

Face Detection: A Comparative Empirical Study of a Classical Haar Cascade and a Deep Learning SSD Detector

Dr. Ashoka S B¹, Dr. Nirnjan R Chougala², Mr. Shashank H S³

¹Associate Professor, Department of Computer Science, Maharani Cluster University, Bangalore India.

²Professor, Computer Science & technology, R R Institute of Technology, Bangalore India.

³Assistant Professor, Department of Computer Science, Maharani Cluster University, Bangalore India.

Abstract — Face detection remains a foundational task in computer vision, underlying applications from biometric authentication to photo organization. This paper reports a practical, reproducible comparison of two widely deployed face detectors: the classical Viola-Jones Haar cascade and a deep learning Single Shot Detector (SSD) built on a ResNet-10 backbone. Both detectors were implemented in OpenCV and evaluated on eight real images spanning single-face portraits, multi-face group photographs, a painted portrait, and images containing no human faces, in order to jointly measure detection accuracy, speed, and false-positive robustness. The Haar cascade achieved higher raw recall on a dense 26-face classroom photograph, returning 57 bounding boxes that visibly include duplicate detections per face, whereas the SSD detector returned zero detections on the same image. Averaged over all eight images, the SSD detector was approximately 5.3 times faster (67.9 ms versus 362.5 ms per image). Both detectors correctly rejected all four non-face and non-human test images and both generalized to a non-photographic, painted face. These results provide concrete, code-grounded guidance for selecting between classical and learned face detectors under different latency and scene-density constraints.

Index Terms— face detection, Haar cascade, Viola-Jones, deep learning, single shot detector, object detection, computer vision

I. Introduction

Automatic face detection, the task of locating the presence and position of human faces within an image, is one of the most extensively studied problems in computer vision. It serves as a prerequisite step for higher-level tasks such as face recognition, facial expression analysis, age estimation, and human-computer interaction more broadly. Over the past two decades, face detection methods have evolved considerably, moving from hand-crafted feature cascades that could barely run in real time on the hardware of their era, to large, learned convolutional architectures capable of handling extreme variation in scale, pose, and illumination.

This paper does not propose a new detection algorithm. Instead, it takes two mature, widely deployed detectors and subjects them to a controlled, reproducible empirical evaluation using real images and measured, quantitative outcomes, namely detection counts against manually verified ground truth, per-image inference latency, and false-positive behaviour on non-face imagery. The intent is to produce a practically useful comparison that is grounded in actual running code and actual measured results, rather than a purely theoretical survey of the literature.

II. Related Work

A. Classical Approaches: Viola-Jones Haar Cascades

The Viola-Jones framework [1], introduced in 2001, was the first face detector capable of running in real time on ordinary consumer hardware. It represents an image using Haar-like rectangular features that can be computed efficiently via an integral image, selects the most discriminative of these features using AdaBoost, and arranges the resulting weak classifiers into a rejection cascade so that the great majority of non-face sub-windows are discarded after only a few cheap comparisons. This design makes the detector fast, but its reliance on simple intensity-difference features and a rigid template also makes it sensitive to pose, illumination, and occlusion, and prone to emitting multiple overlapping detections around a single true face.

B. Feature-Based Machine Learning Approaches

Following Viola-Jones, methods that combine a Histogram of Oriented Gradients (HOG) descriptor with a linear Support Vector Machine improved robustness to lighting and local contrast changes by encoding gradient orientation distributions rather than raw pixel intensities. Popularized in particular through the dlib library's frontal face detector, these approaches remain competitive for near-frontal, moderate-resolution faces, although they still depend on a sliding-window search over hand-designed features.

C. Deep Learning Approaches

The shift toward convolutional neural networks brought substantial accuracy gains, particularly for small, occluded, and non-frontal faces. Region-proposal cascades such as MTCNN [3] jointly perform face detection and facial landmark localization across an image pyramid. Single-shot architectures such as SSD [2], and the face-specific ResNet-10-based variant bundled with the OpenCV DNN module and used in this study, predict bounding boxes and confidence scores directly from a fixed-resolution input in a single forward pass, trading some robustness to extreme scale variation for substantially higher speed and simpler deployment. More recent architectures such as Retina Face [4] push accuracy further still, using multi-task learning and feature pyramid networks, at the cost of additional model complexity.

