A COMPREHENSIVE REVIEW OF MULTI-TASK DEEP LEARNING APPROACHES FOR FACIAL ANALYSIS IN UNCONSTRAINED ENVIRONMENTS

Rajesh Prasad[1], Nitish Gupta[2]

[1,2]Department of Computer Science

NRI institute of information science and technology Bhopal India

Abstract

Facial analysis has progressed rapidly with the adoption of deep learning, yet most existing approaches remain task-specific, focusing individually on detection, landmark extraction, age estimation, gender prediction, pose estimation, or recognition. This fragmentation limits their applicability in real-world scenarios, particularly in crowded or unconstrained environments. This review paper presents a comprehensive analysis of forty state-of-the-art research works covering face detection, multi-task learning, landmark localization, biometric vulnerabilities, age and gender prediction, and robust facial analytics in unconstrained settings. Through systematic comparison, the review identifies major limitations in current systems, including their reduced performance under occlusion, pose variation, illumination changes, and large-group conditions. The study highlights the need for a unified deep-learning-based framework capable of simultaneously extracting multiple facial attributes while maintaining scalability and robustness. Based on the consolidated insights, the paper proposes the conceptual design of a multi-task facial analysis system built on the Buffalo_L architecture, integrating detection, 2D/3D landmarks, age and gender estimation, pose analysis, and face embeddings within a single pipeline. This review aims to guide future research toward developing more comprehensive, scalable, and real-world-ready facial analysis solutions.

Keywords: Face Detection, gender, pose, landmark, Image processing

1. INTRODUCTION

In recent years, intelligent visual interpretation has become essential across multiple domains such as security, healthcare, retail analytics, social media, and human–computer interaction. Among these applications, facial analysis plays a crucial role because the human face provides rich information related to identity, demographics, and behavior. Facial analytics has evolved from merely detecting a face in an image to performing more complex tasks such as recognition, age estimation, emotion understanding, and 3D landmark localization. A visual representation of this evolution is shown in Figure 1.1, where facial analysis has transitioned from classical methods to powerful deep learning–based approaches.

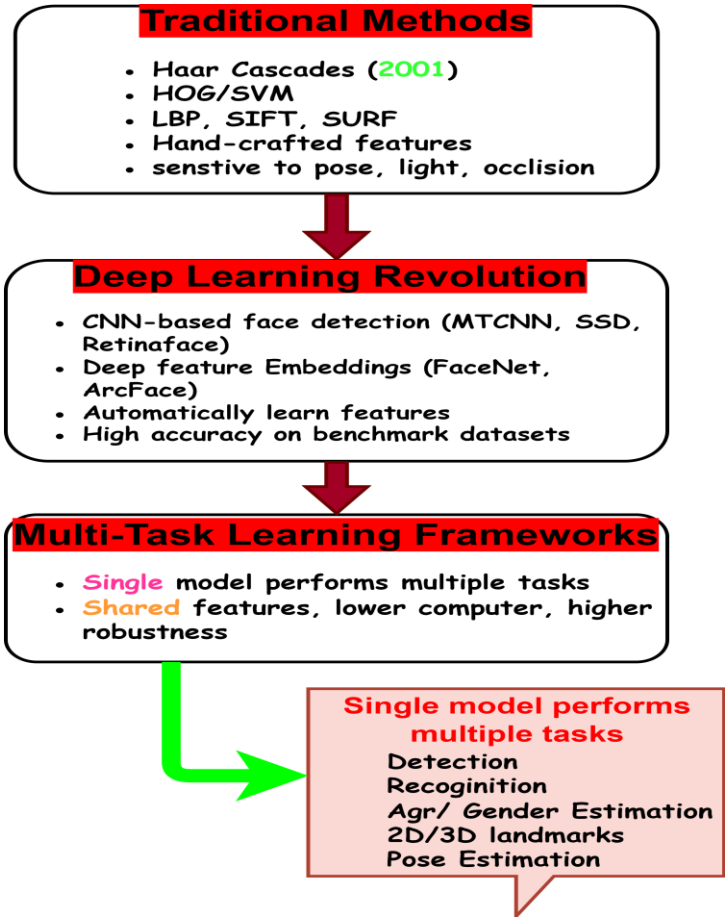

Figure 1.1: Evolution of Facial Analysis Techniques

Traditional computer vision techniques—such as Haar cascades, Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG)—formed the foundation of early facial analysis systems. However, these handcrafted feature-based methods were highly sensitive to variations in illumination, pose, expression, and occlusion, limiting their usability in real-world environments. Deep learning revolutionized the field by enabling models to learn hierarchical features directly from large-scale facial datasets.

