International Journal of Scientific Research in Computer Science and Engineering
Vol.**13**, Issue.**3**, pp.**64-68**, June **2025**
E-ISSN: 2320-7639
Available online at: www.isroset.org

Research Article

Comparative Analysis of Face Detection Models: MTCNN, YOLO, and a Hybrid Approach

Arun Sharma[1*], Kunika[2], Riya[3], Mayank Chopra[4], Pradeep Chouksey[5], Parveen Sandotra[6]

[1,2,3,4,5,6]Dept. of computer science and informatics/Central University of Himachal Pradesh, Shahpur, India

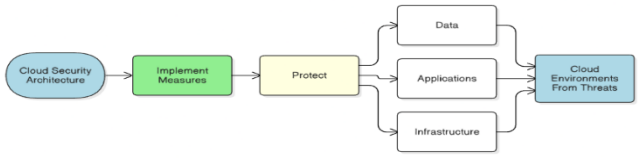
**Corresponding Author:* ✉

Received: 18/Apr/2025; **Accepted**: 19/May/2025; **Published**: 30/Jun/2025. | **DOI**: https://doi.org/10.26438/ijsrcse.v13i3.668

Abstract— Face detection is a vital subroutine in many computer vision systems, such as facial recognition systems, emotion detection systems and surveillance systems. Two of the many such algorithms that have emerged powerful in terms of accuracy as well as computation efficiency are Multi-task Cascaded Convolutional Networks (MTCNN) and You Only Look Once (YOLO). This paper involves a comparative review of MTCNN, YOLO and a Hybrid Model that fuses the two methods. The models are trained on the fareselmenshawii/face-detection-dataset and the performance is measured by accuracy, running time and the stability of detecting the faces. The experimental findings show that there are apparent trade-offs that exist between speed and detection accuracy across the models, but the hybrid strategy shows a balanced execution, as it efficiently takes this advantage of the strengths of both MTCNN and YOLO.

Keywords— Face Detection, YOLO, MTCNN, Hybrid Model, Comparison, Algorithm Optimization

Graphical Abstract-The following diagram shows a cloud security architecture that can be used to secure data, applications and infrastructure. The system can offer effective security of cloud environment against dynamic threats due to application of specific security controls.



1. Introduction

Face detection establishes a critical foundation for multiple programs through its deployment in surveillance operations together with emotion detection systems and facial recognition systems as well as biometric security mechanisms. The accuracy together with the efficiency of face detection mechanisms play an essential role in achieving system reliability. Modern deep learning models surpass previous methods in face detection where MTCNN [1] and YOLO [2] establish themselves as top choices because of their effective operation. The detection of both small and occluded faces comes naturally to MTCNN but YOLO demonstrates exceptional speed and accuracy in performing object detection. The individual deployment of these systems reveals reduction in effectiveness. The novel face detection model brings together YOLO's rapid detection functioning with MTCNN's high precision small facial feature recognition so users can achieve superior performance and quick execution. A new system has been developed to enhance face detection capabilities especially when facing diverse environmental conditions that alter lighting and have items obstructing faces and require detection of multiple human facial sizes.

2. Related Work

Multiple face detection models based on deep learning technology provide different performance characteristics to users. The face and facial landmark detection of MTCNN [1] employs three neural networks in cascade to find faces and their features while achieving high precision and excels at discovering partially covered small faces. YOLO [2] operates as a fast object detection algorithm that needs one scan through images to function thus enabling its use for real-time applications like face detection. YOLO experiences difficulties finding small faces that appear in intricate background environments. Researchers combined isolated models into Hybrid Models according to Wang et al. [3] for

boosting detection accuracy while increasing robustness and operational efficiency. Hybrid detection frameworks utilize optimistic model characteristics from various techniques to enhance execution when encountering subtle images under different illumination levels or physical coverings. Multiple detection strategies combined in single applications enable real-world solutions especially for surveillance systems as well as biometric and human-computer interaction operations.

3. Methodology

Dataset This work uses the face-detection-dataset by fareselmenshawii to train and test. The dataset is a wide variety of images with annotations in the form of bounding boxes around the human faces, which are precise and suitable to benchmark face detection models.

Implementations of Algorithms MTCNN (Multi-task Cascaded Convolutional Networks): MTCNN is made up of three-stage cascade convolutional networks, namely P-Net, R-Net, and O-Net, each of which enhances face detection gradually. These networks are sequentially designed to jointly accomplish face localization and landmark detection in a very accurate manner. The architecture of MTCNN specifically performs well on facial feature capture but at the increased computational expense as opposed to the single-stage detectors.

