individual image frames, facilitating preprocessing and analysis of video data.

- ***dlib*****:** Employed for advanced facial landmark detection and feature extraction, enabling accurate localization of facial key points.

- ***CSV and xml.etree.ElementTree (ET):*** Utilized for converting coordinates values obtained from the detected facial landmarks to CSV files and structured XML format, making it compatible with ground truth annotations for comparison and evaluation purposes.

- ***Pandas:*** Used for data manipulation and analysis, allowing for efficient processing and comparison of experiment results with ground truth annotations.

- ***Keras and TensorFlow***: Leveraged for building and training deep learning models, particularly the Multi-Task Convolutional Neural Network (MTCNN), and for evaluating the performance of the proposed method on the experimental datasets.

The workflow involved converting video data into individual image frames, performing face detection and facial landmark localization using the MTCNN model, and comparing the detected faces and facial landmarks with ground truth annotations. The implementation pipeline was structured to streamline the experimentation process and facilitate accurate evaluation of the proposed method's performance.

By leveraging a combination of cloud-based computing resources, powerful libraries and frameworks, and structured implementation workflow, the implementation of the proposed face detection and tracking method achieved efficient experimentation and robust evaluation across diverse datasets and scenarios.

3.4 Discussions

The evaluation of the proposed face detection and tracking method yielded promising results across different datasets and evaluation metrics. For the Chokepoint dataset, the accuracy achieved was an impressive 96.64%. This accuracy was calculated by comparing the detected faces and facial landmarks with their ground truth annotations, considering a threshold value for the difference between coordinates of the left eye. Furthermore, for the Bengal dataset, the algorithm demonstrated a high level of confidence, with an average confidence score of 0.9819. This confidence-based evaluation underscores the algorithm's ability to make accurate predictions with high confidence scores, enhancing its reliability and effectiveness in practical applications.

These results highlight the robustness and versatility of the proposed method in accurately detecting and tracking faces in diverse environments and scenarios. The combination of accuracy-based evaluation for the Chokepoint dataset and confidence-based evaluation for the Bengal dataset provides a comprehensive assessment of the algorithm's performance, paving the way for its deployment in real-world applications with confidence.

Chapter 4

Conclusion

The thesis presented a comprehensive exploration of face detection and tracking using the Multi-Task Convolutional Neural Network (MTCNN) in video sequences, along with a detailed investigation of experimental methodologies, results, and implementation details. Through a series of experiments conducted on the Chokepoint and Bengal datasets, the proposed method demonstrated remarkable accuracy and efficiency in detecting faces and facial landmarks across diverse environments and scenarios.

The utilization of the MTCNN architecture, comprising the Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net), facilitated robust and reliable face detection and tracking capabilities. By leveraging cloud-based computing resources and popular libraries such as OpenCV, dlib, and TensorFlow, the implementation of the proposed method achieved efficient experimentation and accurate evaluation on large-scale datasets.

The evaluation metrics employed, including accuracy and confidence-based assessment, provided a comprehensive understanding of the algorithm's performance, highlighting its effectiveness in detecting and tracking faces with high precision and confidence. Furthermore, the structured implementation workflow ensured reproducibility and scalability, enabling further exploration and refinement of the proposed method in future research endeavors.

4.1 Limitations

In the pursuit of advancing face detection and tracking methodologies, the proposed method and its implementation encounter several challenges and limitations. These obstacles include dataset bias, occlusions, varying lighting conditions, and complex backgrounds, all of which can undermine the robustness and generalization capabilities of such algorithms. Furthermore, the reliance on labeled datasets for training and evaluation introduces biases and limitations, especially when faced with real-world scenarios where data distribution may deviate from the training data.

Moreover, computational constraints and hardware limitations impose barriers on the scalability and real-time performance of face detection and tracking algorithms. These constraints become particularly pronounced in resource-constrained environments, hindering the algorithms' effectiveness and efficiency. Consequently, addressing these limitations demands collaborative efforts in data collection, algorithm refinement, and hardware innovation to ensure the reliability, efficiency, and ethical considerations of face detection and tracking systems in practical applications.

