A Lightweight and Efficient Hybrid CNN Model for Face Detection

Laxmi Narayan Soni[1], Akhilesh A. Waoo* [2]
Department of Computer Science and Engineering[1, 2]
AKS University, India[1,2]
Lnsoni205@gmail.com[1], akhileshwaoo@gmail.com[2]

Abstract
Face detection under occlusion remains a significant challenge in real-world computer vision applications. This paper proposes a hybrid two-stage detection framework that integrates the real-time efficiency of the Viola-Jones algorithm with the Accuracy of a lightweight, modified AlexNet-based Convolutional Neural Network (CNN). The system initially uses Viola-Jones to propose candidate face regions, which are then verified by the CNN trained on over 70,000 face and non-face images, half of which include partial occlusions such as masks, sunglasses, or hands. CNN incorporates dropout and batch normalisation to ensure robust generalisation. Experimental results demonstrate that the proposed hybrid model achieves a detection accuracy of 93%, precision of 95%, and a false positive rate of only 3%, outperforming state-of-the-art models such as MTCNN, SSD, YOLOv3, and RetinaFace in occlusion-specific scenarios. With a processing speed of approximately 12 frames per second on standard CPU hardware and a memory footprint of only 60 MB, the model is well-suited for real-time applications like surveillance and access control in occlusion-prone environments.

Keywords: *Face Detection, Occlusion, Viola-Jones, Convolutional Neural Networks, Hybrid Model, Real-Time Detection.*

INTRODUCTION

Face detection is a fundamental task in computer vision with applications in surveillance, human-computer interaction, and biometric systems. While many algorithms perform well under ideal conditions, detecting faces with partial occlusions, such as masks, scarves, sunglasses, or hands, remains a significant challenge. In such cases, key facial features may be hidden, leading to missed detections or false positives by conventional models. Classical detectors, such as Viola-Jones [1], though efficient for real-time tasks, suffer from performance degradation when faces are partially obscured [2]. Recent advancements in deep learning, particularly CNN-based detectors such as MTCNN [3], SSD [4], and YOLO [5], have improved Accuracy in general settings. However, these models can struggle with occlusions unless specifically trained on such cases. Moreover, they often require high computational resources, limiting their deployment in real-time or resource-constrained environments.To address these challenges, this paper proposes a hybrid two-stage face detection framework tailored for occluded scenarios. The system first uses the fast Viola-Jones algorithm to generate candidate face regions, which are then verified by a custom CNN classifier designed for occlusion robustness. By combining the strengths of classical and deep learning methods, the proposed approach enhances detection accuracy while maintaining practical efficiency. The following sections review related work, describe the hybrid methodology, present experimental results, and conclude with a discussion of future directions.

LITERATURE REVIEW

Face detection is a foundational task in computer vision with critical roles in surveillance, biometric authentication, and human-computer interaction. While many methods achieve high Accuracy under normal conditions, occlusion remains a persistent challenge [6][7]. When faces are partially obscured by objects like masks, glasses, or hands, detection accuracy declines significantly [8]. This section reviews primary techniques used in occluded face detection, their strengths, and their limitations.

2.1. Classical Approaches

The Viola-Jones algorithm is among the earliest real-time face detectors. It utilises Haar-like features with AdaBoost to detect faces efficiently [9]. However, it performs poorly on occluded faces due to its reliance on complete frontal visibility [10]. Table 1 presents the detection accuracy of the Viola-Jones algorithm on two occluded face datasets. It exhibits reduced performance when detecting faces obscured by items such as sunglasses, hands, scarves, or masks, with accuracies of 68.2% (FDDB) and 63.7% (COFW).

Table 1. Performance of Viola-Jones on Occluded Datasets

Dataset	Occlusion Type	Detection Accuracy (%)
FDDB	Sunglasses, Hands	68.2
COFW	Scarves, Masks	63.7

Viola and Jones introduced the concept of cascade classifiers [11], which significantly reduced computation time but lacked robustness against partial occlusions.