Convolutional Neural Networks (CNNs) and attention-based architectures have significantly improved the accuracy of face detection, recognition, landmark localization, and demographic estimation. Yet, despite these advancements, most modern systems remain task-specific, requiring separate models for detection, alignment, recognition, and age estimation. Deploying multiple models increases computational cost and introduces latency, especially in real-time applications such as smart surveillance or crowd monitoring.

2. CHALLENGES IN IMAGE PROCESSING

Real-world environments introduce additional challenges: occlusion by objects or other faces, extreme pose angles, low-light conditions, low-resolution imagery, and dense groups of people appearing simultaneously. These challenges are illustrated in Figure 1.2, highlighting the complexity of robust facial analysis in unconstrained conditions. Single-task models often fail to generalize across such scenarios, making them unsuitable for large-scale or real-time use. To address these limitations, Multi-Task Learning (MTL) has emerged as a promising paradigm. MTL allows a single neural network to learn shared representations for multiple tasks, enhancing efficiency and improving accuracy across related outputs. By integrating detection, 2D/3D landmark prediction, age and gender estimation, pose estimation, and facial recognition into one unified framework, computational redundancy is reduced while system robustness increases.

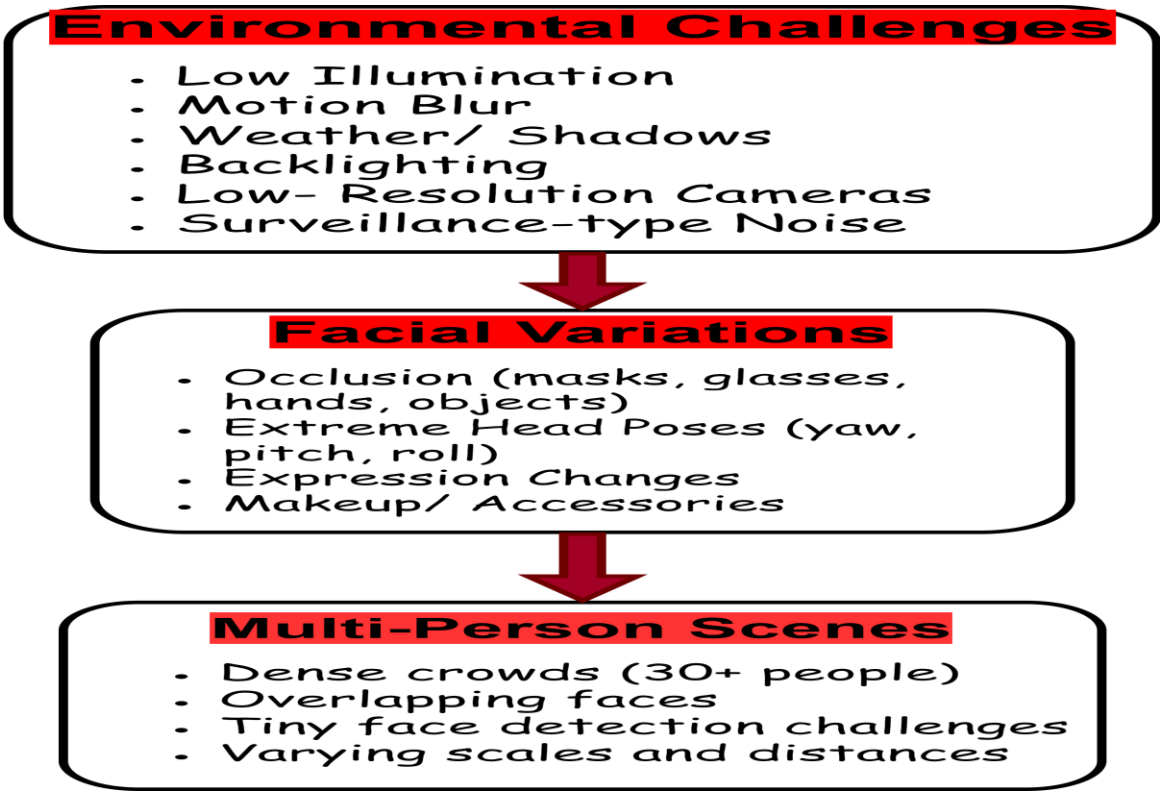

Figure 1.2: Challenges in Real-World Facial Analysis

Over the years, researchers have proposed a wide spectrum of techniques, ranging from classical handcrafted approaches like Haar cascade detectors, skin-color segmentation, and edge-based methods, to more sophisticated deep learning models capable of learning hierarchical facial representations. Parallel to these developments, increased concerns around biometric security, fairness, and privacy have led to focused investigations on morphing attacks, makeup-induced spoofing, template inversion vulnerabilities, and demographic bias in large-scale face

recognition systems. The literature also reflects a growing interest in multi-task learning frameworks that aim to jointly learn facial attributes to improve efficiency and performance.

3. LITERATURE REVIEW

This chapter provides a comprehensive review of key research contributions relevant to face detection and facial analytics. It systematically surveys traditional approaches, deep learning–driven models, multi-task architectures, studies addressing robustness and security, and works highlighting bias and fairness in recognition systems. By critically analyzing more than forty research papers, this chapter identifies conceptual trends, methodological advancements, comparative performance insights, and limitations in existing works. The insights gained from this extensive review form the foundation for the proposed multi-task facial analysis framework presented later in this thesis.