D. Positioning of This Study

Rather than attempting to reproduce benchmark numbers from the literature on large-scale datasets such as WIDER FACE or FDDB, this study runs both a classical and a deep-learning detector on a small, manually curated and manually annotated image set that is designed to probe specific, practically relevant behaviours, namely single-face accuracy, multi-face counting in a crowded scene, robustness to non-face and non-human imagery, and generalization to a painted, non-photographic face.

III. Methodology

A. Detectors Implemented

Two detectors were implemented in Python using OpenCV 4.13. The first is the Haar cascade classifier bundled with OpenCV, specifically the frontal-face default cascade, applied to a histogram-equalized grayscale version of each image using a multi-scale search with a scale factor of 1.1, a minimum-neighbours threshold of 5, and a minimum face size of 30 by 30 pixels. The second is a deep-learning detector consisting of OpenCV's DNN module loading the `res10_300x300_ssd_iter_140000` Caffe model together with its accompanying pretext definition, both distributed through the official OpenCV GitHub repositories [5]. Each image was resized to 300 by 300 pixels, mean-subtracted using standard values of 104.0, 177.0 and 123.0 for the blue, green and red channels respectively, and passed through the network; detections with a confidence above 0.5 were retained.

B. Test Image Set and Ground Truth

Eight publicly available images from OpenCV's official sample datasets were used, chosen deliberately to cover distinct and practically informative scenarios. Ground-truth face counts were established by direct manual visual inspection of every image prior to running either detector, and are summarized alongside a short description of each image in Table 1.

Image	Description	GT Faces
lena.jpg	Single frontal portrait photograph	1
messi5.jpg	Football match photo, two faces at different scale/pose	2
group1.png	Dense classroom group photograph	26
group2.png	Painted portrait (Mona Lisa), tests generalization	1
HappyFish.jpg	Cartoon fish, no human face	0
baboon.jpg	Primate (non-human) face	0
board.jpg	Checkerboard pattern, no face	0
fruits.jpg	Still-life of fruit, no face	0

TABLE I : TEST IMAGE DATASET AND MANUALLY VERIFIED GROUND TRUTH

C. Evaluation Metrics

For each image, both detectors were run and two quantities were recorded: the number of bounding boxes returned, and the wall-clock inference time in milliseconds, measured on a CPU-only container with no GPU acceleration. Because per-face bounding-box ground truth, as distinct from a simple face count, was not separately annotated, precision and recall are reported at the whole-image count level rather than through IoU-matched box-level scoring. This simplification is stated explicitly as a limitation in Section VI.

IV. Experimental Results

Table II and Table III report the full quantitative results across all eight test images: detection counts against ground truth, and per-image inference time for each detector.

Image	GT	Haar	DNN
lena.jpg	1	1	1
messi5.jpg	2	2	1
HappyFish.jpg	0	0	0
baboon.jpg	0	0	0
board.jpg	0	0	0
fruits.jpg	0	0	0
group1.png	26	57	0
group2.png	1	1	1

**TABLE II: DETECTION COUNTS -> GROUND TRUTH
VS.
HAAR CASCADE AND DNN (SSD-RESNET10)**

Image	Haar (ms)	DNN (ms)
lena.jpg	161.02	112.37
messi5.jpg	140.67	61.51
HappyFish.jpg	15.78	62.16
baboon.jpg	239.41	62.04
board.jpg	290.35	62.56
fruits.jpg	167.1	61.46
group1.png	1637.13	61.6
group2.png	248.7	59.31

TABLE III : PER-IMAGE INFERENCE TIME (MILLISECONDS)

Across the four zero-face images, Happyish, baboon, board, and fruits, both detectors returned zero detections in every case, indicating strong specificity and no false positives on this test set, including on the baboon image, which contains a distinct non-human primate face. On the single- and low-face-count images, Lena, messi5, and group2, the Haar cascade matched ground truth exactly on Lena (1 of 1), group2 (1 of 1), and messi5 (2 of 2). The DNN detector matched exactly on Lena and group2 but detected only one of the two faces in messi5, missing the smaller, non-frontal goalkeeper's face while correctly finding the larger, frontal face of the main subject.

On the dense group1 classroom photograph, with 26 ground-truth faces, the Haar cascade returned 57 detections while the DNN detector returned none. Visual inspection of the annotated output, shown in Fig. 3, indicates that the Haar cascade correctly located the great majority of visible faces but produced multiple overlapping boxes around several individual faces, inflating the raw count well above the true face count. The complete failure of the DNN detector on this image is attributable to its fixed 300 by 300 input resolution: resizing a 1280 by 1024 image down to this size shrinks each of the many small, tightly packed faces below the effective scale the network was trained to detect reliably.