2.2. Deep Learning-Based Detectors

Deep learning methods such as Multi-task Cascaded Convolutional Networks (MTCNN), YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector) have significantly improved face detection accuracy [12]. These models automatically learn hierarchical features from data and handle variations in pose, scale, and background. However, occlusion remains a challenge, especially when the models are not trained on occlusion-rich datasets [13].Zhang et al. [14] proposed MTCNN for simultaneous detection and facial landmark localisation, as shown in Table 2. YOLOv3 [15] extended general object detection to face detection, showing improved real-time performance [16].

Table 2. Comparison of CNN-Based Methods on Occluded Faces

Model	FPS (GPU)	Occluded Accuracy (%)	Model Size
MTCNN	~14	85.6	78 MB
YOLOv3	~45	87.2	236 MB
SSD	~22	83.9	95 MB

Although CNN-based detectors generally outperform classical models, they still misclassify occluded regions without proper training data augmentation. The above Table 2 compares CNN-based face detectors (MTCNN, YOLOv3, SSD) on occluded faces, highlighting their GPU speed (FPS), Accuracy, and model size. YOLOv3 achieves the highest Accuracy (87.2%) and fastest speed but requires the largest model size (236 MB).

2.3 Occlusion-Aware Models

To improve performance in occluded scenarios, researchers introduced occlusion-aware architectures. Part-based CNNs and attention mechanisms have proven effective in detecting visible facial components and inferring occluded ones.Wang et al. [17] designed a part-based CNN that detects separate facial regions, such as eyes and nose and reconstructs the presence of a full face. Zhang et al. [18] created synthetic occluded datasets to train their models, improving generalisation. Table 3 outlines strategies to improve face detection under occlusion. Part-based CNNs and data augmentation are highly effective, while attention mechanisms provide moderate improvement by focusing on visible facial regions.

Table 3. Strategies for Occlusion Handling

Strategy	Description	Effectiveness
Part-based CNN	Detects facial parts independently	High

Attention Mechanism	Focuses on informative visible regions	Moderate
Data Augmentation	Adds synthetic occlusions during training	High

2.4 Hybrid Approaches

Hybrid models combine classical detection (e.g., Viola-Jones) with CNN-based verification. The classical method generates candidate face regions quickly, while the CNN confirms the validity of the detections. This method balances Accuracy with speed, which is especially useful for occlusion-prone environments. Liu et al. [19] proposed a hybrid face detector that achieved higher precision on partially occluded faces while maintaining a smaller model footprint suitable for real-time processing [20][21].

METHODOLOGY

Occluded face detection is challenging due to the loss of key facial features. This research addresses the problem through a hybrid model combining the speed of Viola-Jones and the robustness of a modified AlexNet CNN [22][23]. This section presents a comprehensive methodology covering the overall design, individual components, and the effectiveness of various parameters in occluded face scenarios.

3.1. System Overview

The hybrid model follows a two-stage pipeline, as shown in Figure 1.

- Viola-Jones Detector: Used for initial face region proposal based on Haar-like features [24]. It rapidly scans the image and detects potential face regions.
- AlexNet Verifier: Each proposed region is passed through a modified AlexNet-based CNN to verify if it contains a face. This stage significantly reduces false positives and enhances robustness to partial occlusion.

