Face and Object Detection Algorithms for People Counting Applications

S.V. Vasantha   
Department of CSE *Vardhaman College of Engineering*Hyderabad, Telangana, India

s.v.vasantha@gmail.com

Syeda Sufia Hashmi  
Department of IT  
*MJCET*Hyderabad,Telangana,India

sufia29814@gmail.com

B. Kiranmai  
Department of CSE  
*KMIT*Hyderabad, Telangana, India  
kiranmaimtech@gmail.com

Leema Nelson

*Chitkara University Institute of Engineering & Technology, Chitkara University,*

Punjab, India

[Leema.nelson@chitkara.edu.in](mailto:Leema.nelson@chitkara.edu.in)

Mohammed Ashraf Hussain  
Department of CSE *Vardhaman College of Engineering*Hyderabad, Telangana, India  
mdashrafhussain1156@gmail.com

S. Hariharan

Department of CSE

*Vardhaman College of Engineering,*

Hyderabad, India

[hariharan.s@vardhaman.org](mailto:hariharan.s@vardhaman.org)

*Abstract*— This paper delves into the significance of people counting systems, emphasizing their role in furnishing valuable data for operational improvement, security enhancement, and resource optimization in various businesses and organizations. The focus is on a thorough examination of face detection and object detection methodologies rooted in computer vision and deep learning. Specifically, the study scrutinizes Multi-task Cascaded Convolutional Networks (MTCNN) and YuNet for face detection, along with the Histogram of Oriented Gradients (HOG) feature descriptor coupled with Support Vector Machine (SVM) and the real-time capabilities of You Only Look Once version 8 (YOLOv8) for object detection. Through empirical evaluation across diverse conditions and comparative analysis, this research aims to elucidate the strengths and limitations of these algorithms in the context of people counting tasks. The findings provide valuable insights to guide the selection of suitable approaches for specific use cases, contributing to the ongoing progress in the field of computer vision.

Keywords—People Counting, Face Detection, Object Detection, YuNet, Yolo v8.

# Introduction

In the domain of computer vision, the accurate identification and localization of objects and faces represent essential endeavors with profound implications across a myriad of applications. These applications range from enhancing security and surveillance systems to facilitating automated attendance management in academic institutions. It is useful for the estimation of people in public rallies[1], suspicious person detection, resource or food management and tracking of people in emergency circumstances[2]. In recent times, people and crowd analysis has gained significant attention in the field of computer vision and artificial intelligence. This area of study focuses on understanding and extracting valuable information from images and video footage containing people, with a particular emphasis on analyzing the behavior, demographics, and movements of individuals and groups within crowded environments. In the context of this study, we delve into two distinctive yet complementary approaches: face detection and object detection, each employing a set of efficient techniques.

Face detection technology continues to advance drawing from disciplines like image processing [3], neural networks [4], and machine learning [5], boasts a wide range of applications, including access control, emotion analysis, payment solutions, facial recognition, systems for biometric identification, search, and security etc. It utilizes advanced computing algorithms to automatically capture and analyze facial characteristics, enabling the authentication and validation of individuals. Distinguished from other biometric methods, facial recognition possesses unique merits: it enjoys broad public acceptance and yields substantial personal insights. It not only enhances people's everyday lives but also holds considerable economic significance. In surveillance systems, they are frequently employed in tandem with crowd counting to pinpoint specific individuals within a crowd.

Object detection deals with precisely locating and counting objects of interest in digital images forms the basis for numerous technological advancements and solutions. They have seen significant advancements in recent years, primarily due to the adoption of deep learning techniques, such as CNNs[6]. Various architectures, including Region-based CNNs(R-CNN), FasterR-CNN, Single Shot Multi-Box Detector(SSD), and YouOnlyLookOnce(YOLO)[7], have contributed to the improved performance and adaptability of these algorithms. They find applications in diverse fields, from autonomous vehicles and surveillance to healthcare and retail, and continue to evolve to meet the growing demand for accurate and efficient object detection solutions.