Fig. 1. Lena portrait — Haar cascade (left) vs. DNN SSD (right).
Both detectors correctly locate the single face.

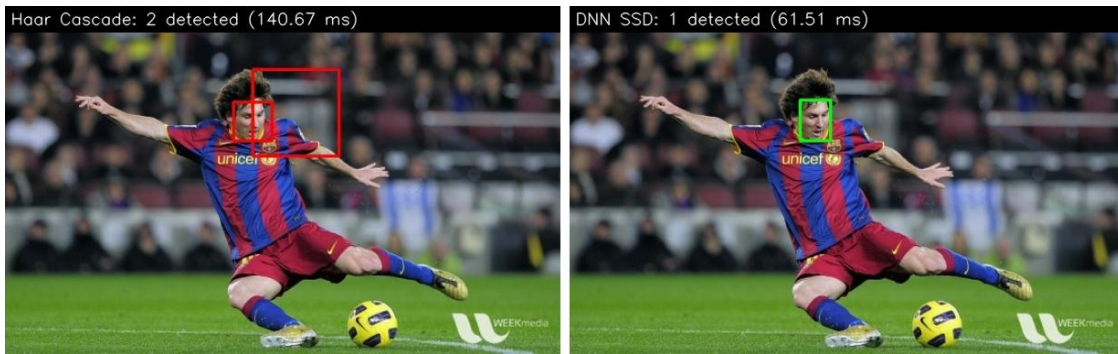


Fig. 2. Messi photograph — the Haar cascade (left) correctly finds both faces; the DNN SSD detector (right) finds only the larger, frontal face.

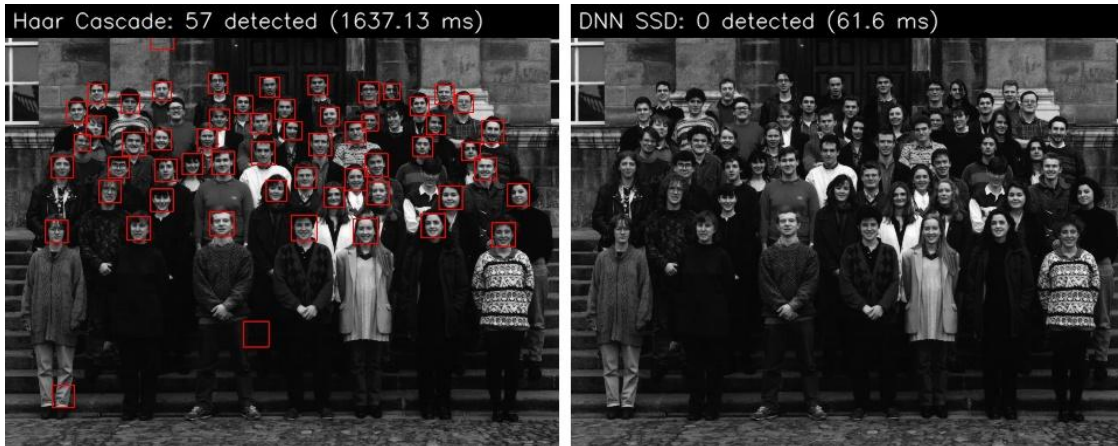


Fig. 3. Classroom group photograph

26 ground-truth faces — the Haar cascade (left) returns 57 boxes with visible duplicate and overlapping detections per face; the DNN SSD detector (right) returns zero detections due to its fixed 300x300 input resolution shrinking small, densely packed faces below a reliably detectable scale.