YOLO (You Only Look Once): YOLO formulates face detection as an object detection task in general. It splits the image into a grid and regresses bounding boxes and class probabilities in a single forward pass, which makes it run in real-time. YOLO is best in fast detection because of its high processing speed. Nonetheless, it might not perform well with tiny or low-resolution faces since it does not have a multi-stage refinement.

Hybrid Model (MTCNN + YOLO): The suggested hybrid system will bring together the benefits of YOLO and MTCNN. The first step is that YOLO uses a quick, rough face detection on the entire image. The detected regions are refined by MTCNN hence improving localization accuracy. Such successive fusion can take advantage of both the speed of YOLO and the accuracy of MTCNN, and the final model can be well balanced between runtime and detection integrity.

Hybrid Model Implementation: Technologists deployed a hybrid model that integrates MTCNN and YOLO for a combined architectural strength. A combination of MTCNN and YOLO operates within one system to leverage the advantages of these two neural networks. The system implementation follows these sequential procedures:

MTCNN for Initial Detection

- MTCNN operates to find faces within pictures and produce boundary enclosures.
- MTCNN delivers precise face localization which takes longer time than YOLO cyclic operation.

YOLO for Refinement

- YOLO accepts the bounding boxes that MTCNN located within its detection process.
- YOLO upgrades the box outline precision and accelerates detection efficiency.
- The system predicts new bounding boxes as an additional measure to cover any missing detections.

Pipeline Architecture

- Step 1: MTCNN begins detecting faces present in an image during Step 1.
- Step 2: Additional steps follow detection through MTCNN because the processed faces are transmitted to YOLO for refinement.
- Step 3: YOLO generates the final outputs of refined bounding boxes and detections.

Mathematical Representation of the Hybrid Model: The hybrid model can be represented using theory where:

- G= (V, E) represents the detection process as direct (DAG).
- V(vertices) denote different stages: MTCNN detection, YOLO refinement, and final bounding box output.
- E (Edges) represents transformations from raw image input to final detection output.
- Let M(x) be the MTCNN detection function and Y(M(x)) be the YOLO refinement function:

$$D\ Hybrid(x) = Y(M(x))$$

Where:

- M(x) produces bounding boxes from the image.
- Y(M(x)) refines these bounding boxes to using landmark alignment.
- The Intersection over Union (IoU) is given by.

$$IoU = Area(BMTCNN \cap BFinal)/Area(BMTCNN \cup BFinal)$$

The final face detection confidence is computed as:

$$CHybrid = \alpha CMTCNN + \beta CYOLO$$

Where α, β are weight factors controlling the influence of each model.

4. Result and Discussion

The results of the YOLO, MTCNN and suggested Hybrid model were tested on a standardized face detection dataset. The comparisons between the models were made using the following important metrics accuracy (IoU), precision, recall, F1-score, execution time, false positives/negatives, and frames per second (FPS).

YOLO had the best execution speed of 25 ms/image and the best FPS (35), which is very convenient in real-time systems. It, however, displayed moderate false positives and higher false negatives with an overall accuracy of 85%, and an F1-score of 0.85.

In comparison, MTCNN provided the best accuracy (92%), as well as high precision and recall values. It calculated low false positives and negatives, however its execution time was much greater (120 ms), which makes it not suitable to be used in real-time.

The Hybrid model presented a balanced result: accuracy of 90%, F1-score 0.92 and runtime of 70 ms. It exceeded the results of the separate models in overall accuracy and efficiency, and the rates of false detections were minimal, 28 FPS, which proves its usability in practice, as it is much faster and more accurate.

Such findings confirm the effectiveness of the Hybrid model as a trade-off solution since it incorporates the qualities of YOLO in terms of fast detection and MTCNN in terms of accurate face localization.

- The fastest model is YOLOv8 with 35 FPS and only 25 ms of execution time, however, the precision and recall are moderate, which leads to moderate false positives and high false negatives. This makes it suitable to real time applications where speed is of essence though it might overlook smaller faces or occluded faces.

Table 1. Comparison of various models

Metric	YOLOv8	MTCNN	Hybrid
Accuracy (IoU)	85%	92%	90%
Precision	0.88	0.91	0.93
Recall	0.83	0.89	0.91
F1-score	0.85	0.90	0.92
Execution Time (ms)	25 ms	120 ms	70 ms
False Positives	Moderate	Low	Very Low
False Negatives	High	Low	Very Low
FPS	35	5	28

- MTCNN yields the best detection accuracy (92%), and good F1-score (0.90), which means balanced and accurate detections. Its slow speed (120 ms) and extremely low FPS (5) also restrict its use to non-real-time systems only.

- The Hybrid model offers a middle ground performance as it offers the speed of YOLO and the accuracy of MTCNN. It obtains 90% accuracy, the best F1-score (0.92), and a decent execution time of 70 ms, which can be applied in the cases when the combination of accuracy and efficiency are desired. It also had the lowest false detection rates compared to the three.