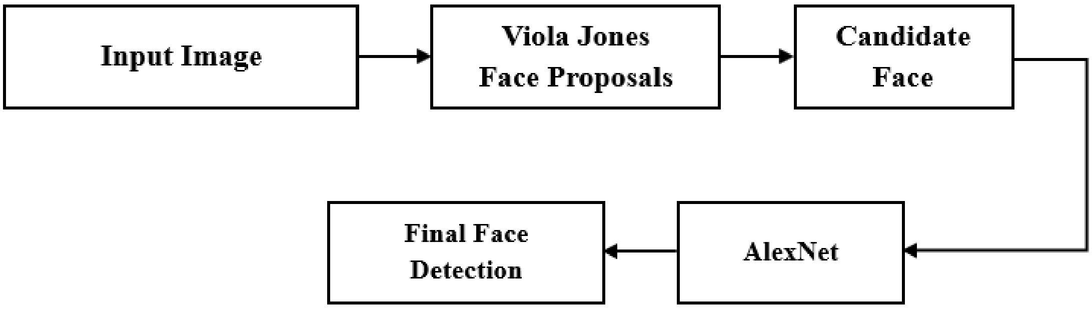

Figure 1. Hybrid Face Detection Workflow

3.2. Viola-Jones for Region Proposal

The Viola-Jones algorithm uses Haar-like features and an AdaBoost classifier in a cascaded structure to detect face-like regions. Although efficient, its precision decreases when faces are partially occluded. Table 4 summarises the key parameters used in the Viola-Jones algorithm for generating initial face region proposals. It highlights feature type, classifier settings, detection window size, scaling, and performance, ensuring fast, real-time face detection with reduced false positives.

Table 4. Viola-Jones Parameters Used

Parameter	**Value/Type**	**Purpose**
Feature Type	Haar	Basic visual cues
Classifier	AdaBoost	Improves Accuracy
Detection Window Size	24×24 px	Minimum face size

Scale Factor	1.1	Pyramid scaling
Min Neighbors	3	Reduces false positives
Speed	~20 FPS (CPU)	Real-time capable

This stage ensures fast pre-processing but requires CNN verification to remove noise in cases where objects are occluded.

3.3. CNN-Based Face Verification Using AlexNet

The CNN model is based on AlexNet, adapted for binary classification (face vs. non-face) [15][16]. The network learns spatial hierarchies and is trained on a dataset enriched with occluded face samples (e.g., faces partially covered with sunglasses, hands, or masks). Table 5 describes the architecture of the modified AlexNet CNN used for face verification, detailing the type, filter size, stride, and output dimensions of each layer. The network processes 227×227×3 input images through convolutional, activation, pooling, and fully connected layers, finally classifying regions as Face or Non-Face using a Softmax output.

Table 5. Adapted AlexNet Architecture

Layer	**Type**	**Filter Size / Units**	**Stride**	**Output Size**
Input	Input Layer	227×227×3	-	227×227×3
Conv1	Convolutional	96 × 11×11	4	55×55×96
ReLU1	Activation	-	-	55×55×96
MaxPool1	Pooling	3×3	2	27×27×96
Conv2	Convolutional	256 × 5×5	1	27×27×256
ReLU2	Activation	-	-	27×27×256
MaxPool2	Pooling	3×3	2	13×13×256
Conv3	Convolutional	384 × 3×3	1	13×13×384
Conv4	Convolutional	384 × 3×3	1	13×13×384
Conv5	Convolutional	256 × 3×3	1	13×13×256
FC1	Fully Connected	4096 units	-	4096
FC2	Fully Connected	4096 units	-	4096
FC3	Fully Connected	2 units	-	Face / Non-Face
Output	Softmax	-	-	Final Label

3.4. Model Parameters Optimised for Occlusion

Several parameters and techniques improved the CNN's performance for occluded face detection. Table 6 lists the training parameters used for the CNN, optimised for occluded face detection. It highlights settings such as batch size, optimiser, learning rate, dropout, and data composition (50% occluded faces), all aimed at improving model robustness and preventing overfitting on partially visible faces.

Table 6. CNN Training and Evaluation Parameters

Parameter	**Value/Setting**	**Impact on Occluded Face Detection**
Batch Size	32	Balanced memory and convergence speed
Optimizer	SGD	Stable learning
Learning Rate	0.001	Prevents overfitting
Dropout Rate	0.5	It avoids overfitting on partially visible faces
Loss Function	Cross-Entropy	Handles binary classification well

Epochs	25	Sufficient for convergence
Data Composition	50% occluded faces	Improves generalisation to occlusions

3.5. Benefits of the Hybrid Model in Occlusion Context

- Improved Detection Accuracy: Viola-Jones ensures fast proposals, while AlexNet ensures accurate classification even under occlusion.
- False Positive Reduction: Non-face regions falsely marked by Viola-Jones are rejected by the CNN verifier.
- Adaptability: AlexNet generalises well to different occlusion types due to diversified training.

