J. Res. Technol. Eng. 6 (3), 2025, 160-175

Journal Of Research Technology & Engineering
www.jrte.org

Low Power Face Recognition Using FPGA and GPU Hybrid

*D T N D Silva, Udara S.P.R. Arachchige, A S Goonetilleke, Prabath Buddika
Faculty of Engineering NSBM Green University
*thusaraka@ieee.org

Received:28 June 2025; Revised: 30 June 2025; Accepted: 08 July 2025; Available online: 10 July 2025

Abstract: The escalating demand for efficient and secure attendance management systems has driven innovation beyond traditional manual methods, which are inherently prone to inaccuracies, inefficiency, and fraud. Artificial intelligence (AI)-driven facial recognition emerges as a transformative solution, offering real-time, non-intrusive, and highly accurate identification. This review paper synthesizes the current state of research in real-time face detection and recognition, particularly for attendance systems, across various hardware platforms and algorithmic approaches. It critically examines existing review papers and individual studies to identify significant research gaps in achieving comprehensive, energy-efficient, and robust solutions for crowded environments. Specifically, this paper highlights the lack of holistic hybrid hardware-software co-optimization, integrated intelligent low-power management with hybrid wake-up mechanisms, and seamless real-time database integration for high-throughput, multi-face data in existing literature. We then position a novel hybrid Field-Programmable Gate Array (FPGA) and NVIDIA Jetson AGX architecture as a promising solution to these identified gaps. This proposed system leverages the FPGA for low-latency image pre-processing and motion-triggered power management, and the Jetson AGX for high-performance deep learning inference and real-time database updates, thereby addressing the complex challenges of real-time attendance in crowded, power-constrained environments.

Index Terms: deep learning, embedded systems, face recognition, FPGA, low-power design

1 INTRODUCTION

Managing attendance is a fundamental administrative task across various institutions, from educational settings to corporate enterprises. Historically, attendance tracking has relied on traditional methods such as manual roll calls, paper-based sign-ins, or ID card-based systems[1]. While seemingly straightforward, these conventional approaches are fraught with inherent limitations that compromise their effectiveness and reliability. They are notably inefficient, consuming significant time that could otherwise be allocated to core activities, and are highly susceptible to human error and deliberate manipulation, such as "buddy-punching". The lack of real-time data availability further impedes prompt decision-making and limits the

ability to gain immediate insights into attendance patterns.

In response to these persistent challenges, modern systems are increasingly embracing advanced technological solutions, with facial recognition technology emerging as a particularly promising innovation. Powered by artificial intelligence (AI) and machine learning (ML), facial recognition offers a non-intrusive, highly efficient, and remarkably accurate alternative for monitoring attendance [2]. These systems leverage state-of-the-art deep learning techniques, such as Convolutional Neural Networks (CNNs), to analyze facial features captured through real-time video streams, enabling automatic identification and verification of individuals as they enter a designated area, marking attendance instantly and seamlessly [2]. The immediate identification and verification capabilities significantly reduce manual effort, virtually eliminating human errors and the potential for fraud [1]. However, due to the increasing complexity of AI workloads, especially at the edge, deploying such systems efficiently requires the integration of heterogeneous hardware. Cole [3] highlights that combining CPUs for control logic, GPUs for parallel data processing, and FPGAs for low-latency inference is critical for achieving performance, energy efficiency, and responsiveness in embedded AI deployments. Tibbetts et al. [4] further extend this perspective by surveying Smart NIC and DPU-based heterogeneous architectures, which emphasize intelligent task delegation and edge-level processing to support scalable, real-time AI workloads under constrained power and latency budgets.

While deep learning models have revolutionized computer vision tasks like face recognition, their inherent computational and memory demands pose significant challenges for deployment on resource-constrained embedded systems. Traditional general-purpose computing engines, such as those based solely on Central Processing Units (CPUs) or Graphics Processing Units (GPUs), often struggle to meet the stringent requirements for real-time processing, energy efficiency, and compact form factors demanded by edge applications.1 This has led to the emergence of heterogeneous architectures, which combine different types of processors (CPUs, GPUs, FPGAs, and specialized AI accelerators) to offload computational stages to the most suitable processing element[1]. This approach aims to balance computational power, energy efficiency, and real-time responsiveness in complex AI applications at the edge[1].

This paper aims to provide a comprehensive review of the current research landscape in real-time face detection and recognition for attendance systems. It will synthesize findings from existing literature, including dedicated survey and review papers, to estab—lish the state-of-the-art in algorithmic approaches, hardware implementations, and power management strategies. Crucially, this review will identify specific research gaps that existing solutions do not fully address, particularly concerning the integration of hybrid hardware architectures for crowded environments and intelligent low-power operation. Finally, we will position a novel hybrid FPGA-Jetson AGX system as a promising approach to fill these identified gaps,

offering a more comprehensive and efficient solution for next-generation attendance management

2. Landscape of Face Recognition Systems: A Review of Reviews

The field of face recognition has seen extensive research, with numerous studies and review papers attempting to categorize and evaluate various approaches. This section synthesizes the findings from these existing reviews and individual research efforts to establish the current state of the art, highlighting both achievements and persistent challenges.