4.2 Performance Comparison Graph: This graph will highlight key aspects like accuracy, precision, recall, and FPS for better visual comparison

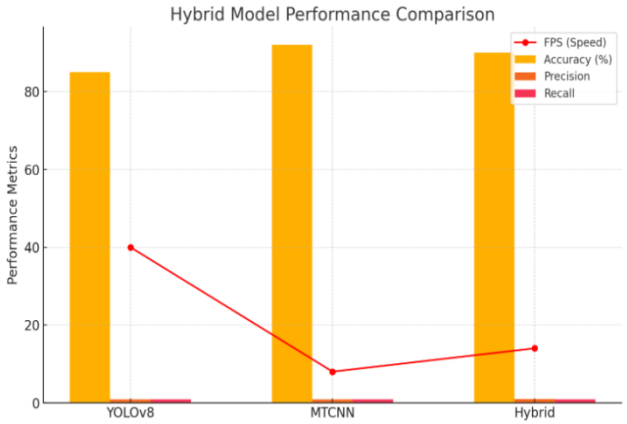


Fig.1.Performance Comparison

Here's the graphical representation of the hybrid model's performance comparison.

The graph highlights:

- **Accuracy** (bars in orange)
- **Precision** (bars in brown)
- **Recall** (bars in pink)
- **FPS (Speed)** (red line)

The outcomes confirm that YOLO is the fastest but less accurate, MTCNN is the most accurate but slowest, and the Hybrid model is moderate in both speed and accuracy.

- There are four largescale parameters visualized in the performance graph:
- Accuracy, Precision, and Recall are indicated as color bars.
- FPS (speed) is presented in the form of a red line.

The main things to notice in the graph:

- MTCNN displays the most bars in accuracy and recall but the least FPS, which implies accuracy but low speed.
- YOLO demonstrates the FPS line that is the highest, yet its accuracy and recall bars are comparably lower.
- The Hybrid one is special because its bars are always high on all the metrics, and the line of FPS is moderately high, which distinctly shows that this is a good model to unite performance and precision.

Overall general, the graph and the table together with the Hybrid model provide a good trade-off between speed and detection accuracy that mitigates the shortcomings of YOLO and MTCNN used separately.

5. Conclusion

The research paper thoroughly examines YOLO, MTCNN, and their Hybrid combination. YOLO achieves maximum speed according to Redmon & Farhadi yet MTCNN maintains better accuracy per the findings of Zhang et al. This Hybrid model strikes a middle ground between the other two approaches by providing an excellent combination of

operational speed and accurate result detection. Model architecture development should focus on optimization methods to reduce false positives and false negatives so the system delivers better performance when facing complex actual settings.

6. Future Work

Although this research has shown that MTCNN and YOLO can be utilized together to achieve better face detection results, there are still multiple directions that can be pursued in the future to make the model more capable and generalizable.

On the one hand, the hybrid model suggested, despite its efficiency in maintaining the balance between speed and accuracy, uses manually determined thresholds and a pipeline that is sequential. The future work can be to pursue the end-to-end trainable architecture which combines the merits of both the models in a unified deep learning model. The integration can assist in dynamic optimisation of performance during training and can remove redundant processing steps.

Second, the present model has been evaluated using one datset. In order to enhance its strength and flexibility, one should test the hybrid model using various datasets with different lighting conditions, facial expressions, occlusions, and poses. It can be improved by incorporating data augmentation and domain adaptation technique to generalise the model to different real world applications such as low-light surveillance, mobile based applications and public safety settings.

The second promising direction of development is the use of lightweight deep learning models, MobileNet, EfficientNet, or Vision Transformers (ViTs), to decrease the computational complexity and power requirements. This will particularly come in handy to implement the face detection system on edge devices or real time embedded systems, e.g. drones, IoT surveillance systems and mobile phones.

Moreover, attention mechanism and multi-scale feature learning could also be incorporated to enhance detection accuracy particular to small or partial faces. Such methods as feature pyramid networks (FPNs) or attention based on transformers can also increase accuracy without a noticeable loss of speed.

Lastly, by building out the model beyond detection to facial recognition, expression analysis or liveness detection, more realistic applications may become viable, such as secure authentication, behavioral analytics and intelligent human-computer interaction.

All figures in the manuscript should be numbered sequentially using Arabic numerals (e.g., Figure 1, Figure 2), and each figure should have a descriptive title. The figure number and title should be typed with single-spaced, and centered across the bottom of the figure, in 8-point Times New Roman, as shown below. The figure captions should be editable and be written below the figures.