Tian & Zhao, (2012) presented a fast face detection method specifically designed for JPEG images by leveraging characteristics of the compressed domain. They proposed reconstructing a low-resolution version of the image using only a small number of DCT coefficients and then applying pixel-domain face detection on this simplified representation. Their study demonstrated that this hybrid strategy significantly accelerates detection while maintaining high precision when compared to conventional pixel-domain approaches. The work highlights the potential of compressed-domain processing for lightweight face detection but remains limited to JPEG-specific operations.

Pamplona Segundo et al., (2011) presented a real-time scale-invariant face detection technique that operates directly on range images. They proposed the use of boosted cascade classifiers that eliminate the need for multi-scale scanning, allowing the system to maintain real-time performance even under substantial pose variations. Their experiments demonstrated impressive results, achieving a detection rate of 99.9% with minimal false detections, while performing equally well on frontal and non-frontal faces. Although highly effective, the dependence on depth imaging hardware may limit broad applicability in typical image-based systems.

Dang & Sharma, (2017) presented a comprehensive review and comparison of classical face detection algorithms including Viola–Jones, SMQT with SNOW classifiers, neural network–based approaches, and SVM-based methods. They analyzed each method using precision and recall measurements generated through the DetEval tool, enabling an objective evaluation of bounding box accuracy. Their study demonstrated the strengths and weaknesses of each approach, showing how different algorithms perform under varying conditions. However, as a review paper, it does not propose new methods but emphasizes the need for more robust detectors as image databases continue to grow.

Alabbasi & Moldoveanu, (2014) presented a skin-color-based face detection method designed to operate under unconstrained illumination and pose conditions. They proposed combining segmentation in RGB, HSV, and YCbCr color spaces—enhanced with elliptical modeling—followed by edge fusion and morphological processing to refine

detected regions. Their results demonstrated high accuracy for single-face images of varying sizes and expressions. Nevertheless, the approach remains sensitive to non-skin background regions and may struggle in multi-face or low-skin-visibility scenarios.

Putro et al., (2016) presented a classifier for distinguishing adult images from benign content using face detection as the primary cue. They proposed integrating the Viola–Jones detector with HS skin-color analysis, extracting features such as face area percentage and skin coverage to support classification. Their study demonstrated an accuracy of 90% on sample images with low false positives and negatives, showing that face-based cues can serve as effective predictors of adult content. However, the small dataset size limits the generalizability of their findings.