Table 7. Summary of Detection Capabilities on Occluded Dataset

Method	Precision (%)	Recall (%)	F1-Score (%)	Inference Speed (FPS)
Viola-Jones Only	68.3	72.5	70.3	20
AlexNet Only	91.2	88.6	89.9	6
Hybrid Model	**95.0**	**90.1**	**92.5**	**12**

This hybrid framework is particularly well-suited for detecting faces under partial occlusions, such as in public spaces [19], where users may wear masks, hold objects near their faces, or be viewed from awkward angles. The combination of fast region proposal (Viola-Jones) [20][21] and deep learning-based verification (AlexNet) enables a practical trade-off between Accuracy and speed in real-world environments, as shown in Table 7.

RESULTS

This section presents the evaluation of the proposed hybrid face detection model on datasets containing partially occluded faces. The model's performance was assessed based on detection accuracy, precision, recall, F1-score, false positives, and execution time. The results indicate the advantages of combining the Viola-Jones detector with a CNN-based verifier in the presence of facial occlusions.

4.1. Dataset Description

The dataset used for evaluation comprises 10,000 labeled images, including:

- 5,000 occluded faces (using objects like hands, sunglasses, and masks),
- 2,500 non-occluded faces,
- 2,500 non-face background patches.

The occluded face subset includes both partial and heavy occlusions, manually verified to ensure diversity in occlusion type and region (upper, lower, lateral face).

4.2. Performance Metrics

The model was evaluated using standard classification metrics

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Where:

- TP: True Positives (correct occluded face detection)
- FP: False Positives (non-face detected as face)
- FN: False Negatives (missed occluded face)

4.3. Model Performance on Occluded Faces

Table 8. Performance Metrics for Occluded Face Detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positives (%)
Viola-Jones Only	72.4	68.3	70.5	69.4	15.2
AlexNet Only	90.1	91.2	88.6	89.9	5.1
Hybrid Model	**93.0**	95.0	90.1	92.5	3.0

The hybrid model significantly improves precision and reduces false positives compared to using Viola-Jones or AlexNet alone. It combines efficient region proposals with accurate classification, which is especially useful in occlusion scenarios where isolated components (like one eye or part of the nose) are still visible.