2.1. General Face Recognition Algorithms and Performance

Face recognition systems typically involve three main stages: face detection, feature extraction, and face recognition. A wide array of algorithms has been developed for these tasks, ranging from traditional machine learning methods to advanced deep learning models.

Early approaches often relied on traditional machine learning techniques. For instance, the Viola-Jones algorithm, introduced in 2001, became famous for its real-time performance and high training rate, utilizing Haar features and a cascaded classifier [5]. While effective for simple images, it struggles with variations like glasses or masks and is less accurate than modern deep learning methods [5]. Other traditional methods include Principal Component Analysis (PCA) and Local Binary Patterns (LBP). PCA is a statistical approach for dimensionality reduction, less sensitive to noise and memory-efficient, but requires large training datasets. LBP describes image texture and shape, is invariant to luminosity and rotations, and is suitable for highly parallel architectures. However, these traditional methods often lack robustness against real-world variations such as low resolution, complex backgrounds, and significant occlusions [6].

A promising approach involves hybrid models that integrate statistical methods with neural networks. For instance, systems combining Discrete Cosine Transform (DCT) for feature extraction with neural networks for classification have demonstrated improved performance in constrained environments, offering a balance between accuracy and computational efficiency [7].

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized face recognition, offering superior accuracy and robustness by learning intricate features directly from raw images. The evolution of deeper convolutional architectures, such as the Inception model proposed by Szegedy et al., has further enhanced the efficiency and scalability of CNN-based face recognition systems [8]. Prominent deep learning models for face detection include Multi-task Cascaded Convolutional Networks (MTCNN), Single Shot Detector (SSD), and the You Only Look Once (YOLO) family [7]. YOLO variants, like YOLOv3 and YOLOv7, are highly regarded for their balance of speed and accuracy,

making them suitable for real-time applications, even recognizing partially obscured faces. MTCNN excels at simultaneous face detection, landmark localization, and alignment, showing robustness against pose, illumination, and partial occlusions [9]. For face recognition, FaceNet is a state-of-the-art deep convolutional network that maps face images into a compact Euclidean embedding space, achieving high accuracy (e.g., 99.63% on the LFW dataset) [10]. Sharma and Kumar [11] demonstrate the effectiveness of an optimized deep learning model specifically designed for real-time object detection in surveillance environments, highlighting the feasibility of deploying lightweight yet highly accurate architectures on embedded platforms. Recent review articles have synthesized the evolution of deep learning for face recognition. For example, Tolu et al. [12] conducted a systematic review on the application of deep learning techniques to face recognition, discussing various CNN architectures and training methodologies tailored for face identification and verification tasks. Their analysis emphasizes the importance of dataset diversity, data augmentation, and the use of transfer learning to enhance model generalization across variable lighting, pose, and occlusion conditions. However, as facial recognition systems grow increasingly accurate and widely adopted, they also raise serious privacy concerns. Moens and Kornet [13] critically examine whether these systems can be reconfigured to preserve user privacy, highlighting recent approaches like federated learning and differential privacy to mitigate the risks associated with centralized biometric data processing. Recent innovations like **YuNet++**, a millisecond-level face detector, showcase how efficient and lightweight architectures can further reduce latency while maintaining accuracy, making them highly suitable for real-time embedded deployment [14].

To further enhance face recognition under adverse conditions such as low light or occlusion, multimodal approaches that fuse infrared and visible spectrum images have shown promising results. Li et al. [15] proposed a CNN-based fusion model that combines spatial and feature-level data from both modalities, significantly improving recognition accuracy in low-visibility environments. Complementing this, recent work by Zhang et al. [16][17] presents an infrared-visible image fusion technique leveraging hierarchical deep feature extraction and joint multi-scale attention mechanisms, further enhancing performance under variable environmental conditions.

Several review papers have extensively discussed these algorithms. Baobaid et al. (2022) provide a comprehensive review of recent face recognition algorithms, categorizing them into neural network and non-neural network approaches, and comparing their accuracy and processing time [18]. They highlight that neural network algorithms generally outperform non-neural network ones in accuracy, with FaceNet achieving the highest reported accuracy on the LFW dataset [1]. Similarly, other surveys by Li et al. (2020), Beham and Roomi (2013), and Hasan et al. (2021) categorize and compare algorithms based on appearance, features, and soft computing, or feature-based vs. image-based approaches [1]. These reviews

consistently conclude that deep learning-based methods offer superior performance, especially in terms of accuracy [1]. A bibliometric analysis by Ramnath and Dhanabalachandran further reinforces this, identifying CNNs and deep learning frameworks as central to face recognition advancements and highlighting environmental, hardware, and algorithmic factors as primary influences on system accuracy [18].