The choice between face and object detection for people counting applications depends on the characteristics of the environment in which the counting is to be conducted. These approaches have made significant advancements but are not without their challenges and issues. Some of the notable problems and concerns include, privacy, bias, accuracy, objects viewed from different angles may look completely different, person ageing, changes in appearance, complex backgrounds, object annotation, object occlusion, interoperability challenges across different devices and application tasks [8]. Face detection techniques are particularly suitable for specific environments where there is relatively good visibility of individual’s faces. These methods work well when the counting scenario allows for clear and unobstructed views of people's faces. For example, in controlled indoor settings, like retail stores or office lobbies, where individuals are expected to face cameras when passing through entrances or checkpoints, face detection can be highly effective. This method relies on recognizing and tracking faces in the field of view of the camera.

On the other hand, object detection techniques are more appropriate in scenarios where large crowds are present, or where visibility of individuals' faces may be limited. Object detection doesn't rely solely on identifying faces but can detect and count individuals based on their entire body or other distinguishing features. In crowded public spaces such as stadiums, airports, or outdoor events, object detection can provide accurate people counting even when people are facing different directions, partially obscured, or in motion. The key distinction here is that face detection primarily focuses on recognizing and counting individuals based on their facial features, which are visible in specific, controlled settings. In contrast, object detection takes a broader approach, considering the entire body or other distinctive characteristics, making it more suitable for situations where large crowds or challenging viewing conditions are common.

# Related Work

TensorFlow-based algorithm developed to detect people and track animated object movements in video sequences. It divides a video into a series of individual images, identifies objects or entities within these images, and records their descriptions in a standardized XML file [9]. The development of a straightforward yet efficient auto-scale L2S (learn-to-scale) module presented the automatic rescaling of densely populated areas to establish practical proximity levels that correspond to distances between adjacent individuals in the image plane[10].

Counting people through face detection is a complex and unsolved challenge in the field of computer vision. This research explores the use of two object detection models for the purpose of detecting and tallying people's faces. The initial model is based on Faster-RCNN, and the second one relies on SSD, which are trained for object detection on a wide range of facial conditions. Both models resulted in an accuracy of 0.5, MAE obtained by F-RCNN was 7.5, and by SSD was 8.6 [11]. The system is developed for counting students through accurate face detection. It was designed to capture and recognize student faces effectively using Neural Networks-based object classification within digital images with 0.99 training accuracy and 0.98 testing accuracy. On RaspberryPi testbed, the first accurate label for facial recognition was found at a distance of 1 to 5 meters, except 4 meters [12].

Face detection encompasses a range of facial technologies, including authentication, recognition, and clustering. Effective preparation is essential for identification and comprehension. Unfortunately, conventional techniques have not yielded satisfactory results in terms of face recognition accuracy. To recognize faces within datasets, the model employs the CNN approach coupled with the well-established Max Pooling process to enhance the precision of face detection. The model is rigorously trained and validated using the LFW dataset. The training phase achieved an impressive accuracy rate of 95.72%, and during validation, the accuracy soared to 96.27%. [13]. The task of displaying the targeted object from the videos faces a detection accuracy challenge. An approach capable of classifying vehicles in videos was developed, which involved HOG features, a Linear SVM classifier and YOLOv3 algorithms and obtained improved training accuracy[14].

An investigation was carried out by contrasting three models—Faster-RCNN+ResNet50, SSD+Mobile Net, and EfficientDet+D0 to recognize various objects. Each model was trained on TensorFlow using a custom dataset. Experimental results revealed that in terms of mAP, Faster-RCNN+ResNet50 achieved the highest score of 0.45, while the lowest score was recorded for EfficientDet+D0 at 0.27 [15]. For precise detection in real-time over edge devices a method was designed that minimizes unnecessary detection entities and classifies essential objects of interest in application contexts. To achieve this, it employs a transfer learning technique that hones the model's learning exclusively on these selected objects. [16] The enhanced MTCNN was developed for the application environment of processing multi-face detection. This was achieved by applying the data augmentation, rebuilding the detection network with a deep separable convolutor, and optimising NMS. The rate of accuracy was 92% [17].