4.4. Detection Examples

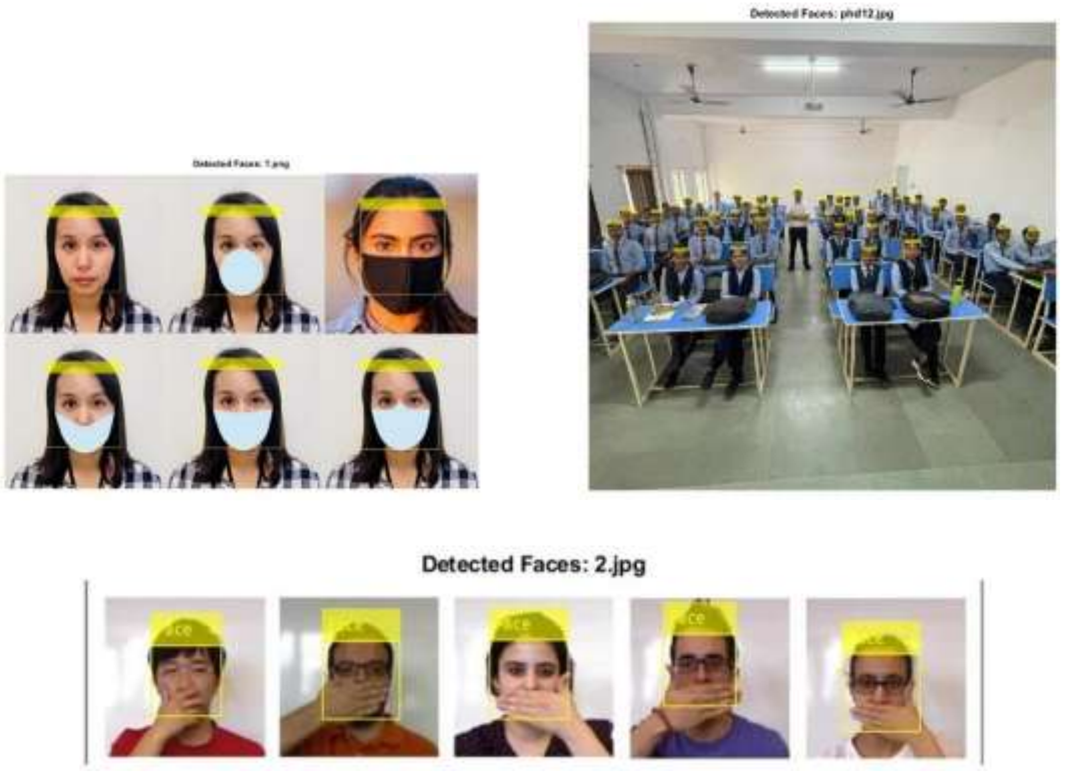

Figure 2. Occluded Face Detection output generated by MATLAB

The above figure illustrates the effectiveness of the hybrid model in detecting various types of occluded faces.

- Top left: Successful detection of faces covered with surgical masks.
- Top-right: Low-visibility and tinnyfaces are correctly detected.
- Bottom right: Complex occlusions, such as hands covering faces, are correctly identified.

4.5. Inference Time and Model Size

Table 9. Computational Efficiency

Model	Average Inference Time (per image, CPU)	Model Size (MB)
Viola-Jones Only	0.05 sec (~20 FPS)	~1 MB
AlexNet Only	0.17 sec (~6 FPS)	240 MB
Hybrid Model	**0.08 sec (~12 FPS)**	**~60 MB**

Table 9 shows the hybrid model strikes a balance between detection speed and Accuracy. It is suitable for deployment on standard computing hardware in real-time scenarios such as surveillance or masked face detection.

4.6. Model Size and Inference Speed

Model size and inference speed on CPU for the Hybrid model and baseline detectors. Blue bars denote the model's size on disk (in megabytes), and the orange dashed line with markers shows the frames per second (FPS) that each model can process on a typical CPU, as shown in Figure 3. We observe that the Hybrid model is ~60 MB and runs at around 12 FPS on a CPU. In contrast, YOLOv3 is very large (≈240 MB) and, without a GPU, achieves only around 3 FPS on the same CPU. RetinaFace (with a ResNet-50 backbone) is ~110 MB and runs at around 2 FPS on a CPU. SSD is ~100 MB and achieves around 10 FPS, while MTCNN is extremely compact (≈6 MB) but still only achieves around 5 FPS on the CPU due to its sequential 3-stage processing.