However, a common critique across these reviews is that while they detail algorithmic enhancements and their accuracy, they often lack in-depth discussion on the specific hardware accelerators required for real-time deployment, particularly for complex hybrid architectures [19]. Furthermore, while challenges like occlusion, pose variation, and lighting changes are acknowledged [20], a comprehensive, end-to-end solution specifically optimized for

crowded, multi-face environments with high throughput remains an area of ongoing challenge[21].

2.2. Embedded System Implementations for Face Recognition

The demand for real-time face recognition at the edge has spurred significant research into embedded system implementations, leveraging various hardware platforms beyond traditional CPUs.

FPGA-Only Implementations: FPGAs are increasingly recognized for their low latency, energy efficiency, and inherent parallelism, making them suitable for real-time embedded systems. They excel at hardware-accelerated image pre-processing tasks like resizing, filtering, and feature extraction (e.g., FAST features, optical flow). For instance, FPGA-based accelerators can reduce image front-end processing latency from seconds on CPUs to milliseconds. Some studies have implemented entire CNNs or parts of them on FPGAs, demonstrating high throughput and energy efficiency, especially for lightweight models. However, FPGAs often have built-in memory limitations that restrict them to lighter, less accurate CNN models, requiring extensive external memory for larger networks, which can increase latency [22]. Jiang et al. [23] provide a comprehensive review of FPGA-based CNN acceleration techniques, outlining key strategies such as loop tiling, data reuse, parallelism, and quantization. Their findings highlight the trade-offs in latency, throughput, and resource utilization across different FPGA architectures and reinforce the relevance of FPGAs in energy-efficient CNN deployment for embedded face recognition systems [23]. Additionally, FPGA-based face recognition systems have been successfully developed for real-time environments, demonstrating that compact, hardware-efficient designs can still deliver accurate performance under constrained conditions [23]. Architectures like GF-YOLO, accelerated using hybrid overlapping strategies on FPGA platforms, have also proven highly effective for small-target detection in remote sensing applications. These designs achieve real-time inference with optimized memory access and parallelization, balancing power efficiency and detection accuracy in constrained embedded

environments [24].

Jetson-Based Implementations: NVIDIA Jetson platforms (e.g., Jetson Nano, Xavier, AGX Orin) are powerful edge AI devices specifically designed for deep learning inference. They offer substantial performance gains for deep learning tasks through their GPUs, Tensor Cores, and dedicated Deep Learning Accelerators (DLAs). Jetson devices can run complex CNN models like ResNet50, MobileNetV3, and YOLOv4 with various optimizations (e.g., FP16, INT8 quantization, TensorRT) to achieve high inference speeds. For multi-face detection in crowded scenes, Jetson AGX Orin can achieve impressive throughputs (e.g., 290 FPS for 1920×1080 frames with 6 faces) by leveraging all its hardware engines and integrating a face tracker module to avoid redundant recognition. This is consistent with the findings of Ghosh et al. [24], who demonstrate that Jetson-based platforms can deliver reliable real-time face recognition performance when combined with quantization, efficient model architectures, and hardware-specific optimization. Baobaid and Meribout [25][26] further extend this by implementing a face recognition pipeline on Jetson AGX Orin that achieves 298 FPS using integrated accelerators (Tensor Cores, NVDEC, VIC, and DLA) and a face tracker module, illustrating the impact of fine-grained hardware–software co-design in edge AI In parallel, several low-cost embedded platforms such as Raspberry Pi have also been explored for face-based attendance, leveraging traditional methods like Haar cascades and LBPH due to their lightweight nature and low computational demands. For instance, Gupta et al. [27] present a smart attendance system utilizing OpenCV libraries and face detection techniques on embedded Linux systems, demonstrating cost-effective real-time recognition suitable for educational environments. Similarly, the system developed by D. Gupta et al. uses OpenCV and the LBPH algorithm on a Raspberry Pi to implement a smart, real-time attendance system, emphasizing cost-effectiveness and ease of deployment in academic settings [27]. Aishwarya et al. [28] further reinforce this trend by proposing an embedded facial recognition system for attendance that combines OpenCV with a simple UI interface, enabling real-time monitoring and automated record generation while keeping hardware and software complexity low, making it particularly attractive for classroom and campus-wide deployment scenarios. Debadrita Ghosh [29] further proposed an OpenCV-based real-time attendance system using LBPH on a Raspberry Pi, which not only marks attendance but also sends email notifications to guardians, offering improved transparency and parental engagement in institutional settings. Hong Zhao et al. [30] built an embedded face recognition system on the Tiny6410 platform using Haar-like features for detection, PCA for feature extraction, and Euclidean distance for recognition. Their system achieves high recognition accuracy and stable performance while being suitable for portable and mobile scenarios. Extending this approach to smart education, Prasanna et al. [31] developed a hybrid face recognition system that uses CNNs and ensemble classifiers to detect and recognize students' faces in real time. Their smart classroom model improves attendance monitoring

efficiency and provides automated data storage, enabling educators to manage classroom analytics effectively. Andriyanov and Dementiev [32] contribute further by analyzing the trade-offs between algorithmic complexity, memory usage, and inference time in embedded face recognition systems, proposing an optimized pipeline for low-power platforms without compromising detection accuracy.