A sophisticated face-detection based attendance system was developed. It incorporates MT-CNN for face detection and utilizes the VGGFace2 large-scale face recognition dataset model. It can be effectively demonstrated in educational institutions, offering a solution to address issues related to fraudulent attendance and proxy attendance. It enables us to analyze the attendees at various events. Subsequently, the detected faces can be cross-referenced with the student face image database, ensuring that attendance records are effectively maintained[18].

In the context of people counting applications, two primary techniques are often considered: face detection and object detection. The choice between face detection and object detection for people counting depends on the nature of the environment and the specific requirements of the application, ensuring that the method selected aligns with the visibility and tracking challenges posed by the given scenario. In this paper, a comparative study of these techniques, analyzing their strengths, limitations, and real-world applicability is performed to contribute insights that can guide the selection of the most suitable detection method for specific use cases.

# Proposed Methodology

People counting using ML empowers businesses, institutions, and public agencies with valuable insights and tools for better decision-making, resource allocation, security, and operational efficiency. It is an essential technology for optimizing a wide range of industries and improving the overall quality of services provided to individuals in various settings. To implement these applications the two basic approaches employed are: face detection and object detection. The choice between these techniques depends on the specific requirements of the application and the environmental conditions. While face detection is well-suited for controlled environments with clear views of faces, object detection offers a broader and more flexible approach, making it suitable for scenarios with diverse challenges. Understanding the strengths and limitations of each technique is essential for effectively implementing people-counting solutions in various contexts.

The two state-of-the-art algorithms MTCNN, and YuNet based on face detection are employed as these techniques have demonstrated remarkable prowess in detecting faces under diverse conditions, setting the stage for comprehensive face recognition and analysis[17]. On the other hand, HOG-SVM and Yolov8 models are taken for examination under the object detection category. HOG-SVM excels in detecting objects based on gradient information, while YOLOv8, renowned for its speed and accuracy, brings real-time object detection capabilities to the forefront.

## People Counting using MTCNN

In the contemporary landscape of counting methods, crowd counting through CNN has emerged as the start-of-art approach. CNN is a sophisticated DL model composed of multiple layers of interconnected neural nodes. One of its primary strengths lies in its ability to extract crucial features from raw images and utilize these features to identify various objects and elements within those images. This technology has played a transformative role in the field of crowd counting, and this transformation became even more pronounced with the advent of MultiTask-Cascaded CNN, or MTCNN. It introduced a paradigm shift by enabling the network to simultaneously perform multiple tasks related to object recognition, localization, and classification within the same framework. This advancement allowed for more comprehensive and accurate crowd counting by incorporating a broader range of capabilities.

The MTCNN shown in Fig. 1, short for MultiTask Cascaded Convolutional Networks, utilizes a multi-stage architecture that consists of three distinct networks: the P-Net (Proposal Network), R-Net (Refine Network), and O-Net (Output Network). The P-Net is responsible for generating candidate facial regions, effectively reducing the area to be searched. Following this, the R-Net takes on the role of refining the proposals, thereby eliminating false positives. Ultimately, the O-Net specializes in fine-grained facial feature analysis, facilitating precise facial landmark localization and attribute recognition.

A diagram of a process flow

Description automatically generated

Fig. 1. MultiTask Cascaded CNN(MTCNN) Architecture

## People Counting using YuNet

YuNet, a lightweight and efficient face detector, achieves the optimal balance between accuracy and speed. Traditional face detection methods were successful when dealing with frontal faces, as they could reliably identify and analyze facial features in this orientation. However, they often faced difficulties when confronted with side-angle faces. These traditional detectors struggled to maintain the same level of accuracy and speed when dealing with profiles and varying head orientations, which are common scenarios in real-world applications. The introduction of YuNet in OpenCV's release (version 4.5.4, in October 2021) marked a notable milestone in addressing this challenge. YuNet has been specifically designed to overcome the limitations associated with side-angle face detection. It excels in accurately identifying and locating faces even when they are at different angles, providing a robust solution for real-world applications.