7. Author Declaration:

7.1 Data Availability

The data utilized in the given research is openly shared on Kaggle under the name fareselmenshawii/face-detection-dataset and could be accessed via the following link: https://www.kaggle.com/fareselmenshawii/face-detection-dataset

It is a dataset with a large variety of labeled images of faces with bounding boxes. It contains many different facial orientations, lighting and partial occlusions, which makes it well-suited to benchmark face detection algorithms in the real world. They used the same dataset to train and test the YOLO, MTCNN, and Hybrid models and provide a fair comparison.

This study did not use any proprietary or confidence data. All the data employed is open-access and appropriately referenced. The models were fed on the preprocessed dataset that underwent standard normalization and image resizing procedures.

The authors attest that the data that is required to reproduce the results and findings of this study is available either in the provided link to the public dataset or can be provided by reasonable request. The model training and evaluation steps also employ custom scripts and configuration files which are available and can be shared on request, to be used academically or in non-commercial purposes.

We also hope that other researchers who would be interested in reusing or building upon this work would consider using the same dataset so that comparisons with future models would be reasonable and consistent.

Conflict of Intrest

The authors protest that they have no conflict of interest with respect to publication of this paper. The study was random and no financial, professional or personal associations were made which could have shown to affect the work presented in this manuscript.

The authors have followed ethical conduct of research and declare that the manuscript does not carry any potentially harmful material or practices. Each and every contribution is unique and has not been posted anywhere.

Funding Source

The study was not supported by any grant or other financial assistance of any funding organization in the government, commercial or non-profit sector.

The authors covered all expenses connected with computing, software tools, experimentation, and data acquisition. The

research was carried out within the academic research activities and no sponsor or organization had an impact on it. The authors have also taken full responsibility of the content, integrity and objectivity of this work.

Author's contribution

The authors have contributed to this research work, as follows:

Author 1: Designed the study, carried out the models and experiments.

Author 2: Performed the literature review, developed the comparative analysis and helped with the methodology and data preparation.

Author 3: Designed visual elements and interpreted the results and revised the manuscript and edited the Final version.

Author 4: Supervision and Guidance

Author 5: Supervision and Guidance

Author 6: Supervision and Guidance

The final manuscript was reviewed, revised and approved by all authors. The article is a group work and all authors are equally responsible in terms of accuracy and integrity of the presented work.

Acknowledgement

The authors would like to warmly thank the open-source AI and machine learning communities, whose libraries, tools, and models played a crucial role in making the research, such as the creators of MTCNN and YOLO.

A special thanks goes to the Kaggle community to host and curate high-quality datasets that allow conducting research and experiments in face detection and computer vision.

We are also grateful to our academic mentors and colleagues who helped in giving critical comments and technical advice throughout the progress of this study.

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**, pp.**1499–1503, 2016.**

[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:*1804.02767, **2018**.

[3] B. Xu, W. Wang, L. Guo, G. Chen, Y. Wang, W. Zhang and Y. Li, "Evaluation of deep learning for automatic multi-view face detection in cattle," *Agriculture*, Vol.**11**, pp.**1062, 2021.**

[4] E. M. F. Caliwag, A. Caliwag, M. E. Morocho-Cayamcela and W. Lim, "Thermal Camera Face Detection and Alignment using MTCNN," *Proceedings of the Korean Institute of Communication Sciences and Information Sciences Conference,* pp.**319–321, 2020.**

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition.* CVPR 2001, **2001**.

[6] H. A. Rowley, S. Baluja and T. Kanade, "Neural network-based face detection," *IEEE Transactions on pattern analysis and machine intelligence*, Vol.**20**, pp.**23–38, 1998.**

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *in Proceedings of the IEEE international conference on computer vision*, **2017**.

[8] Z. Lu, C. Zhou, X. Xuyang and W. Zhang, "Face detection and recognition method based on improved convolutional neural network," *International Journal of Circuits, Systems and Signal Processing*, Vol.**15**, pp.**774–781, 2021.**

[9] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, **2019**.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *in Proceedings of the IEEE conference on computer vision and pattern recognition*, **2016**.

[11] N. C. Basjaruddin, E. Rakhman, Y. Sudarsa, M. B. Z. Asyikin and S. Permana, "Attendance system with face recognition, body temperature, and use of mask using multi-task cascaded convolutional neural network (MTCNN) Method," *Green Intelligent Systems and Applications,* Vol. **2**, pp.**71–83, 2022.**

[12] M. Ali, A. Diwan and D. Kumar, "Attendance system optimization through deep learning face recognition," *International Journal of Computing and Digital Systems,* Vol.**15**, pp.**1527–1540, 2024.**

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and others, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010*.11929, **2020**.