Hybrid CPU-GPU-FPGA Systems: Recognizing the limitations of single-accelerator solutions, research has increasingly explored heterogeneous architectures combining CPUs, GPUs, and FPGAs. These systems aim to offload specific computational stages to the most suitable processing element to enhance overall performance and energy efficiency. For instance, FPGAs can handle memory-intensive streaming computations and pipelined tasks, while GPUs excel at massively parallel data processing. Studies have shown that such hybrid systems can outperform FPGA/GPU-only accelerators in terms of runtime and energy efficiency for tasks like image classification and feature detection [1]. Existing reviews on embedded vision and heterogeneous computing discuss the benefits of combining these platforms. However, they often focus on general computer vision tasks or specific CNN architectures (e.g., ResNet18, MobileNetV2) rather than the precise, end-to-end pipeline required for a real-time attendance system in crowded environments. Furthermore, while they acknowledge communication latency as a challenge, they don't always detail the specific communication mechanisms (e.g., Ethernet vs. PCIe) and their impact on a hybrid FPGA-Jetson system for attendance. Shankar [33] complements this perspective by evaluating the design of AI-driven embedded systems that prioritize both high-performance computing and low-power operation, validating the role of heterogeneous architectures in achieving scalable, energy-efficient AI deployments at the edge.

2.3. Power Management in Embedded AI Systems

Energy efficiency is a critical design consideration for embedded systems, especially those intended for continuous operation or deployment in remote locations. Modern edge AI platforms offer sophisticated power management capabilities. Ali et al. [34] emphasized the importance of designing energy-aware computer vision deployments tailored for heterogeneous architectures, highlighting that dynamic adaptation of workloads across FPGAs, GPUs, and CPUs can significantly reduce power consumption while maintaining high performance.

Jetson Power States: NVIDIA Jetson platforms provide various power states to optimize energy consumption. The most power-efficient state is Deep Sleep (SC7), where core power rails are turned off while essential components like RTC and DRAM remain powered to enable rapid wake-up. In SC7 mode, Jetson AGX Orin can consume as little as ~300mW (CVM power) or ~1.1W (total CVM+CVB power). Deep sleep can be initiated programmatically using Linux commands like sudo systemctl suspend.

Motion-Triggered Wake-up: The system's ability to enter a low-power state during inactivity and wake up upon detecting motion is crucial for energy conservation. Common wake sources on Jetson platforms include power button presses, RTC alarms, USB hotplug, and Wake on LAN. For motion-triggered wake-up, external interrupts via GPIO pins are a common mechanism. Ultra-low power accelerometers (e.g., ADXL362) can function as motion-activated power switches, consuming very little current (e.g., <300nA in sleep mode) while continuously monitoring for motion. Upon detecting motion, such a sensor can generate a signal (e.g., AWAKE bit mapped to an output pin) that can be used to trigger the Jetson's wake-up.

Reviews on low-power embedded AI systems discuss general power optimization techniques like model pruning, quantization, knowledge distillation, and dynamic voltage and frequency scaling (DVFS). They also cover power-gating and clock-gating to reduce power consumption in unused components. However, these reviews typically discuss power management in a general sense for embedded AI. They do not often delve into the specific integration of an FPGA as a dedicated, ultra-low power motion-sensing guardian that precisely controls the wake-up and sleep cycles of a more powerful Jetson AGX for an attendance system. The detailed interplay between FPGA-based motion sensing and Jetson's deep sleep modes for this specific application remains an area where integrated solutions are less explored in existing reviews. Several VLSI-oriented studies have also addressed low-power strategies, emphasizing techniques like voltage scaling, clock gating, and architectural-level power reduction, which can be foundational to energy-aware embedded systems design [35]. **Jetson Power States:** NVIDIA Jetson platforms provide various power states to optimize energy consumption. The most power-efficient state is Deep Sleep (SC7), where core power rails are turned off while essential components like RTC and DRAM remain powered to enable rapid wake-up.[57] In SC7 mode, Jetson AGX Orin can consume as little as ~300mW (CVM power) or ~1.1W (total CVM+CVB power).[57] Deep sleep can be initiated programmatically using Linux commands like sudo systemctl suspend.[58]

Motion-Triggered Wake-up: The system's ability to enter a low-power state during inactivity and wake up upon detecting motion is crucial for energy conservation. Common wake sources on Jetson platforms include power button presses, RTC alarms, USB hotplug, and Wake on LAN. For motion-triggered wake-up, external interrupts via GPIO pins are a common mechanism. Ultra-low power accelerometers (e.g., ADXL362) can function as motion-activated power switches, consuming very little current (e.g., <300nA in sleep mode) while continuously monitoring for motion. Upon detecting motion, such a sensor can generate a signal (e.g., AWAKE bit mapped to an output pin) that can be used to trigger the Jetson's wake-up.

