International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Design and Simulation of a Lightweight Edge–Cloud Face Recognition System for Smart Attendance Management

H Krupa[1], G Sudha Gowd[2], M Vidhyanand Reddy[3], S Arshad Ali[4], G Rajesh[5], V Thanvi[6]

[1]*PVKK Institute of Technology Anantapuramu, Andrapradesh*
[2]*PVKK Institute of Technology Anantapuramu, Andrapradesh*
[3]*PVKK Institute of Technology Anantapuramu, Andrapradesh*
[4]*PVKK Institute of Technology Anantapuramu, Andrapradesh*
[5]*PVKK Institute of Technology Anantapuramu, Andrapradesh*
[6]*PVKK Institute of Technology Anantapuramu, Andrapradesh*

Abstract - Attendance management remains a significant challenge in educational institutions due to proxy entries, human errors, and the absence of real-time monitoring. Conventional roll-call methods are time-consuming and unreliable, particularly in large classrooms. This paper presents a Hybrid Edge-Cloud Face Recognition-Based Smart Attendance Monitoring System implemented as a prototype and evaluated in a simulated deployment environment. The proposed framework integrates lightweight Convolutional Neural Networks (CNNs) for face detection and recognition within a virtualized edge computing setup. Model optimization techniques, including structured pruning and INT8 quantization, are applied to reduce computational complexity and enable deployment on resource-constrained edge devices. To minimize bandwidth usage between edge and cloud components, facial embeddings are compressed from 512 to 128 dimensions. A hybrid loss function combining ArcFace and Center Loss is employed during training to enhance robustness against variations in lighting, pose, and multi-face classroom conditions. The architecture distributes processing between simulated edge nodes and a centralized cloud server to evaluate latency, bandwidth consumption, and scalability. An adaptive multi-frame attendance validation mechanism further reduces false positives. Experimental results demonstrate improved efficiency and reduced computational and network overhead without compromising recognition accuracy.

Key Words: Smart Attendance System, Face recognition, Convolutional Neural Network, Model Quantization, Edge cloud Architecture, Automated Attendance Management

1. INTRODUCTION

Maintaining the attendance regularly in the schools is highly important because in the institutions where the government allows grants, faculty benefits, and other performance reviews to be made, attendance must occur regularly [1]. Universities and colleges have a plethora of academic and financial incentives for students and faculty members, yet in most instances regular attendance is a prerequisite before the benefits can be enjoyed. The attendance monitoring is thus obligatory to the delivery of equitable advantages, educational advancement and administration judgement [2]. Register systems have always been used to keep the attendance manual. Such systems will most probably contain proxy entries, data alterations and human error [3]. Paper records require a lot of time to keep even during the working days, and long-term storage of paper records might be challenging to manage in the large institutions [4]. Moreover, the attendance records are usually linked with the processing of payroll, grading systems and performance evaluation, and thus the quality of the data is a factor. The administrative people may fail to identify the irregular attendance pattern and take the right steps due to the lack of real-time implementation and automated analysis [1][3][5]

To overcome the drawback of the manual systems, attendance systems based on biometrics, such as fingerprint recognition, have been introduced. The fingerprint systems are, however, prone to failure due to wet or dry or damaged fingers; physical contact is also required, and physical hardware must also be installed and maintained at any given moment [6]. Environmental factors and sensor failures may also reduce system reliability. The face recognition attendance system is an alternative to the contactless system. Nevertheless, traditional face recognition

algorithms are vulnerable to light variations, face position and camera distance [7]. The deterioration of performance may also be achieved in overcrowded and traditional approaches are not always resistant to dynamism and the real-world scenario [9]. Therefore, the appropriate time is to have a more efficient, more accurate and more scalable attendance management system that can be implemented in the varying environmental conditions and with little dependence on hardware and little cost of operation [2][10].

In our study, we have developed a deep learning face recognition smart attendance system. In place of NFC-authentication, we have employed a Convolutional Neural Network (CNN) model to do face recognition[11]. The system uses CNN based embedding model in the process of extracting facial features. To raise the efficiency, embedding size is dropped to 128 dimensions as opposed to 512 dimensions[13]. In addition, the INT8 quantization serves to reduce the complexity of computation and accelerate the inference. All the methods are integrated to further enhance the accuracy and added functionality on the attendance management [12]. To address the drawbacks of the traditional systems, we developed an edge-cloud architecture, an architecture that integrates INT8 quantization and facial detection modules into one. The planned architecture to be implemented will manage the attendance of all employees in an expandable and structured manner[14].