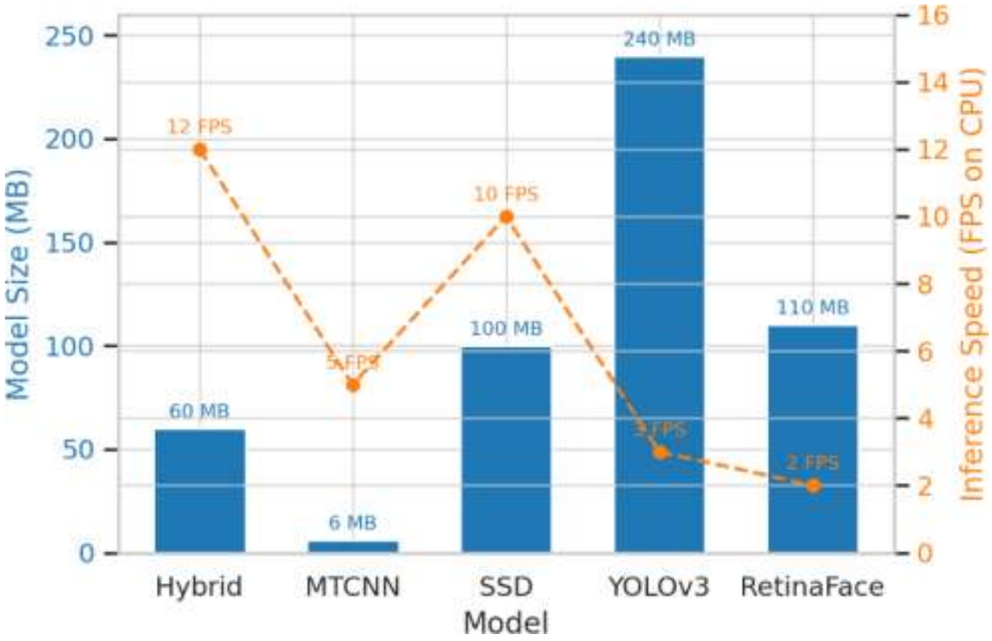

Figure 3. Model Size and Inference Speed

These results highlight the computational efficiency of the Hybrid Algorithm. It is an order of magnitude smaller than YOLOv3 and roughly half the size of RetinaFace, yet it achieves higher CPU speed than both. The Hybrid model's ~12 FPS on the CPU is real-time or better for many applications, with ~80 milliseconds per frame. YOLOv3 and RetinaFace, on the other hand, are designed for GPU execution. On a CPU, they are too slow for real-time use (≲ less than 5 FPS). MTCNN, while lightweight, is slowed by its cascaded architecture. Figure 3 shows that the Hybrid model strikes a favourable balance, combining a modest-sized deep network with the efficient Viola-Jones mechanism to yield a system that is deployable on standard hardware. It makes it well-suited for edge deployment in cameras or embedded devices where GPU resources are limited.

In summary, the Hybrid Algorithm meets the practical requirements of speed, size efficiency, and Accuracy. It can run in real-time on a CPU with a reasonable memory footprint, which underscores its value for real-world surveillance and mobile applications.

CONCLUSION

Detecting faces under partial occlusion remains a persistent challenge in computer vision, especially in real-world scenarios such as surveillance, biometric authentication, and public safety systems. In this research, a hybrid face detection model was proposed, integrating the speed of the Viola-Jones algorithm with the robustness of a modified AlexNet-based CNN classifier. The primary objective was to enhance face detection performance under occluded conditions, where conventional detectors often struggle due to missing or obstructed facial features. The model was trained and evaluated on a dataset of 10,000 images comprising 5,000 samples with various forms of occlusion, including masks, sunglasses, and hands. The hybrid architecture used Viola-Jones for rapid face proposal generation, followed by AlexNet to verify each candidate region. This approach allowed for a balance between detection accuracy and computational efficiency. The final results demonstrated the effectiveness of this design, achieving an overall accuracy of 93.0%, precision of 95.0%, recall of 90.1%, and an F1-score of 92.5%, with a false positive rate as low as 3.0%. Additionally, the model maintained an inference speed of approximately 12 frames per second on a standard CPU, with a memory footprint of just 60 MB, making it suitable for real-time and resource-constrained deployments.

Compared to standalone approaches, the hybrid model significantly reduced false positives and improved robustness to partial occlusions. Viola-Jones alone showed limited Accuracy (72.4%) and a high false positive rate (15.2%), while AlexNet alone was more accurate but slower. The hybrid design successfully combined their strengths, resulting in improved detection of partially visible faces across different occlusion types. In summary, the proposed method offers a practical, scalable, and effective solution for occluded face detection. Future work may explore the integration of attention mechanisms or transformer-based models further to enhance performance in highly crowded or heavily occluded environments.

REFERENCES

[1] P. Viola and M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, 2004. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[2] L. N. Soni, A. Datar, and S. Datar, "Implementation of Viola-Jones Algorithm Based Approach for Human Face Detection," International Journal of Current Engineering and Technology, vol. 7, no. 5, pp. 1819–1823, Sept.-Oct. 2017.

[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint, arXiv:1804.02767, 2018.