Reviews on low-power embedded AI systems[4] discuss general power optimization techniques like model

pruning, quantization, knowledge distillation, and dynamic voltage and frequency scaling (DVFS). They also cover power-gating and clock-gating to reduce power consumption in unused components. However, these reviews typically discuss power management in a general sense for embedded AI. They do not often delve into the specific integration of an FPGA as a dedicated, ultra-low power motion-sensing guardian that precisely controls the wake-up and sleep cycles of a more powerful Jetson AGX for an attendance system. The detailed interplay between FPGA-based motion sensing and Jetson's deep sleep modes for this specific application remains an area where integrated solutions are less explored in existing reviews.

2.4. Database Integration in Real-Time Attendance Systems

For a real-time attendance system, efficient database management and synchronization are paramount. Edge computing plays a crucial role by shifting data processing closer to the source of data generation, offering faster insights, reduced response times, and optimized bandwidth utilization.

Edge databases are designed to operate effectively on devices with limited resources and support offline operations, ensuring data availability even when disconnected from central servers. SQLite is a widely deployed embedded database known for its small footprint and robust SQL capabilities, making it ideal for edge deployments. Other options include document-oriented databases like CouchDB/PouchDB with sophisticated synchronization mechanisms.

While local processing is beneficial, synchronization with a central database is still required for comprehensive record-keeping and reporting. Edge databases typically implement sophisticated data synchronization mechanisms to ensure data consistency across distributed systems, including conflict resolution and intelligent data prioritization. The goal is to maintain a complete audit trail of all changes, allowing any system state to be reconstructed from event history.

Existing research on real-time attendance systems often mentions database integration and real-time updates [3]. However, specific challenges related to handling the high throughput of data generated by a multi-face detection system in crowded environments, and the seamless synchronization of this high-volume, real-time data from the edge to a central database, are not always the primary focus of these papers [35]. Reviews on embedded databases [36] discuss their general capabilities but do not specifically address the unique demands of a hybrid FPGA-Jetson system processing numerous faces simultaneously. A recent study by Ray [37] presents a system that enhances attendance monitoring by integrating face recognition with geolocation tagging and real-time action logging, offering an additional layer of accountability and system responsiveness.

3. Identified Research Gaps and the Proposed Hybrid FPGA-Jetson AGX System

Based on the comprehensive review of existing literature, while significant advancements have been made in individual components of real-time face recognition and embedded systems, several critical research gaps remain, particularly when considering the stringent requirements of a real-time, low-power attendance system operating in crowded environments. This section identifies these gaps and positions the proposed hybrid FPGA-Jetson AGX system as a novel solution.

3.1. Gap 1: Lack of Comprehensive End-to-End Hybrid Optimization for Crowded, Multi-Face Real-Time Attendance

Existing hybrid systems (CPU-GPU-FPGA) demonstrate the potential of heterogeneous computing by partitioning tasks across different accelerators [1]. However, many of these studies focus on general computer vision tasks (e.g., SIFT, general CNN inference) or specific object detection scenarios (e.g., small-target remote sensing). While some research explores multi-face detection on Jetson platforms, it often focuses on optimizing the GPU's internal engines or integrating a face tracker, without explicitly detailing the role of an FPGA for initial image pre-processing and resizing to offload the Jetson in a hybrid pipeline specifically for attendance in crowded environments. The challenge of accurately detecting and recognizing many faces in a single frame in real-time, under varying conditions (occlusion, pose, lighting), with a fully optimized end-to-end hybrid architecture, is not comprehensively addressed in existing literature.

Recent efforts such as E Crowd Vision have highlighted the feasibility of combining face recognition and crowd counting in dynamic scenes, demonstrating that hybrid systems can enable scalable and responsive performance in real-time environments [38]. However, even these systems do not fully integrate FPGA-level front-end optimization or detailed energy-aware hardware co-design strategies required for ultra-efficient attendance systems. There is a need for a system that meticulously balances the computational load between the FPGA (for efficient, low-latency front-end processing) and the Jetson (for complex deep learning inference) to achieve optimal performance for high-density face detection and recognition in attendance scenarios.