The paper has the following structure. Section II is a system overview and gives a discussion on the basics of CNN and proposed network architecture diagram. It also outlines the system requirements and the functionality of each module. Part III will consider the traditional attendance systems and observe and experimental studies done on them and its difficulty. Section IV presents the manner in which the proposed better system will be developed and deployed along with the functionality of each of the modules and the algorithms that will be used. Section V will have the experiment results and discuss the performance of the full face recognition based attendance architecture. Finally, the paper includes the conclusion and future research directions and system improvements in the form of Section VI.

1. RELATED WORK

As I noticed in my assignment in writing about different approaches of various research papers we should give the relation that each of the modules is underway and what are the limitations. The attendance management system which involves the use of the live video streaming was developed in [1] to automate the attendance of students. The OpenCV system was used to extract video frames and dlib library was used to perform face detection and recognition. The student database identified faces and matched them in the stored database of students to update on attendance automatically. Though it reduced the number of proxies attending and error in manual counting, it depended on classical face recognition libraries and was not scalable, embedding optimization, and performance on a large scale.

The proposal on face recognition-based attendance monitoring system was one of the suggestions in [2], which was aimed at supplementing the traditional attendance processes in learning institutions. The system built facial database and compared the captured images with the stored in the attendance sessions. The known data were entered on an excel sheet and sent to the faculty members. Even though the system increased automation and reduced ambiguity, it was still incapable of strong deep learning optimization, model compression plans, and performance evaluation in various environmental conditions. In [3], a smart attendance system, which was built on a CNN, the high-definition cameras were utilized to take the pictures of the students and then compared them with the already existing datasets to update a centralized database. The system improved the precision of identification and minimized the human factor through convolutional neural networks. However, the study failed to explore the model design of lightweight models and quantization, embedding dimensionality reduction or scalable architecture concerns to be executed in real time.

In [4] the 6th International Conference on Cyber and IT Service Management proposed in 2018, a mobile-based attendance system that uses a combination of NFC and face authorization was introduced. The system consisted of Raspberry Pi and the cloud storage to manage the attendance data. Although the NFC integration provided an improved security level to the authentication process, the system was very reliant on the hardware infrastructural support as well as a lack of the advanced

machine learning models and optimization algorithms that would offer fruitful face recognition. In [5], which was published in the 2024 6th International Conference on Cybernetics and Intelligent System an attendance system is proposed that relies on the HAAR Cascade Classifier face detector. The system was determined to work well under controlled situations of light and fixed capture distance. Still, it was susceptible to lighting variations and pose change and not deep learning based embeddings and computational optimization methods, as it relied on traditional computer vision methods.

2. PROPOSED WORK

According to the pertinent work presented in [1] 5 there are a number of limitations, which are not addressed by the current advances in the automation of attendance management based on face recognition and other technologies. Most of the current systems are premised on traditional approaches to computer vision or are simply concerned with integrating hardware without far attention to lightweight deep learning optimization. These issues have not been studied in detail, such as the dimensionality reduction embedding, computational performance, scalability to real-time, and the capacity to execute with broad environmental variations is not detailed extensively.[2][15] Moreover, less has been done to explore problems with model quantization in an effort to reduce the overall cost of inference, and support deployment to large-scale institutional settings. Accordingly, the optimized the software-based deep learning-enabled attendance system with built-in lightweight CNN models, compression, INT8 quantization, and scalable edge -cloud implementation is needed to ensure a certain increase in the accuracy, efficiency, and adaptability levels[3][7].

It is possible to develop the proposed system based on the challenges that were identified in the previous experimental studies. The architecture can be further broken down into three major sections and they include: CNN-based face recognition model with embedding optimization, INT8 quantization and hybrid edge cloud architecture.[16]. All the modules are aimed at upgrading the precision of recognition, dimension, and scalability of the whole attendance administration system. This architecture will primarily be designed to reduce proxies attendance, maximize recognition performance, decrease the computational cost and real time processing of attendance which is also provided as support [17]. Hardware dependence and cost of running will also be minimized by the system. Face detection and recognition module captures the facial image and identifies the part of face using real time inputs [18]. CNN model carries out the extraction of features on the identified pictures of the faces, and it generates embedding vectors of unique characteristics of the face. The embedding dimensionality is now reduced to 128 dimensions so as to increase efficiency. This reduction saves on the use of memory and in addition, the rate of comparison and speed of processing is also enhanced [1][2][3].

INT8 quantization is further applied to the model so as to perform even better. The floating-point weights are converted to integer values through quantization and this can be exploited to decrease the size and computation bicycle of the model[19]. This promotes the inference rate and the accuracy of recognition is acceptable. This discrete model is lightweight therefore it can be used in implementation of an edge computing [12]. The proposed hybrid 2 edge-cloud design would have face detection and recognition carried out at edge layer and an attendance record in cloud layer and the centralized database and reporting module controlled at the cloud layer[20]. With this distribution, there is less centralized workload processing and is more scalable. Overall, the proposed architecture can become a successful, accurate, and scalable automated attendance control system.