Luo et al., (2019) presented the Small Faces Attention (SFA) detector, designed to overcome the challenge of detecting small-scale faces that degrade the performance of traditional CNN-based detectors. They proposed a multi-branch architecture where each branch specializes in a specific face scale, combined with feature fusion from adjacent branches to enhance small-face detection. Their experiments demonstrated substantial performance improvements on benchmarks such as WIDER FACE and FDDB while maintaining competitive runtime. The approach is highly effective but introduces architectural complexity requiring careful training.

Storey et al., (2018) presented the Integrated Deep Model (IDM), which fuses Faster R-CNN with a stacked hourglass network to jointly improve face detection and facial landmark localization. They proposed a novel optimization mechanism to integrate the two networks, reducing false positives by 62% and improving localization accuracy with minimal added error. Their study demonstrated high recall and precision across multiple challenging datasets, highlighting the value of multi-task integration. Despite strong results, the system remains computationally heavier than single-stage detectors.

Lu & Chuang, (2022) presented a deep learning–based method for human and face detection under varying infrared illumination conditions. They proposed using the MI3 multi-intensity IR dataset and introduced a fusion technique that combines multiple IR exposure levels to overcome underexposure and overexposure in surveillance scenarios. Their experiments demonstrated improved detection across long-range and near-range conditions when using detectors such as SSD, YOLO, Faster R-CNN, and Mask R-CNN. This work contributes a valuable dataset but is largely dependent on IR-specific imaging scenarios.

Tan et al., (2018) presented the first systematic evaluation of face detection and verification using lensless cameras, exploring the feasibility of deploying such systems in low-cost IoT devices. They proposed using deep learning models adapted to handle the noise and low resolution inherent in lensless FlatCam images. Their study demonstrated that both face detection and verification remained feasible with high accuracy despite the significant degradation in visual clarity. The work opens a new direction in ultra-compact camera systems, though resolution limitations continue to pose challenges.

Wu et al., (2025) presented an end-to-end deep network for jointly detecting faces and facial landmarks in real time. They proposed modifying the YOLO architecture by incorporating multitarget labels and adding a specialized head for landmark localization. Their enhancements, including structural re-parameterization and channel-shuffling modules, significantly improved accuracy and inference speed. Experimental evaluations demonstrated strong performance on datasets such as 300W, COFW, and AFLW2000-3D, outperforming many state-of-the-art models while remaining efficient for edge devices. This work underscores the growing relevance of unified face-and-landmark architectures for real-time applications.

Some more papers were analysed and a comparative table is presented as table 1, which show the proposed work improvement which can be done agaist the work presented in the articles

Table 1: Comparative Review of Existing Facial Analysis Models

Ref	Work Done	Limitations	Features of Proposed work
Sahoo & Banka, (2018)	Proposed multi-feature fusion (local+global) and modified EM feature filling for age estimation; hierarchical SVM/SVR on FG-NET, MORPH.	➢ Focused on age only ➢ does not provide 2D/3D landmarks, pose, embedding or recognition. ➢ limited evaluation on multi-face/group images.	We integrate age estimation within a full multi-task pipeline (detection + 2D/3D landmarks + pose + gender + embedding) and evaluate on single, small-group and large-group images.
Jing et al., (2023)	Comprehensive survey of 3D face recognition methods, datasets and challenges; comparative analysis of accuracy/robustness.	➢ Survey, no single system implementation. ➢ lacks evaluation of unified 2D+3D multi-task pipelines on crowd images.	We combine 3D landmark outputs with 2D features and recognition embeddings in one operational pipeline and test on large-group images to evaluate practical robustness.
Ranjan et al., (2019)	Multi-task CNN fusing intermediate layers to jointly do face detection, 2D landmark localization, pose estimation and gender recognition; shows benefits of joint learning.	➢ Does not include age estimation, 3D landmarks, or large-crowd testing. ➢ recognition embedding integration is absent.	Our work extends HyperFace-style multi-task learning by adding age estimation, 3D landmarks, identity embeddings and large-group evaluations to improve completeness and real-world applicability.
Prasher et al., (2024)	Experimented with CNNs for age/gender classification; practical/real-time considerations and dataset experiments.	➢ Single-task focus ➢ limited multi-face/crowd evaluation ➢ often sensitive to pose/occlusion and lacks 3D features.	We embed age/gender estimation as part of a single multi-task Buffalo_L pipeline so demographic predictions benefit from joint features (landmarks, pose, embedding).
Rohith et al., (2023)	Explores ML/DL models for real-time age estimation; deployment aspects for real-time apps.	➢ Only addresses age; limited to per-face evaluation and not a unified multi-task system. ➢ crowd scenes not evaluated.	Our system performs age estimation with simultaneous landmark/pose/detection so it is robust for multi-face and crowd contexts.
Poddar & Sharma, (2024)	Evaluated gender recognition CNNs on balanced datasets and suggested architectural tradeoffs for efficiency and accuracy.	➢ Gender only ➢ single-task ➢ not integrated with recognition or 3D features ➢ limited crowd testing.	We integrate gender prediction in a holistic pipeline and test on varying group sizes to show scalability and reduced redundancy.