3.2. Gap 2: Limited Integration of Advanced Low-Power Management with Hybrid Wake-up Mechanisms

While Jetson platforms offer deep sleep modes for power saving and motion sensors can trigger wake-ups , the existing reviews on low-power embedded AI systems generally discuss these concepts in isolation or in less integrated contexts. There is a significant gap in research that explicitly details and evaluates an intelligent power management scheme where a

low-power FPGA acts as a dedicated guardian, continuously monitoring motion via a sensor, and then precisely controlling the *wake-up and sleep cycles of the more power-hungry Jetson AGX*. This specific hybrid wake-up mechanism, designed to maximize energy savings during inactivity while ensuring

immediate responsiveness upon detecting an object, is not a common or thoroughly explored integrated solution in the context of real-time attendance systems. The reliability and latency of such a hybrid wake-up system, especially considering potential issues with Jetson's wake-up behavior, represent a clear area for novel contribution.

3.3. Gap 3: Insufficient Focus on Seamless Real-Time Database Integration for High-Throughput Crowded Scene Data

Existing attendance systems discuss real-time database updates and the benefits of edge computing for local data processing. However, the specific challenges of handling the
volume, velocity, and variety of data generated by a high-throughput, multi-face detection and recognition system in crowded environments are not extensively addressed in the context of database integration. The seamless, low-latency update of attendance records for "many faces in a frame" [User Query], coupled with robust edge-to-cloud synchronization mechanisms that can manage potential data conflicts and ensure consistency, remains an area requiring more dedicated research. While edge databases are discussed, their optimization for the specific demands of a hybrid FPGA-Jetson system processing numerous simultaneous identifications and updating a real-time attendance database is a less explored domain.

3.4. The Proposed Hybrid FPGA-Jetson AGX System: Addressing the Gaps

The proposed research directly addresses the aforementioned gaps by developing a real-time face detection and recognition attendance system that leverages a hybrid FPGA-Jetson AGX architecture with intelligent low-power capabilities.

- **Addressing Gap 1 (Comprehensive End-to-End Hybrid Optimization):**
    - **FPGA's Role:** The FPGA is responsible for identifying the real-time camera image, resizing it, and sending it to the Jetson via Ethernet [User Query]. This offloads computationally intensive, low-level image pre-processing tasks (e.g., acquisition, resizing, quality enhancement) from the Jetson, ensuring that the Jetson receives optimized data streams. This strategic task partitioning is crucial for maintaining real-time performance in crowded environments with many faces.
    - **Jetson AGX's Role:** The Jetson then crops faces from the resized images and identifies individuals (e.g., Amal, Kamal, Vimal, Jorn) [User Query]. It utilizes its powerful GPU, Tensor Cores, and DLAs for high-performance deep learning inference, specifically optimized for crowded areas with many faces in the frame. This includes leveraging multi-face detection algorithms and integrating a face tracker module to enhance throughput and accuracy by avoiding redundant recognition calls.
    - **Communication:** High-speed Ethernet communication ensures rapid data transfer between the

FPGA and Jetson, preventing bottlenecks [39], [40]. Lokhande and Ingole [41] demonstrated that FPGA-based implementation of Gigabit Ethernet offers high-speed, low-latency data transfer suitable for vision systems and high-bandwidth embedded applications. These FPGA-based transmission systems, as demonstrated by Wang et al. [42], can support efficient, reliable image data streaming, which is crucial for sustaining real-time face recognition in high-density attendance environments.

- **Addressing Gap 2 (Integrated Low-Power Management with Hybrid Wake-up):**
  - The project is designed as a low-power system where the Jetson goes to deep sleep mode to save power if no object is detected for one minute [User Query].
  - A motion sensor connected to the FPGA detects motion and sends a signal to the FPGA, which then wakes the Jetson to start capturing images [User Query]. If there is inactivity again, the FPGA stops capturing images, and the Jetson returns to deep sleep [User Query]. This intelligent power management scheme, with the FPGA acting as a low-power, always-on guardian, is a key novel contribution to energy efficiency.
- **Addressing Gap 3 (Seamless Real-Time Database Integration):**
  - The system updates attendance in real-time in the database as faces are identified [User Query]. This requires efficient edge database management and synchronization mechanisms to handle the high volume of data generated by continuous, multi-face detection in crowded areas. The proposed system aims to demonstrate robust real-time updates and data consistency, even with the demanding data rates from crowded scenes.

By meticulously partitioning tasks and leveraging the complementary strengths of each platform, the proposed hybrid FPGA-Jetson AGX system aims to achieve a challenging balance of real-time performance, high accuracy in crowded and occluded environments, and stringent low-power operation, thereby filling critical gaps in existing research.

4. Conclusion

This review paper has systematically examined the current landscape of real-time face detection and recognition systems for attendance, drawing insights from a wide array of academic literature and existing reviews. We have highlighted the significant progress made in algorithmic development, particularly with deep learning models, and the increasing adoption of embedded hardware platforms like FPGAs and Jetson devices for edge AI deployment.