The YuNet architecture, as depicted in Fig. 2, consists of three main components: a backbone, a tiny feature pyramid network (TFPN) neck, and a head. The backbone serves as the central component of the network, responsible for feature extraction in the detection process. It's crucial for this part to be efficient and lightweight, particularly when deploying the model on edge devices. The 3×3 depthwise(DW) separable-conv. offers substantial gains in terms of computational efficiency and reduced parameters. DWBlock, which encompasses two DWUnits, further enhances the efficiency and effectiveness of the backbone. The neck component in the architecture is responsible for amalgamating multiscale features to create a more advanced level of feature representation. It leverages techniques such as FPN and depthwise(DW) separable conv to achieve this. In head section, simOTA is used as anchor matching to minimize losses.

A diagram of a network

Description automatically generated

## Fig. 2. YuNet Architecture

## People Counting using HOG-SVM

In the context of object detection, the Histogram of Oriented Gradients (HOG) technique is a valuable approach that leverages gradient-based features extracted from image blocks. These features are then organized into a feature vector that encapsulates essential information related to edges and object shapes. The HOG technique is distinguished by its ability to excel in the detection of objects under various conditions, including changes in scale, rotation, and lighting. This adaptability makes it particularly effective in scenarios where the appearance of objects can vary significantly. The strength of the HOG technique lies in its capacity to capture edge and shape information effectively. This involves analyzing the intensity gradients within localized regions of an image. The gradient direction and magnitude provide insights into the presence of edges and patterns, which are critical for identifying objects within images.

One of the key advantages of HOG is its ability to work across different scales, meaning it can detect objects whether they are large or small within an image. It is also robust against rotations, meaning objects can be oriented in various ways and still be accurately detected. Furthermore, the technique is resilient to changes in lighting conditions, allowing it to function well in environments with varying illumination. When used in conjunction with SVM as depicted in Fig.3, it provides a reliable framework for detecting objects. SVM serves as a classifier that, when coupled with the discriminative features extracted by HOG, can effectively distinguish between object and non-object regions in images.

A diagram of process flow

Description automatically generated

Fig. 3. HOG coupled with SVM Architecture

## People Counting using Yolo v8

Yolov8 (You Only Look Once Version8) represents a significant leap in the evolution of object detection methods, offering a remarkable balance between speed and accuracy. Its primary objective is to swiftly and accurately identify objects in real-time scenarios, making it a crucial advancement in computer vision and image analysis. Yolo employs a single pass, streamlining the process significantly. During this single pass, the network predicts both the classes of detected objects and their corresponding bounding box coordinates simultaneously. This innovative approach ensures that object detection occurs swiftly, and practically in real-time, while maintaining a high degree of precision and accuracy.

Yolo v8's architecture depicted in Fig.4 encompasses a CSPDarknet53 backbone responsible for feature extraction, utilizing cross-stage partial connections for enhanced gradient flow. The novel C2F mechanism enables effective feature fusion across scales, facilitating multi-level information integration. The head employs convolutions and spatial attention for feature refinement prior to bounding box prediction, involving class probabilities and coordinates, collectively enabling efficient object detection.

A diagram of a computer

Description automatically generated

## Fig. 4. Simplified Yolov8 Architecture

# Results Discussion and Comparative Analysis

This section presents an empirical evaluation conducted under a range of conditions, accompanied by a comparative analysis. This evaluation serves the purpose of offering valuable insights into the strengths and limitations of these approaches when applied to people counting tasks. It aids in guiding the selection of these methods for specific use cases, ensuring their suitability and effectiveness. It includes both results and a comparative analysis of methods within and across the two fundamental approaches, namely, face detection and object detection.