4. PROPOSED SOLUTIONS

Motivated by these advantages, the present research proposes a comprehensive multi-task facial analysis system using the Buffalo_L deep learning model, capable of simultaneously performing all major facial tasks within a single pipeline. The high-level architecture of this unified framework is presented in Figure 1.3.

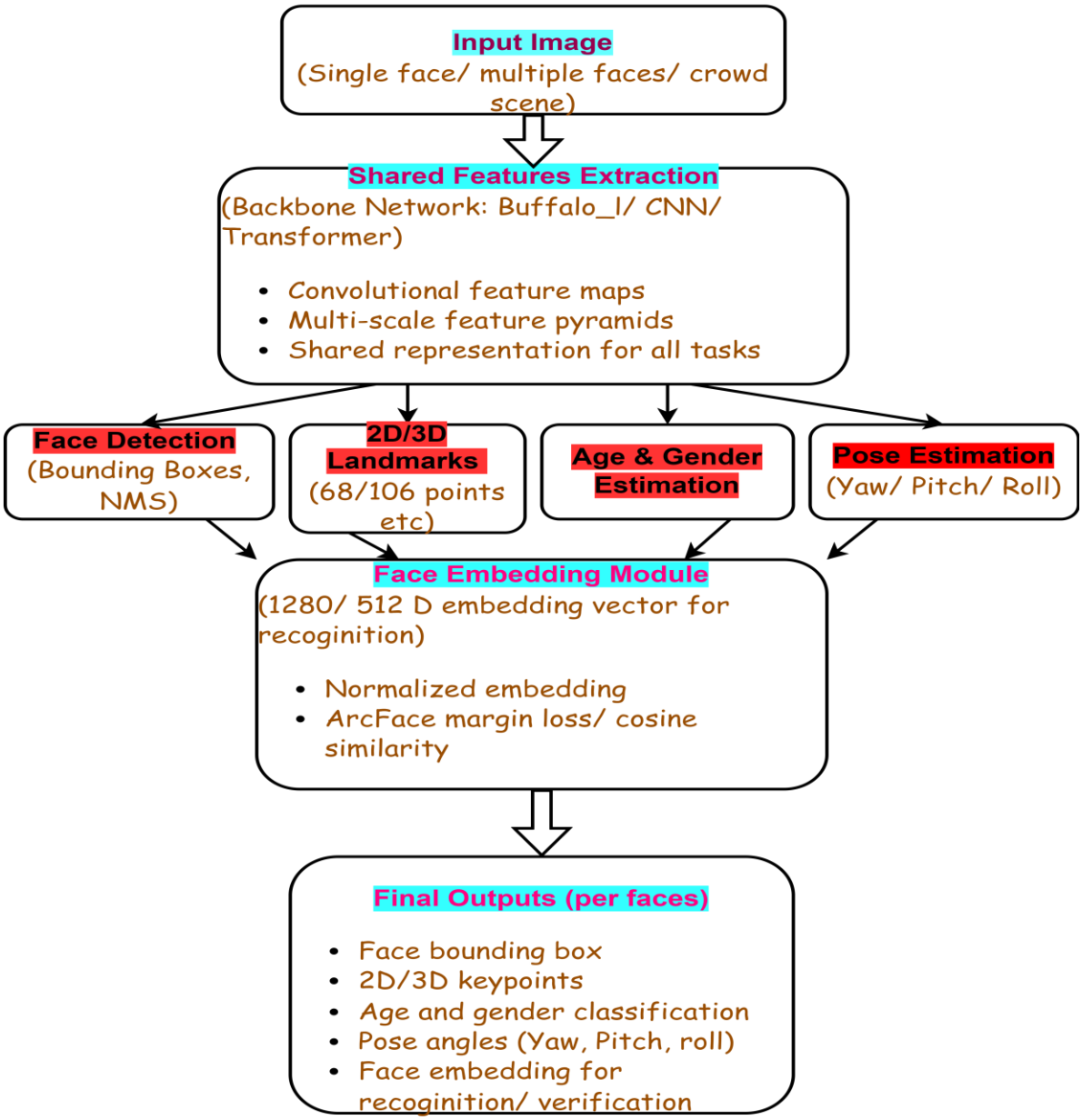

Figure 1.3: Proposed Multi-Task Facial Analysis Framework

The extensive review of existing literature leads to the following major findings:

- **Single-task models dominate the field**, with most studies addressing detection, age estimation, or gender classification in isolation rather than as part of an integrated system.
- **Deep learning significantly outperforms handcrafted methods**, yet even advanced CNN-based models often struggle in crowded scenes or under occlusion, pose variation, and extreme illumination.
- **Very few works incorporate both 2D and 3D landmarks**, or evaluate performance on large-group images exceeding 20–30 individuals.
- **Bias and fairness issues** persist across demographics, hairstyles, facial attributes, and age groups, suggesting the need for fairness-aware training strategies.
- **Security vulnerabilities remain substantial** due to makeup attacks, morphing attacks, and template inversion, which several reviewed papers confirm.
- **Multi-task learning approaches are promising**, but existing ones (e.g., HyperFace) lack age estimation, 3D landmarks, or evaluation on group images.
- **No reviewed system provides a single unified pipeline** capable of handling detection, landmarks, pose, age, gender, and recognition simultaneously.
- **Crowd-based and real-world datasets** remain limited, making it difficult to evaluate scalability and generalization.