Despite these advancements, our analysis reveals distinct research gaps that impede the development of truly comprehensive and robust attendance solutions for complex real-world scenarios. Specifically, there is

a notable absence of:

1. **Comprehensive end-to-end hybrid optimization** fully integrates FPGA-based image pre-processing with Jetson-based deep learning inference, specifically tailored for high-throughput, multi-face detection and recognition in crowded attendance environments.
2. **Integrated and intelligent low-power management** strategies that leverage a low-power FPGA as a dedicated motion-sensing guardian to precisely control the deep sleep and wake-up cycles of a more powerful Jetson AGX.
3. **Sufficient focus on seamless real-time database integration** capable of handling the high volume and velocity of data generated by multi-face detection in crowded scenes, ensuring immediate updates and robust synchronization.

The proposed research, a hybrid FPGA-Jetson AGX system, directly addresses these identified gaps. By assigning the FPGA to real-time image acquisition, resizing, and motion-triggered power management, and the Jetson AGX to advanced deep learning inference for multi-face detection and recognition, coupled with real-time database updates, this system offers a novel and holistic solution. This integrated approach not only promises to enhance real-time performance and accuracy in challenging crowded environments but also ensures significant energy efficiency, paving the way for next-generation attendance management systems that are both powerful and sustainable. Future work will involve the detailed implementation and empirical validation of this hybrid architecture to demonstrate its full potential in real-world deployments.

REFERENCES

1. J. Zhang, S. Xiong, C. Liu, Y. Geng, W. Xiong, S. Cheng, and F. Hu, “FPGA-based feature extraction and tracking accelerator for real-time visual SLAM,” Sensors, vol. 23, no. 9, pp. 1–21, 2023.
2. K. B. Dharmaraj, C. N. Dhanushree, H. V. Revanth, N. Shashank, and P. M. Tejaswini, “Real-Time Student Face Recognition Attendance System Using AI,” Int. Adv. Res. J. Sci. Eng. Technol., vol. 12, no. 5, pp. 1–2, May 2025, doi: 10.17148/IARJSET.2025.125183.
3. J. Cole, “Heterogeneous Computing for AI: Integrating CPUs, GPUs, and FPGAs,” ResearchGate, Sep. 2022.
4. N. Tibbetts, S. Ibtisum, and S. Puri, “A survey on heterogeneous computing using SmartNICs and emerging data processing units (expanded preprint),” Missouri Univ. of Sci. and Technol., Rolla, MO, 3 Mar. 2025.
5. A. Mishra and A. Shrivastava, “Low power design techniques: A review,” in Proc. Int. Symp. Electron. Syst. Design (ISED), 2013, pp. 1–6.
6. S. Sharma and A. Kumar, “Optimized deep learning model for real-time object detection in

surveillance applications," Appl. Sci., vol. 15, no. 4, Art. no. 422, 2025.

7. S. Tariyal, R. Chauhan, Y. Bijalwan, R. Rawat, and R. Gupta, "A comparative study of MTCNN, Viola Jones, SSD and YOLO face detection algorithms," in Proc. IEEE Int. Conf. Innovative Trends in Computer and Engineering Applications (IITCEE), Jan. 2024. doi: 10.1010/iitese59897.2024.20447445.
8. Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," Sensors, vol. 20, no. 2, pp. 342, Jan. 2020.
9. C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.
10. K. M. Ponnmoli and A. Pandian, "Comparative evaluation of face detection algorithms: Accuracy, efficiency, and robustness," Indica Journal, vol. 6, no. 3, pp. 1–8, 2025.
11. A. N. Anjeana and K. Anusudha, "Survey on various face detection methods," Int. J. Creat. Res. Thoughts, vol. 11, no. 3, Mar. 2023.
12. F. Tolu, C. Giuffrida, and D. Giordano, "Deep learning techniques for face recognition: A systematic review," J. Imaging, vol. 11, no. 1, 2023, Art. 58.
13. T. G. Aishwarya, V. N., and P. Asritha, "Face Recognition Attendance Management System," Int. J. Res. Publ. Rev., vol. 6, no. 5, pp. 11049–11052, May 2025.
14. D. Ray, "A Face Recognition Based Attendance System with Geolocation and Real Time Action Logging," unpublished.
15. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., June 2015, pp. 815–823.
16. A. Baobaid and M. Meribout, "Edge GPU based face tracking for face detection and recognition acceleration," arXiv:2505.04524, May 2025.
17. N. A. Andriyanov and V. E. Dementiev, "Optimization of face recognition systems for implementation in embedded systems," Pattern Recognit. Image Anal., vol. 34, no. 4, pp. 1245–1254, Apr. 2025.
18. T. Ali, D. Bhowmik, and R. Nicol, "Energy aware computer vision algorithm deployment on heterogeneous architectures," Discover Electronics, vol. 2, no. 1, Jun. 2025. doi: 10.1007/s44291-025-00078-7.
19. M. Eid et al., "A novel lightweight CNN model for efficient COVID-19 detection using chest X-ray images," Electronics, vol. 12, no. 882, Mar. 2023.
20. J. Jiang et al., "FPGA based acceleration for convolutional neural networks: A comprehensive review," unpublished.