The prediction outcome is represented using a green colored box on the face. In assessing the algorithm performance, Ground Truth (GT) and Predicted Count (PC) serve as the foundational metrics. The results of MTCNN, YuNet, HOG-SVM, and Yolov8 are shown in Fig. 5a, 5b, 5c, and 5d respectively. Compared to MTCNN, YuNet performs better in some situations, notably when presented with photos taken from side angles. Even when face characteristics are not apparent, it excels at locating people inside a frame, ensuring nearly complete coverage of people in the scene. Compared to HOG-SVM, Yolov8 performs better in real-time images captured from running video, notably when people are partly overlapped.

|  |  |
| --- | --- |
| FaceDetection Algorithm Results | |
|  |  |
| a) MTCNN Results: GT- 5, PC- 4 | b) YuNet Results: GT- 5, PC- 5 |
| ObjectDetection Algorithm Results | |
|  |  |
| c)HOG-SVM Results:GT-10,PC-9 | d) Yolov8 Results: GT-10, PC-10 |
| Face Detection(FD) vs Object Detection(OD) | |
|  |  |
| e) FD Results: GT-6, PC-5 | f) OD Results: GT-6, PC-6 |
|  |  |
| g) FD Results: GT-1, PC-1 | h) OD Results: GT-1, PC-3 |
|  |  |
| i) FD Results: GT-3, PC-3 | j) OD Results: GT-3, PC-5 |

## Fig. 5. Performance Analysis of People Counting Algorithms

Face detection(FD) methods perform well when the scene captured contains objects like mannequins whereas Object detection(OD) performance degrades in this case whose results are presented from Fig 5g to Fig 5j. Conversely, Object Detection (OD) methods exhibit the capability to accurately identify individuals, even when their photographs are taken from side or rear perspectives whereas FD methods fail completely on the rear side, whose results are presented in Fig. 5e and Fig. 5f. People counting algorithms are analyzed in terms of Average Precision (AP) at easy level i.e., at AP50, which is detailed in Table 1. It is noted that the YuNet model exhibits superior average precision compared to other people counting algorithms.

1. Comparison of people counting methods

| Method | AP50 |
| --- | --- |
| MTCNN | 0.94 |
| YuNet | 0.98 |
| HOG-SVM | 0.91 |
| Yolov8 | 0.92 |
| Face Detect Net[19] | 0.80 |
| Multitask Point Supervision[20] | 0.75 |

# Conclusion

A practical evaluation of people counting algorithms was conducted under diverse conditions. In the context of face detection algorithms, YuNet demonstrated superior performance in side-angle photos compared to MTCNN. When dealing with real-time images extracted from video streams, YOLOv8 surpasses HOG-SVM, especially in situations with partial overlaps among people. Notably, face detection (FD) methods excel in scenes that include objects such as mannequins, while object detection (OD) faces challenges in such scenarios. Conversely, face detection (FD) algorithms struggle with rear-side photos, whereas object detection (OD) algorithms exhibit enhanced performance. Upon examining the experimental results critically, it is evident that, for people-counting tasks, object detection algorithms exhibit superior performance in both small and large gatherings, regardless of the viewpoint (front, rear, or side), with the exception of shopping areas. In these varied scenarios, YOLOv8 stands out as the superior model for people counting, while YuNet demonstrates its effectiveness as the preferred model in shopping areas.