- **Modern architectures often demand high computational resources**, limiting deployment on embedded or edge devices.
- These insights collectively motivate the **proposed conceptual framework** based on Buffalo_L, addressing all major gaps through a unified multi-task approach.

5. CONCLUSION

This review paper provides a detailed examination of contemporary developments in facial analysis, synthesizing findings from more than forty influential studies across deep learning–based detection, recognition, landmark localization, age and gender estimation, pose analysis, presentation attack detection, and fairness assessment. The review demonstrates the substantial progress made in facial analytics while simultaneously exposing critical limitations in existing approaches, including fragmentation across tasks, declining performance in unconstrained settings, demographic bias, and limited scalability to crowded scenes.

Based on the identified gaps, the paper proposes the conceptual foundation for a unified multi-task framework built on the Buffalo_L deep-learning architecture. This framework integrates face detection, 2D/3D landmark extraction, pose estimation, age and gender prediction, and face embedding generation into a single operational pipeline. Such an approach promises improved efficiency, accuracy, and robustness, especially for real-world applications in surveillance, crowd monitoring, and human–computer interaction.

Although the present work does not include the implementation or experimental evaluation of the proposed system, it establishes a clear research direction and provides a comprehensive foundation for future studies. The insights and synthesized evidence presented in this review offer a valuable roadmap for researchers aiming to develop next-generation multi-task facial analysis systems capable of operating reliably in complex, real-world environments.