21. J. Xiao et al., “Adaptive interpolation algorithm for real time image resizing,” in Proc. IEEE Int. Conf. Innovative Comput., Inf. and Control (ICICIC), vol. 2, 2006, pp. 221–224.
22. A. Baobaid and M. Meribout, “Edge-GPU based face tracking for face detection and recognition acceleration,” arXiv:2505.04524v1, May 2024.
23. Debadrita Ghosh, "Real-time attendance system using face recognition technique," Int. J. Eng. Appl. Sci. Technol., vol. 5, no. 9, pp. 258–261, Jan. 2021.
24. A. Baobaid and M. Meribout, “Edge-GPU based face tracking for face detection and recognition acceleration,” arXiv:2505.04524v1, May 2025.
25. S. Alha and I. Khan, “A comprehensive review of face detection using deep learning techniques,” Int. J. Appl. Res., vol. 11, no. 6, pp. 270–275, Apr. 2025.
26. K. Moens and A. Kornet, “Can facial recognition be reconfigured for privacy? A survey,” arXiv preprint arXiv:2505.04524, 2024.
27. M. Manogaran, R. Varatharajan, and V. Chang, “A survey on data fusion and mining in Internet of Things,” Future Internet, vol. 17, no. 175, pp. 1–17, Jun. 2025. DOI: 10.3390/fi17070175.
28. A. Mishra, “Power and performance trade-offs in real-time embedded systems,” in Proc. Int. Symp. on Electronic Design (ISED), Dec. 2013, doi: 10.1109/ISED.2013.6827844.
29. R. S. Thakur, U. Choubey, S. Tripathi, V. Rajput, S. Pagare, and D. Chauhan, “E Crowd Vision: Real-time face recognition and crowd counting in dynamic environments,” Int. J. Eng. Appl. Sci. Manag., vol. 6, no. 1, Jan. 2025.
30. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, 2016, 1499–1503.
31. V. Shankar, “Design and evaluation of AI driven embedded systems for high performance, low power applications,” Int. J. Adv. Res. Sci. Commun. Technol., vol. 5, no. 4, 331–338, Apr. 2025.
32. C. Stentoumis, D. Tzovaras, and A. Daras, “Face recognition using depth-assisted training,” Electronics, vol. 14, no. 7, 2023, Article 1736.
33. A. Baobaid and M. Meribout, “Edge-GPU based face tracking for face detection and recognition acceleration,” unpublished.
34. A. Jurado, A. Pomares-Garcia, A. Ruiz-Gonzalez, M. D. R-Moreno, and J. A. Gómez-Pulido, “Real-time edge AI: Implementation of YOLOv5s object detection on the Jetson Nano,” Remote Sens., vol. 17, no. 3, Art. no. 494, 2025.
35. Z. Cheng, J. Gu, H. Fan, H. Cai, B. Wu, J. Lin, and Y. Tian, “Fast and accurate face detection with neural architecture search,” arXiv preprint arXiv:2307.16834, Jul. 2023.
36. M. Li, Y. Li, B. Sun, and W. Jin, “A multimodal fusion model based on infrared and visible images

for face recognition," Sensors, vol. 25, no. 8, Art. no. 3126, 2025.

37. A. Ramnath and M. M. Dhanabalachandran, "Face recognition methods: A comprehensive review based on bibliometric analysis," Mathematics, vol. 13, no. 13, 2023, Article 2095.
38. Pedram Ghazi, Antti P. Happonen, Jani Boutellier, and Heikki Huttunen, "Embedded implementation of a deep learning smile detector," in Proc. European Workshop on Visual Information Processing (EUVIP), 2018, 1–6.
39. Anjali S. S., Rejani Krishna P., and Aparna Devi P. S., "High speed data transfer using FPGA," Int. J. Eng. Res. Gen. Sci., vol. 4, no. 3, pp. 543–549, May–June 2016.
40. Hao Wang, Zhi Weng, and Xiaochun Li, "High-speed image acquisition and network transmission system based on FPGA," in Proc. Int. Conf. Elect., Mech. and Ind. Eng., Apr. 2016, pp. 72–75.
41. Hong Zhao, Xi-Jun Liang, and Peng Yang, "Research on face recognition based on embedded system," Math. Probl. Eng., vol. 2013, Art. ID 519074, 2013.
42. V.R. Gad, R.S. Gad, and G.M. Naik, "Implementation of Gigabit Ethernet Standard using FPGA," Int. J. Mobile Netw. Commun. Telematics, vol. 2, no. 4, pp. 31–44, Aug. 2012.