##### References

1. A. Jalal, M. A. K. Quaid, and A. S. Hasan, “Wearable Sensor-Based Human Behavior Understanding and Recognition in Daily Life for Smart Environments, in proc. on FIT, 2018.
2. K. Kim, A. Jalal and M. Mahmood, “Vision-based Human Activity recognition system using depth silhouettes: A Smart home system for monitoring the residents, JEET, 2019.
3. Zeng J, Qiu X, Shi S,“Image processing effects on the deep face recognition system”, Math. Biosci. and Engg., vol. 18, no. 2, pp. 1187–1200, 2021.
4. Liu J and Ashraf MA,“Face recognition method based on GA-BP neural network algorithm,” Open Phys., vol. 16, no. 1, pp.1056 – 1065, 2018.
5. Wang Q, Guo G,“Benchmarking deep learning techniques for face recognition”, Journal of VisualComm. and Image Representation, vol.65,pp.102663,2019.
6. Vasantha SV, Samreen S, Aparna YL. Rice Disease Diagnosis System (RDDS). Computers, Materials & Continua. 2022 Oct 1;73(1)
7. Mao QC, Sun HM, Liu YB, Jia RS. Mini-YOLOv3: real-time object detector for embedded applications. Ieee Access. 2019 Sep 16;7:133529-38.
8. M.Bodke, C. Patil, P. Chopade, Y. Patil and O. Patil, "A Review Paper on Object-Detection using the DeepLearning Approach," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10151610.
9. Bornia J, Frihida A and Claramunt C, "Detecting objects and people and tracking movements in a video using tensorflow and deeplearning," 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC\_ASET), Hammamet, Tunisia, 2020, pp. 213-218, doi: 10.1109/IC\_ASET49463.2020.9318253.
10. Xu C, Liang D, Xu Y, Bai S, Zhan W, Bai X, Tomizuka M. Autoscale: learning to scale for crowd counting. International Journal of Computer Vision. 2022 Feb;130(2):405-34.
11. Y. al Atrash, M. Saad and I. H. Alshami, "Detecting and Counting People's Faces in Images Using Convolutional Neural Networks," 2021 Palestinian International Conference on Information and Communication Technology (PICICT), Gaza, Palestine, State of, 2021, pp. 116-122, doi: 10.1109/PICICT53635.2021.00031.
12. M.Irsan, R. Hassan, M. K. Hasan and M. C. Lam, "The Process of Using Face Detection Through Convolutional Neural Network," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICBATS54253.2022.9759092.
13. F. M. J. Mehedi Shamrat, M. A. Jubair, M. M. Billah, S. Chakraborty, M. Alauddin and R. Ranjan, "A Deep Learning Approach for Face Detection using Max Pooling," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 760-764, doi: 10.1109/ICOEI51242.2021.9452896.
14. A. S. Abdullahi Madey, A. Yahyaoui and J. Rasheed, "Object Detection in Video by Detecting Vehicles Using Machine Learning and Deep Learning Approaches," 2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era (FoNeS-AIoT), Nicosia, Turkey, 2021, pp. 62-65, doi: 10.1109/FoNeS-AIoT54873.2021.00023.
15. C.Hary and S. Mandala, "Object Detection Analysis Study in Images based on Deep Learning Algorithm," 2022 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2022, pp. 226-231, doi: 10.1109/ICoDSA55874.2022.9862922.
16. D. Kim, S. Lee, N. -M. Sung and C. Choe, "Real-time object detection using a domain-based transfer learning method for resource-constrained edge devices," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 457-462, doi: 10.1109/ICAIIC57133.2023.10067064.
17. Q. Guo, Z. Wang, C. Wang and D. Cui, "Multi-face detection algorithm suitable for video surveillance," 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 2020, pp. 27-33, doi: 10.1109/CVIDL51233.2020.00013.
18. V.Bittal, V. Jagdale, A. Brahme, D. Deore and B. Shinde, "Multifarious Face Attendance System using Machine Learning and Deep Learning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 387-392, doi: 10.1109/ICICCS56967.2023.10142759.
19. Gorbatsevich VS, Vizilter YV. FaceDetectNet: Face detection via fully-convolutional network. Компьютерная оптика. 2019;43(1):63-71.
20. M. Zand, H. Damirchi, A. Farley, M. Molahasani, M. Greenspan and A. Etemad, "Multiscale Crowd Counting and Localization By Multitask Point Supervision," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 1820-1824, doi: 10.1109/ICASSP43922.2022.9747776.