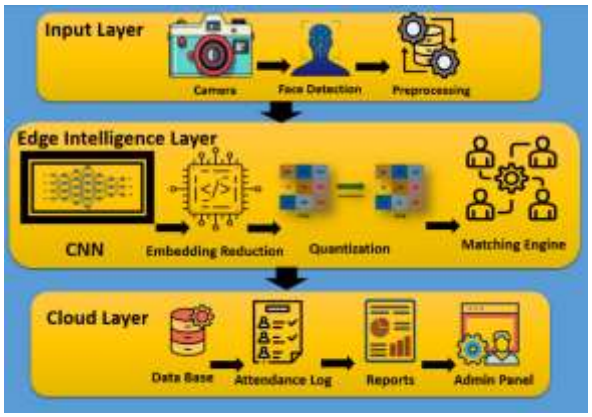

Fig.1 Proposed architecture

The proposed architecture consists of three levels of the input acquisition, edge intelligence processing and the cloud management[6]. The edge layer will be based on face-detecting, CNN-based feature-extracting, embedding optimization, and INT8 quantized inference. Attendance, report and analytics are stored in the cloud layer. It is a hybrid design that ensures scalability, less complicated computational load, and effective real-time attendance control.[1][4][5].

According to the suggested fig 1 smart attendance system is a scalable architecture that consists of the deep learning face recognism and a hybrid edge-cloud architecture [7][15]. The system aims at helping in the proper identification process and reducing the complexity of the computation and offering a real time performance. This overall design encompasses modules of input acquisition, edge-level intelligence processing modules and cloud management modules [21][22].

A. Input Acquisition and Face Detection

The system begins with image acquisition subsystem that obtains facial image of a person through a camera interface in a simulation system. The frames captured pass through the face detection module that identifies and removes the facial area in the background. The above measure will be applied to make sure that only relevant facial information is forwarded to the next processing. Recognition that is correctly identified increases the validity of the recognition process and reduces the unnecessary calculation costs.

B. Preprocessing and Feature Extraction

Computerized tomography (CT) scans are significant in the diagnosis of osteoporosis since it can demonstrate the bone mineral density and bone mass.

After recognizing the face, the preprocessing processes of the extracted portion of the face are performed such as resizing, normalization, and alignment. These functions normalize input data and insert it into Convolutional Neural Network (CNN).

The CNN model performs the hierarchical feature extraction by use of convolutional, activation and pooling layers. The network generates embedding of 512-dimensional, which is a representation of distinctive face characteristics. Such embeddings may be treated as numerical comparison of identities.

C. Embedding Optimization

Dimensionality reduction is incorporated in order to make things more efficient. It is a smaller variant of feature V dimension, rather than 512 dimensions, 128 which reduces the memory used, as well as similarity comparison is faster and overall performance of processing is improved without having a severe impact on recognition rate.

D. Model Quantization

To further optimize the system, INT8 quantization is used. The quantization device converts floating point model weights into a 8-bit integer values.[4][10] This reduces model sizes, reduces the size of the memory used and makes inference faster. Each of the quantized models is made lightweight and can be run at the edge level without any significant impact on recognition accuracy.

E. Face Matching and Attendance Decision

The optimized embedding vector is compared to existing embedding vectors in the database by use of similarity measurement technique. When the similarity score exceeds some given limit, the system validates the identity and shows the person as available. The attendance will not be logged in case the match fails, a scenario that will reduce the possibility of the proxy entry[9].

F. Hybrid Edge–Cloud Architecture

This system uses a hybrid edge-cloud system to attain scalability and resource-efficient utilization. The edge layer does face detection, CNN-based feature extraction, optimization of embedding, quantized inference, and matching. The cloud layer has the responsibility of managing the attendance storage, database maintenance, generation of reports and administrative analytics [4] [11].

G. Functional Workflow

General way of operation of the system is:

Input: Live facial image III, Stored embedding database DDD, Similarity threshold TTT

Output: Attendance status (Present / Not Recognized)

Step 1: Get real time and camera image frame III.

Step 2: Face detection of III. Abort in case no face is found.

Step 3: Face is detected and processed:

(a) Resize image

(b) Normalize pixel values

(c) Align face orientation

Step 4: CNN model takes input of preprocessed image of face.

Step 5: Extract 512-dimensional embedding vector E512.

Step 6: Apply embedding dimensionality reduction: Generate optimized embedding E128.

Step 7: Apply INT8 quantized inference for efficient computation.

Step 8: For each stored embedding $E_i \in D$: Compute similarity score S=similarity (E128,Ei)

Step 9: If $S \geq T$: Mark attendance as **Present**. Update cloud database. Generate attendance log.