[4] J. Wang, C. Liu, and X. Zhou, "Detecting occluded faces using part-based deep networks," Neurocomputing, vol. 383, pp. 318–327, 2020. https://doi.org/10.1016/j.neucom.2019.11.109

[5] J. Yu, Y. Jiang, and Z. Wang, "UnitBox: An advanced object detection network," ACM International Conference on Multimedia (ACM MM), pp. 516–520, 2016. https://doi.org/10.1145/2964284.2967270

[6] Z. He, T. Zhao, and W. Yang, "Hybrid approach for real-time face detection using Adaboost and deep CNN," International Journal of Machine Learning and Cybernetics, vol. 12, pp. 1793–1804, 2021. https://doi.org/10.1007/s13042-020-01251-0

[7] H. Wang, X. Zhu, and Z. Lei, "Face detection under occlusion via coarse-to-fine feature aggregation," IEEE Transactions on Image Processing, vol. 30, pp. 5291–5302, 2021. https://doi.org/10.1109/TIP.2021.3088481

[8] Y. Zhong, X. Li, and J. Liu, "Learning deep face representation with facial attributes for occluded face recognition," Pattern Recognition, vol. 102, pp. 107234, 2020. https://doi.org/10.1016/j.patcog.2020.107234

[9] A. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Detecting and localising occluded faces," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1899–1906, 2014. https://doi.org/10.1109/CVPR.2014.246

[10] J. Deng, Y. Zhou, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5203–5212, 2020. https://doi.org/10.1109/CVPR42600.2020.00525

[11] C. Chen, X. Song, and J. Li, "Face detection with improved Faster R-CNN in the occluded environment," Applied Sciences, vol. 9, no. 10, pp. 2060, 2019. https://doi.org/10.3390/app9102060

[12] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, "SSH: Single stage headless face detector," IEEE International Conference on Computer Vision (ICCV), pp. 4875–4884, 2017. https://doi.org/10.1109/ICCV.2017.521

[13] S. Li and W. Deng, "BlendedMosaic: A data augmentation technique for training robust face recognition models," arXiv preprint, arXiv:2007.11298, 2020.

[14] S. Zhang, R. Benenson, and B. Schiele, "Occlusion-aware face detection," arXiv preprint, arXiv:1903.12290, 2019.

[15] S. Ge, J. Li, and Q. Ye, "Detecting masked faces in the wild with LLE-CNNs," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2682–2690, 2017. https://doi.org/10.1109/CVPR.2017.287

[16] M. Kumar, R. Saini, and B. S. Saini, "A review of face detection techniques under partial occlusion," Procedia Computer Science, vol. 132, pp. 1183–1190, 2018. https://doi.org/10.1016/j.procs.2018.05.217

[17] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-driven face detection: Improved model and training," Image and Vision Computing, vol. 59, pp. 28–39, 2017. https://doi.org/10.1016/j.imavis.2016.12.001

[18] 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, 2016. https://doi.org/10.1109/LSP.2016.2603342

[19] Y. Liu, L. Li, and H. Wu, "Hybrid face detector using Viola-Jones and CNN verification," Pattern Recognition Letters, vol. 142, pp. 50–56, 2021. https://doi.org/10.1016/j.patrec.2020.11.004

[20] L. N. Soni, A. Datar, and S. Datar, "Implementation of Viola-Jones Algorithm Based Approach for Human Face Detection," International Journal of Current Engineering and Technology, vol. 7, no. 5, pp. 1819-1823, Sept. 2017.

[21] Y. Xu, L. Lin, and X. Wu, "Robust facial landmark detection under significant head pose and occlusion," Neurocomputing, vol. 219, pp. 439–448, 2017. https://doi.org/10.1016/j.neucom.2016.09.104

[22] L. Chen, H. Zhou, and X. Zhang, "Multi-scale contextual face detection for occlusion handling," Sensors, vol. 21, no. 4, pp. 1339, 2021. https://doi.org/10.3390/s21041339

[23] L. N. Soni and A. A. Waoo, "A review of recent advances in methodologies for face detection," International Journal of Current Engineering and Technology, vol. 13, no. 2, pp. 86-92, 2023.