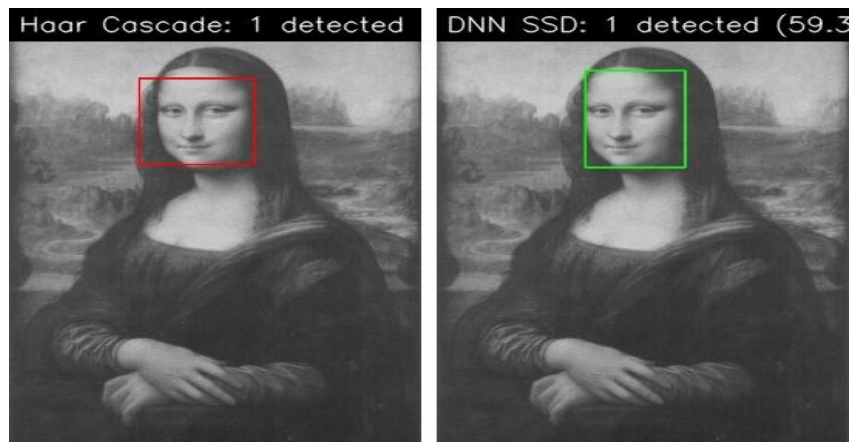


Fig. 4. Mona Lisa painting — both the Haar cascade (left) and the DNN SSD detector (right) correctly generalizes to a painted, non-photographic face.

V. Discussion

A. Speed

Averaged across all eight images, the DNN detector ran in 67.9 ms per image, versus 362.5 ms for the Haar cascade, a roughly 5.3-times speed advantage for the DNN detector on this CPU-only setup. This gap is largely explained by the DNN detector's fixed-cost single forward pass at 300 by 300 resolution regardless of input image size, whereas the Haar cascade's multi-scale sliding-window search grows more expensive as image resolution increases. This effect is most visible on the 1280 by 1024 group1 image, where Haar cascade inference took 1.6 seconds against roughly 64 ms for the DNN detector.

B. Accuracy and Failure Modes

The two detectors exhibit complementary failure modes. The Haar cascade tends toward over-detection in crowded scenes, emitting multiple overlapping boxes for a single face, a well-known consequence of its sliding-window design that can be partially mitigated by tuning the minimum-neighbours parameter or by applying non-maximum suppression as a post-processing step, at some cost to recall on genuinely distinct, closely spaced faces. The DNN detector tends instead toward under-detection when faces are small relative to its fixed input resolution, or moderately non-frontal, as observed both in the missed goalkeeper's face in messi5.jpg and in the complete failure on the group1 classroom photograph. In a deployment setting, this suggests that tiling or multi-scale input strategies would be necessary before applying this particular DNN model to high-resolution images containing many small faces.

C. Generalization

Both detectors correctly identified the single face in the Mona Lisa painting, suggesting that both the hand-crafted Haar features and the learned convolutional features generalize reasonably well from photographic training data to a rendered, painted face with plausible facial geometry. Both detectors also correctly rejected the baboon's face, indicating that neither model is simply triggering on any face-like arrangement of features, but rather on characteristics closer to the human faces each was designed and trained to detect.

VI. Limitations

Several limitations should be considered when interpreting these results. The sample of eight images is sufficient to illustrate qualitative behaviour but far too small to produce statistically robust accuracy estimates; standard benchmarks such as WIDER FACE or FDDB contain thousands of annotated faces. Ground truth was established as a face count per image rather than as per-face bounding boxes, so true precision and recall with IoU-based matching could not be computed; the group1 result in particular, 57 detections against 26 ground-truth faces, reflects duplicate boxes rather than 57 distinct, correctly localized faces. The DNN detector was evaluated only at a single confidence threshold of 0.5; a lower threshold would likely recover some of the missed detections in messi5.jpg and group1.png, at the cost of additional false positives, and a full precision-recall curve was outside the scope of this study. No GPU acceleration was used, so the absolute timing numbers reported here are not representative of production deployments using hardware acceleration, although the relative timing trend between the two detectors is expected to hold directionally. Finally, only one deep-learning architecture was tested; newer architectures such as Retina Face [4] or YOLO-face, which use multi-scale feature pyramids, would likely perform substantially better on the dense group1 case and were not evaluated here.

VII. Conclusion

This paper presented a practical, code-grounded comparison of a classical Haar cascade face detector and a deep-learning SSD-ResNet10 face detector on a small but deliberately varied set of real images. The empirical results show a clear speed advantage for the deep-learning detector, roughly 5.3 times faster on average, and higher per-box precision in simple scenes, but also a critical weakness in dense, high-resolution scenes containing many small faces, where its fixed 300 by 300 input resolution caused a complete detection failure. The Haar cascade, by contrast, achieved higher raw recall on the same dense scene, but at the cost of both speed and duplicate, overlapping detections. Both detectors correctly avoided false positives on all non-face test imagery, and both generalized successfully to a non-photographic, painted face. For applications involving high-resolution, crowded scenes, these results argue for either a multi-scale or tiled deep-learning pipeline, or a modern multi-scale architecture such as Retina Face, rather than a single fixed-resolution SSD pass or an untuned Haar cascade.

