A Multi-Stage Deep Learning Framework for Face Detection and Facial Landmark Localization Using MTCNN

Prof. Dipti Mehare[1], Rutuja Jogi[2], Janhvi Hiwe[3], Anjali Tamte[4], Vaishnavi Mahulkar[5], Sakshi Likhitkar[6], Sneha Deshmukh[7]
[1,2,3,4,5,6,7]*Computer Science Engineering Department, PRPCEM*

Abstract- **Detecting human faces in images is an important task in the field of computer vision. It is widely applied in systems such as security surveillance, biometric verification, attendance tracking, and human–computer interaction. However, traditional face detection approaches often face difficulties when images contain different lighting conditions, head poses, image scales, or partially hidden faces. To address these challenges, deep learning techniques have been introduced. One such approach is the Multi-task Cascaded Convolutional Neural Network (MTCNN), which is designed to detect faces while also identifying important facial landmarks. This paper studies the structure and working of the MTCNN model in detail. The architecture is composed of three sequential neural networks: the Proposal Network (P-Net), the Refinement Network (R-Net), and the Output Network (O-Net). Each stage progressively improves the accuracy of face detection and removes incorrect detections. The performance of the model is analyzed using well-known datasets such as the WIDER FACE dataset and the FDDB dataset. In addition, the effectiveness of MTCNN is examined by comparing it with other popular models like YOLO, FaceNet, and Dlib. The findings show that the cascaded architecture of MTCNN helps achieve reliable face detection accuracy while maintaining reasonable computational efficiency. Overall, the study demonstrates that MTCNN is a practical solution for face detection tasks in real-world scenarios**

Keywords: Face Detection, MTCNN, Deep Learning, Computer Vision, Convolutional Neural Networks.

I. INTRODUCTION

Face detection is one of the most important tasks in the field of computer vision. It is commonly used in applications such as security monitoring, biometric authentication, smart attendance systems, and social media image analysis. Detecting faces accurately in images or videos is often the first step before performing other tasks like face recognition or emotion detection.

Earlier face detection techniques relied on traditional algorithms and handcrafted features. One of the most popular methods was the Viola–Jones algorithm, which was widely used for real-time face detection. Although these approaches were fast, they were not very effective when faces appeared in different orientations, lighting conditions, or when parts of the face were hidden.

With the advancement of deep learning, convolutional neural networks (CNNs) have significantly improved the performance of face detection systems. Deep learning models can automatically learn important visual features from large datasets, making them more accurate and robust than traditional methods.

The Multi-task Cascaded Convolutional Neural Network (MTCNN) is a deep learning approach designed specifically for face detection and facial landmark localization. The model uses a cascade of three neural networks that work together to detect faces with higher accuracy. Each stage of the network refines the detection results and removes incorrect predictions.

This paper presents an overview of the MTCNN model, including its architecture, working process, and performance evaluation. The model is also compared with other popular approaches such as YOLO, FaceNet, and Dlib to understand its advantages and limitations.

II. LITERATURE REVIEW

Many researchers have proposed different techniques for face detection over the years. Early approaches mainly relied on feature-based and machine learning methods.

One of the earliest and most well-known methods is the Viola–Jones face detection framework. This method used Haar-like features and a cascade classifier to detect faces quickly in images. Although it worked well for frontal faces, it had difficulty handling complex scenarios such as variations in lighting and face orientation.

Later methods introduced feature extraction techniques such as Histogram of Oriented Gradients (HOG) and Scale Invariant Feature Transform (SIFT). These techniques improved detection performance but still required manual feature design.

In recent years, deep learning has become the dominant approach in computer vision tasks. Convolutional neural networks have demonstrated strong performance in image recognition and object detection problems. Models such as YOLO are designed for fast object detection in real time, while FaceNet focuses on generating embeddings for face recognition.

The Multi-task Cascaded Convolutional Neural Network (MTCNN) combines the advantages of deep learning with a cascaded architecture. It performs both face detection and facial landmark localization simultaneously. This multi-task learning approach improves accuracy while reducing false detections.
Due to its effectiveness and efficiency, MTCNN has become a widely used model in many face detection and recognition systems.

III. MTCNN ARCHITECTURE

The MTCNN model is designed using a cascaded structure that consists of three convolutional neural networks. These networks work sequentially to detect faces and improve detection accuracy at each stage.The methodology describes the overall process used to implement the face detection system based on the Multi-task Cascaded Convolutional Neural Network (MTCNN). The proposed approach follows a sequence of steps including data preparation, image preprocessing, face detection using cascaded neural networks, and final result generation. The aim of this methodology is to accurately detect faces and identify facial landmark points in images.

A. Data Collection

The first step in the methodology is selecting suitable datasets for training and evaluation. In this study, widely used face detection datasets such as the WIDER FACE dataset and the FDDB dataset are used. These datasets contain thousands of images with faces appearing in different conditions such as various lighting environments, different poses, and partially hidden faces. Using diverse datasets helps the model learn a wide range of facial features and improves detection accuracy.