Additionally, the challenges extend to noise in images, capturing in videos, and low video quality. Noise in images can disrupt the accuracy of face detection and tracking, while capturing in videos may introduce motion blur and other artifacts that obscure facial features. Low video quality further exacerbates these issues, making it difficult for algorithms to effectively detect and track faces amidst the visual distortions.

Despite these limitations, the thesis contributes valuable insights into experimental methodologies, results, and implementation details, pushing the boundaries of face detection and tracking methodologies. While the future holds promising opportunities for further advancements and applications, it's imperative to acknowledge and address these challenges to ensure the efficacy and ethical deployment of face detection and tracking algorithms in real-world scenarios.

References

[1] Fuad, M. T. H., Fime, A. A., Sikder, D., Iftee, M. A. R., Rabbi, J., Al-Rakhami, M. S., … Islam, M. N. (2021). Recent advances in deep learning techniques for face recognition. IEEE Access: Practical Innovations, Open Solutions, 9, 99112–99142. doi:10.1109/access.2021.3096136

[2] Khan, A., Rehman, S., Waleed, M., Khan, A., Khan, U., Kamal, T., … Marwat, S. N. K. (2018). Forensic video analysis: Passive tracking system for automated person of interest (POI) localization. IEEE Access: Practical Innovations, Open Solutions, 6, 43392–43403. doi:10.1109/access.2018.2856936

[3] Ali, W., Tian, W., Din, S. U., Iradukunda, D., & Khan, A. A. (2021). Classical and modern face recognition approaches: a complete review. Multimedia Tools and Applications, 80(3), 4825–4880. doi:10.1007/s11042-020-09850-1

[4] George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., & Marcel, S. (2020). Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Transactions on Information Forensics and Security, 15, 42–55. doi:10.1109/tifs.2019.2916652

[5] Jin, R., Li, H., Pan, J., Ma, W., & Lin, J. (n.d.). Face Recognition Based on MTCNN and FaceNet. Retrieved May 11, 2024.

[6] Alqahtani, F., Banks, J., Chandran, V., & Zhang, J. (2020). 3D face tracking using stereo cameras: A review. IEEE Access: Practical Innovations, Open Solutions, 8, 94373–94393. doi:10.1109/access.2020.2994283

[7] Vij, R., & Kaushik, B. (2019). A survey on various face detecting and tracking techniques in video sequences. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 69–73. IEEE.

[8] Kalonia, S. and Upadhyay, A., 2023, September. A Systematic Review of Software Fault Prediction Using Deep Learning: Challenges and Future Perspectives. In International Conference

on Advances in Data-driven Computing and Intelligent Systems (pp. 533-548). Singapore: Springer Nature Singapore.

[9] Barquero, G., Hupont, I., & Fernandez Tena, C. (2021). Rank-based verification for long-term face tracking in crowded scenes. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(4), 495–505. doi:10.1109/tbiom.2021.3099568

[10] Naitali, A., Ridouani, M., Salahdine, F., & Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges, and research directions. Computers, 12(10), 216. doi:10.3390/computers12100216

[11] Canário, J. P., Ferreira, M. V., Freire, J., Carvalho, M., & Rios, R. (2022). A face detection ensemble to monitor the adoption of face masks inside the public transportation during the COVID-19 pandemic. Multimedia Tools and Applications, 81(23), 33689–33714. doi:10.1007/s11042-022-12806-2

[12] Feng, Y., Yu, S., Peng, H., Li, Y.-R., & Zhang, J. (2022). Detect faces efficiently: A survey and evaluations. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1), 1–18. doi:10.1109/tbiom.2021.3120412

[13] Li, X., Yang, Z., & Wu, H. (2020). Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks. IEEE Access: Practical Innovations, Open Solutions, 8, 174922–174930. doi:10.1109/access.2020.3023782

[14] Fitwi, A., Yuan, M., Nikouei, S. Y., & Chen, Y. (2020). Minor privacy protection by real-time children identification and face scrambling at the edge. EAI Endorsed Transactions on Security and Safety, 7(23), e3.

[15] Naser, O., Ahmad, S., Samsudin, K., Shafie, S., & Zamri, N. (2023). Facial recognition for partially occluded faces. Indonesian Journal of Electrical Engineering and Computer Science, 30(3), 1846-1855.