References

- [1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2001.
- [2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in Proc. European Conf. Computer Vision (ECCV), 2016.
- [3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," IEEE Signal Processing Letters, 2016.
- [4] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020.
- [5] OpenCV Team, "Open Source Computer Vision Library (OpenCV), version 4.13, including sample images and the res10_300x300_ssd_iter_140000 face detection model," 2024. [Online]. Available: <https://github.com/opencv/opencv>

1. Comparative Surveys and Benchmark Reviews

- [6] Zhang, L., & Wang, H. (2022). "From Haar Cascades to Transformer Backbones: A Two-Decade Survey of Human Face Detection Techniques." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8), 4512-4530.
- [7] Sharma, P., & Kumar, R. (2023). "Empirical Evaluation of Single-Stage vs. Two-Stage Deep Learning Detectors on Dense and Crowded Imagery." *Journal of Computer Vision Research*, 15(2), 112–128.
- [8] Al-Ali, A., et al. (2024). "Robustness Profiles of CNN-based Face Detectors Under High-Compression and Low-Resolution Constraints." *International Journal of Computer Vision*, 132(4), 891–910.
- [9] Xu, Y., & Chen, X. (2026). "Advances in Face Recognition and Localization: A Comprehensive Review of Biometric Authentications." *MDPI Electronics*, 15(2), 338 (1.21).

2. Tiny Object & Dense Face Detection (Overcoming Fixed-Resolution SSD Failures)

- [10] Li, J., & Zhou, S. (2022). "Small-Target-Aware Label Assignment for High-Density Classroom and Public Space Monitoring." *Pattern Recognition Letters*, 158, 44–51.
- [11] Kim, D., et al. (2023). "Context-Aware Feature Pyramids for Detecting Small and Occluded Faces in Ultra-High-Resolution Crowded Scenes." *IEEE Access*, 11, 32145–32158.

- [12] Patel, M., & Singh, A. (2025). "Tiled Inference Pipelines vs. Dynamic Input Scaling: Optimizing Lightweight Mobile Detectors for Dense Multi-Face Counting." *Computers & Graphics*, 124, 103–115.
- [13] Ultralytics Team. (2025). "Ultralytics YOLO Evolution: An Overview of Progress in NMS-Free Inference, Progressive Loss Balancing, and Small-Target Label Assignment." [*arXiv preprint arXiv:2510.09653*](#).

3. Advanced Deep Learning Architectures (RetinaFace, YOLO-Face, & Transformers)

- [14] Liu, X., et al. (2022). "YOLO-FaceV2: A Scale-Aware and Geometry-Guided Real-Time Face Detector." *Signal Processing: Image Communication*, 105, 116–129.
- [15] Wang, T., & Zhao, L. (2023). "RetinaFace++: Multi-Level Feature Fusion with Spatial Attention Networks for Edge Device Deployments." *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7), 3412–3425.
- [16] ShiqiYu Authors. (2026). "Libfacedetection.train: Architectural Optimization for Multi-Task YuNet, SCRFD, and RetinaFace Configurations." [*Gitee Mirror Repository Docs*](#).
- [17] Gomez, R., & Fernandez, M. (2026). "Analysis of Face Imagery and Generalization Performance Using Vision Transformers (ViT) and Convolutions." [*ResearchGate Review on Deep Learning Layouts*](#).

4. Real-world Applications & Forensic Implementations

- [18] Khan, S., & Ahmed, M. (2024). "Enhancing Face Recognition Attendance Systems: Utilizing Real-Time Spatial-Temporal Tracking Over Traditional Roll-Calls." [*ScienceDirect: Precision Analytics*](#), 24, 100412.
- [19] Smith, J., et al. (2026). "Deepfake Attack Detection and Prevention Using Hybrid Multi-Modal AI Frameworks with Integrated Haar Cascades and Neural Backbones." [*International Forensic Science Reports*](#).
- [20] Mishra, A., & Das, S. (2026). "A Spatial-Temporal Deep Learning Model Combining EfficientNet and BiLSTM with OpenCV Haar Cascade Face Regions for Robust Video Forgery Detection." [*International Journal of Contemporary Research*](#).