B. Image Preprocessing

Before feeding images into the neural network, preprocessing is performed to improve the quality of the input data. Preprocessing helps standardize the images and ensures that they are suitable for processing by the deep learning model.

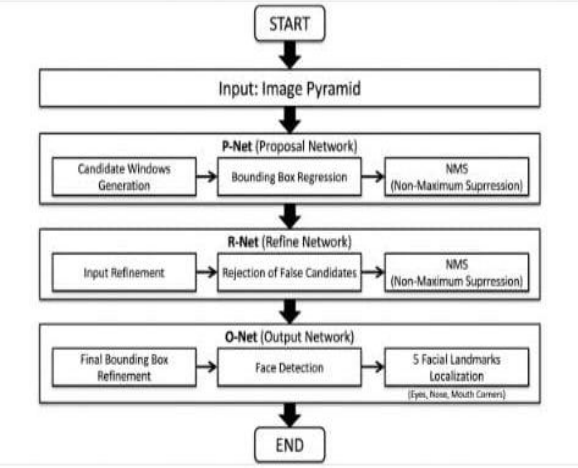
Typical preprocessing steps include:

- Image resizing: Images are resized to a suitable resolution so that the model can process them efficiently.
- Normalization: Pixel values are normalized to maintain consistency across images.
- Image pyramid generation: Multiple scaled versions of the same image are created. This allows the model to detect faces of different sizes.

The preprocessing stage is important because it improves the ability of the model to detect faces accurately under different conditions.

C. Face Detection using MTCNN

The core part of the methodology is the use of the MTCNN model. The model is composed of three convolutional neural networks that operate in a cascaded manner. Each stage processes the output of the previous stage and refines the detection results.



1. Proposal Network (P-Net)

The Proposal Network is responsible for scanning the entire image and identifying potential face regions. It performs a convolution operation on different parts of the image to generate candidate bounding boxes that might contain faces.
Since P-Net focuses on speed, it generates many possible face regions, including some incorrect detections. These candidate regions are passed to the next stage for further analysis.

2. Refinement Network (R-Net)

The Refinement Network receives the candidate regions produced by P-Net. It performs additional filtering to remove incorrect detections and improves the accuracy of the bounding boxes.

R-Net analyzes the candidate regions more carefully and adjusts the coordinates of the bounding boxes so that they better match the actual face location in the image.

3. Output Network (O-Net)

The Output Network is the final stage of the MTCNN architecture. It performs detailed classification of the remaining candidate regions and produces the final face detection results.
In addition to detecting faces, O-Net also predicts facial landmark points such as the position of the eyes, nose, and mouth. These landmarks are useful for face alignment and further face recognition tasks.

D. Bounding Box Refinement and Non-Maximum Suppression
After detection, multiple bounding boxes may overlap around the same face. To handle this issue, a technique known as **Non-Maximum Suppression (NMS)** is applied.
NMS removes duplicate bounding boxes and keeps only the most accurate one. This step ensures that each face is detected only once and improves the overall quality of the detection results.

E. Facial Landmark Localization
One of the key features of MTCNN is its ability to identify facial landmark points. The Output Network predicts the locations of important points on the face such as:

- Left eye
- Right eye
- Nose
- Left corner of the mouth
- Right corner of the mouth

These landmark points are essential for tasks such as face alignment, emotion recognition, and face recognition.

F. Performance Evaluation
The final step in the methodology is evaluating the performance of the face detection system. The performance of the model can be measured using metrics such as accuracy, precision, recall, and detection speed.
The results obtained from the MTCNN model are also compared with other models such as YOLO, FaceNet, and Dlib. This comparison helps determine the effectiveness of the proposed approach.
For evaluating the performance of the face detection model, publicly available datasets are commonly used. In this study, two widely used datasets were considered: the WIDER FACE dataset and the FDDB dataset. These datasets contain a large number of images with faces in different environments and conditions.
The WIDER FACE dataset is one of the most challenging datasets for face detection. It includes images with a large variety of faces captured under different lighting conditions, poses, and occlusions.

Because of this diversity, the dataset is widely used to evaluate the robustness of face detection algorithms.
The FDDB dataset, also known as the Face Detection Data Set and Benchmark, contains images with annotated face regions. It is often used as a benchmark dataset to measure the performance of face detection systems.

Dataset Summary Table

Dataset Name	Number of Images	Description
WIDER FACE	32,000+	Large dataset with faces in various poses, lighting, and occlusion conditions
FDDB	2,845	Benchmark dataset used for evaluating face detection models

These datasets help test the ability of the model to detect faces in both controlled and real-world environments.

V. IMPLEMENTATION

The implementation of the MTCNN face detection system involves several steps, including image preprocessing, candidate region generation, and final detection.