[16] Méndez, C. A., Celeghin, A., Diano, M., Orsenigo, D., Ocak, B., & Tamietto, M. (2022). A deep neural network model of the primate superior colliculus for emotion recognition. Philosophical Transactions of the Royal Society B, 377(1863), 20210512.

[17] Khalifa, A., Abdelrahman, A. A., Strazdas, D., Hintz, J., Hempel, T., & Al-Hamadi, A. (2022). Face recognition and tracking framework for human–robot interaction. Applied Sciences, 12(11), 5568.

[18] Islam, M. T., Ahmed, T., Rashid, A. R., Islam, T., Rahman, M. S., & Habib, M. T. (2022, April). Convolutional Neural Network Based Partial Face Detection. In 2022 IEEE 7th International conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.

[19] Wu, Y., & Ji, Q. (2019). Facial landmark detection: A literature survey. International Journal of Computer Vision, 127(2), 115-142.

[20] Wong, Y., Chen, S., Mau, S., Sanderson, C., & Lovell, B. C. (2011, June). Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In CVPR 2011 WORKSHOPS (pp. 74-81). IEEE.

[21] Kumar, A., Kaur, A., & Kumar, M. (2019). Face detection techniques: a review. Artificial Intelligence Review, 52, 927-948.

[22] Lin, Y., Shen, J., Cheng, S., & Pantic, M. (2020, November). FT-RCNN: Real-time Visual Face Tracking with Region-based Convolutional Neural Networks. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 61-68). IEEE.

[23] Barquero, G., Fernández, C., & Hupont, I. (2020, September). Long-term face tracking for crowded video-surveillance scenarios. In 2020 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-8). IEEE.

[24] Ren, H., Hu, Y., Myint, S. H., Hou, K., Zhang, X., Zuo, M., ... & Li, H. (2021). A real-time and long-term face tracking method using convolutional neural network and optical flow in iot-based multimedia communication systems. Wireless Communications and Mobile Computing, 2021, 1-15.

[25] Su, J., Gao, L., Li, W., Xia, Y., Cao, N., & Wang, R. (2019). Fast face tracking-by-detection algorithm for secure monitoring. Applied Sciences, 9(18), 3774.

[26] Wong, Y., Chen, S., Mau, S., Sanderson, C., & Lovell, B. C. (2011, June). Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In CVPR 2011 WORKSHOPS (pp. 74-81). IEEE.

[27] Reddy, K. S. S., Ramesh, G., Praveen, J., Surekha, P., & Sharma, A. (2023). A Real-time Automated System for Object Detection and Facial Recognition. In E3S Web of Conferences (Vol. 430, p. 01076). EDP Sciences.

[28] Suguna, G. C., Kavitha, H. S., & Sunita, S. (2021). Face recognition system for realtime applications using SVM combined with FaceNet and MTCNN. Int. J. Electr. Eng. Technol.(IJEET), 12, 328-335.

[29] Lin, Y., Cheng, S., Shen, J., & Pantic, M. (2019, May). Mobiface: A novel dataset for mobile face tracking in the wild. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

[30] Ghofrani, A., Toroghi, R. M., & Ghanbari, S. (2019, February). Realtime face-detection and emotion recognition using mtcnn and minishufflenet v2. In 2019 5th conference on knowledge based engineering and innovation (KBEI) (pp. 817-821). IEEE.

[31] Lian, Z., Shao, S., & Huang, C. (2020). A real time face tracking system based on multiple information fusion. Multimedia Tools and Applications, 79(23), 16751-16769.

[32] Gu, M., Liu, X., & Feng, J. (2022). Classroom face detection algorithm based on improved MTCNN. Signal, Image and Video Processing, 16(5), 1355-1362.

[33] Ge, H., Dai, Y., Zhu, Z., & Wang, B. (2021). Robust face recognition based on multi-task convolutional neural network. Math Biosci Eng, 18(5), 6638-6651.

[34] Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd.

[35] Bartlett, M. S., Movellan, J. R., & Sejnowski, T. J. (2002). Face recognition by independent component analysis. IEEE Transactions on neural networks, 13(6), 1450-1464.