REFERENCES

1. Alabbasi, H. A., & Moldoveanu, F. (2014). Human face detection from images, based on skin color. *2014 18th International Conference on System Theory, Control and Computing, ICSTCC 2014*, 532–537. https://doi.org/10.1109/ICSTCC.2014.6982471
2. Dang, K., & Sharma, S. (2017). Review and comparison of face detection algorithms. *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, 629–633. https://doi.org/10.1109/CONFLUENCE.2017.7943228
3. Jing, Y., Lu, X., & Gao, S. (2023). 3D face recognition: A comprehensive survey in 2022. *Computational Visual Media 2023 9:4*, *9*(4), 657–685. https://doi.org/10.1007/S41095-022-0317-1
4. Lu, P. J., & Chuang, J. H. (2022). Fusion of Multi-Intensity Image for Deep Learning-Based Human and Face Detection. *IEEE Access*, *10*, 8816–8823. https://doi.org/10.1109/ACCESS.2022.3143536
5. Luo, S., Li, X., Zhu, R., & Zhang, X. (2019). SFA: Small Faces Attention Face Detector. *IEEE Access*, *7*, 171609–171620. https://doi.org/10.1109/ACCESS.2019.2955757
6. Pamplona Segundo, M., Silva, L., & Bellon, O. R. P. (2011). Real-time scale-invariant face detection on range images. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 914–919. https://doi.org/10.1109/ICSMC.2011.6083768
7. Poddar, K., & Sharma, B. (2024). Efficient Gender Recognition with Deep Learning Models on Balanced Facial Image Datasets. *8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2024 - Proceedings*, 1439–1443. https://doi.org/10.1109/I-SMAC61858.2024.10714830
8. Prasher, S., Nelson, L., & Arumugam, D. (2024). Deep Learning Models for Age and Gender Prediction Using Facial Images. *2024 5th International Conference for Emerging Technology, INCET 2024*. https://doi.org/10.1109/INCET61516.2024.10593323
9. Putro, M. D., Adji, T. B., & Winduratna, B. (2016). Adult image classifiers based on face detection using Viola-Jones method. *Proceeding of 2015 1st International Conference on Wireless and Telematics, ICWT 2015*. https://doi.org/10.1109/ICWT.2015.7449208
10. Ranjan, R., Patel, V. M., & Chellappa, R. (2019). HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 121–135. https://doi.org/10.1109/TPAMI.2017.2781233
11. Rohith, S. V., Khan, L. A., Archana, V., Jayanthi, M. G., & Kannadaguli, P. (2023). Human Age Estimation from Images in Real-Time Application Using Machine Learning and Deep Learning Models. *2023 International Conference on Network, Multimedia and Information Technology, NMITCON 2023*. https://doi.org/10.1109/NMITCON58196.2023.10275901

12. Sahoo, T. K., & Banka, H. (2018). Multi-feature-Based Facial Age Estimation Using an Incomplete Facial Aging Database. *Arabian Journal for Science and Engineering 2018 43:12*, *43*(12), 8057–8078. https://doi.org/10.1007/S13369-018-3293-0
13. Storey, G., Bouridane, A., & Jiang, R. (2018). Integrated Deep Model for Face Detection and Landmark Localization From "In The Wild" Images. *IEEE Access*, *6*, 74442–74452. https://doi.org/10.1109/ACCESS.2018.2882227
14. Tan, J., Niu, L., Adams, J. K., Boominathan, V., Robinson, J. T., Baraniuk, R. G., & Veeraraghavan, A. (2018). Face Detection and Verification Using Lensless Cameras. *IEEE Transactions on Computational Imaging*, *5*(2), 180–194. https://doi.org/10.1109/TCI.2018.2889933
15. Tian, Q., & Zhao, S. (2012). A fast face detection method for JPEG image. *International Conference on Signal Processing Proceedings, ICSP*, *2*, 899–902. https://doi.org/10.1109/ICOSP.2012.6491725
16. Wu, Q., Wang, X., Li, N., Fong, S., Zhang, L., & Yang, J. (2025). Real-Time Face and Facial Landmark Joint Detection Based on End-to-End Deep Network. *IEEE Transactions on Instrumentation and Measurement*, *74*. https://doi.org/10.1109/TIM.2025.3541698