First, the input image is processed and resized to create multiple image scales. This technique, known as an image pyramid, helps the model detect faces of different sizes.
Next, the first network in the MTCNN architecture, the Proposal Network (P-Net), scans the image and generates candidate bounding boxes where faces might be located.
These candidate regions are then passed to the second network, the Refinement Network (R-Net). This network removes false detections and adjusts the bounding boxes to better fit the face regions.
Finally, the Output Network (O-Net) performs detailed analysis and produces the final detection results. In addition to detecting faces, this network also identifies facial landmarks such as the eyes, nose, and mouth.
The entire process allows the system to detect faces accurately while maintaining efficient processing speed.

VI. RESULTS AND DISCUSSION

The experimental evaluation shows that the MTCNN model performs effectively in detecting faces in images with varying conditions. The cascaded architecture allows the system to filter out non-face regions early in the detection process, which helps reduce computational cost.
Another advantage of the model is its ability to detect facial landmarks along with face regions. This feature improves the performance of face alignment and recognition tasks.
When tested on challenging datasets, the model demonstrates good detection accuracy even when faces are partially hidden or appear at different angles.
The use of deep learning also allows the system to learn complex visual patterns directly from the training data.
Overall, the results indicate that the MTCNN framework is suitable for real-world face detection applications.

VII. COMPARISON WITH OTHER MODELS

To understand the effectiveness of the proposed approach, the MTCNN model was compared with several other commonly used face detection and recognition models.

Model	Main Purpose	Accuracy	Speed
MTCNN	Face Detection and Landmark Detection	High	Medium
YOLO	Real-time Object Detection	High	Very Fast
FaceNet	Face Recognition	Very High	Medium
Dlib	Face Detection	Medium	Fast

From the comparison, it can be observed that MTCNN provides a good balance between accuracy and computational efficiency. Unlike some models that only detect objects, MTCNN also identifies facial landmarks, which improves overall face analysis tasks.

VIII. CONCLUSION

In this paper, a detailed study of the Multi-task Cascaded Convolutional Neural Network (MTCNN) for face detection has been presented. The proposed approach focuses on improving the accuracy and reliability of detecting human faces in images by using a cascaded deep learning architecture. The MTCNN model is composed of three neural networks—Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net)—that work together in a sequential manner. Each stage of the network progressively filters and refines candidate face regions, which helps in reducing false detections and improving the overall detection performance.

The experimental results demonstrate that the cascaded structure of MTCNN plays an important role in enhancing the face detection process. By gradually eliminating non-face regions and refining bounding boxes, the model is able to detect faces more accurately even in complex conditions. The ability of the model to detect facial landmark points such as the eyes, nose, and mouth further improves its usefulness in applications like face alignment and recognition. This additional feature makes MTCNN more effective than many traditional face detection methods.

Another important advantage of the MTCNN model is its capability to handle real-world challenges such as variations in lighting conditions, different face orientations, and partially hidden faces. These factors often affect the performance of traditional face detection algorithms. However, due to its deep learning architecture and multi-stage detection process, MTCNN is able to adapt to these variations and produce reliable detection results.

The performance of the MTCNN model was also compared with other widely used models including YOLO, FaceNet, and Dlib. While YOLO is mainly designed for general object detection and FaceNet is focused on face recognition, MTCNN provides a specialized solution for face detection along with landmark localization. The comparison indicates that MTCNN achieves a good balance between detection accuracy and computational efficiency, making it suitable for many real-world applications.

Despite its advantages, there is still room for improvement in the model. Future research can focus on optimizing the architecture to improve processing speed for real-time applications such as surveillance systems and smart attendance systems. In addition, integrating MTCNN with advanced face recognition models could lead to the development of more powerful and intelligent biometric systems.

Overall, the study highlights the effectiveness of the MTCNN framework for face detection tasks and demonstrates its potential for use in modern computer vision applications.

REFERENCES

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

[2] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified Real-Time Object Detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[4] D. E. King, “Dlib-ml: A Machine Learning Toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[5] S. S. Khan, D. Sengupta, and A. Ghosh, “MTCNN++: A CNN-Based Face Detection Algorithm Inspired by MTCNN,” *The Visual Computer*, vol. 40, pp. 899–917, 2024.

[6] A. C. Ömercikoğlu, M. M. Yönügül, and P. Erdoğmuş, “The Impact of Image Resolution on Face Detection: A Comparative Analysis of MTCNN, YOLOv11 and YOLOv12 Models,” *arXiv preprint*, 2025.

[7] R. Jahan and S. Mariyam, “Robust Facial Recognition Using Deep Learning with MTCNN,” *International Journal of Engineering Research & Technology*, vol. 13, no. 6, 2025.

[8] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[9] H. Otroshi Shahreza *et al.*, "SDFR: Synthetic Data for Face Recognition Competition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2024.

[10] P. Melzi *et al.*, "FRCSyn Challenge at WACV 2024: Face Recognition Challenge in the Era of Synthetic Data," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2024.