Step 10: Else: Mark as **Unknown / Not Recognized**.

Step 11: End process.

The design system guarantees that there is accurate attendance administration, scalability and effectiveness and low operation cost of the system in addition to low hardware dependency.

3. IMPLIMENTATION

The smart attendance management system that was proposed and developed in the form of a modular cloud-based system with integration of RESTful backend services, INT8-quantized face recognition model, relational database (hosted on a cloud) and role based web dashboards. The system architecture is divided into four significant layers, i.e., (1) frontend interface layer, (2) backend API layer, (3) AI inference engine and (4) cloud database layer. This three-level approach offered scalability, maintainability and separation of duties in the components of the system. This dataset is designed to support the development of automated face-based attendance systems in academic environments. It contains labeled face images of students captured in classroom-like settings, reflecting real-world challenges such as Varying lighting conditions, Diverse facial expressions and orientations, Multiple students in a single frame, Low-resolution and surveillance-style imagery, The dataset aims to replicate a typical classroom surveillance setup to help researchers and developers build and evaluate systems that can reliably identify students and record attendance automatically.

The AI component is implemented using the convolutional neural network which is trained to produce face embedding and identify the identity. After model training in floating-point precision (FP32), the model was converted to INT8 format with post-training quantization after following the TensorFlow Lite optimization tricks. The process of quantization demonstrated a significant decrease in the size of the model and the computational burnt as well as preserving the quality of representation of features. The quantized models are installed in the platform of the server on the backend, which enables real-time inference by sending captured facial images on the frontend interface.

The backend is realized with the assistance of API based on the REST and comprising image uploading, pre-processing, embedding, similarity matching, and communicating with the database. When an attempt is made by a user to mark attendance, the image that is received is forwarded to the backend server in a secure way. The inference engine develops facial embedding and compares them to those in the stored embedding on the cloud in the MySQL database. In case identity is checked successfully the attendance records are automatically updated with the information about the time. Role based authentication schemes also ensure that administrators, employees and students are the only persons who should access certain authorized features of the system.

The cloud database layer was implemented to ensure the high availability and remote access as well as centralized storage. The database contains user profiles, facial embedding, attendance record, salary, award records and metadata of system. Secure API endpoints, which are backed by authentication tokens and access control policies, are used to carry out data transactions that form data security. It is not based on the special biometric hardware and can be used in the distributed institutional settings.

Prolonged tests on the system to test the performance of the models, backend responsiveness as well as stability of the system. According to the results of the experiments, INT8 quantization reduced the size of the model more than 70 percent compared to the original FP32 one. The time spent on infusing images also dropped to significant extent and real time attendance checks could be implement. The memory consumption and CPU utilization that was reduce also proved that it is more efficient in computation which is better suited to cloud and mobile accessible situations.

The recognition performance of the quantized model was high with a minimal degradation relative to the baseline model. The institutions had the acceptable operation limits of False Acceptance Rate (FAR) and of False Rejection (FRR). The Recall and F1-Score measures established that the system is capable of ensuring reliable identity discrimination when the optimization is done using integers.

End to end system delay: This includes the transmission of images, pre-processing, inference, validation of database and dashboard update and this remained constant even in a normal network. The architecture also displayed scalability since there was little degradation in response time as simultaneous authentication requests could be serve. The web-based dashboard gave a chance to the administrators to monitor the level of attendance, control employees and student records, pay salaries and performance indicators. This observation supports the fact that the integration of INT8 model quantization and cloud-based service structure is an effective, scalable, and secure smart attendance management system. The proposed structure offers the most appropriate balance between the degree of calculations and the accuracy of recognition, and that is why it might be introduced to learning institutions and business organizations that would require automatic storage of attendance.

4. CONCLUSION AND FUTURE WORK

This research proposed the development of a smart attendance control system in the form of a cloud-computed platform, which is built on an INT8-quantized deep learning model of face recognition. The proposed system will consist of the frontend dashboards, RESTful API on the backend, and a MySQL database on the cloud that will provide a hardware-independent and scalable institutional attendance monitoring system. Moreover, edge-computing support would be worth considering to be of lightweight to support partial inference processing in mobile gadgets to reduce the network latency and reliance on clouds.

A multi-factor authentication with the combination of facial recognition and liveness detection and OTP verification or behavioral biometrics can also be implemented in the future as another use to prevent a serious step against the spoofing attack. The process of administrative decision making can also be enhanced by expanding the system to accommodate huge institutional networks as the distributed distributed institutions with automated analytics dashboard and AI-centered attendance prediction models. Also, privacy-sensitive techniques, such as encrypted facial embeddings and federated learning, can be used to improve compliance and trust in data protection and user confidence.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to the institution and department for providing the necessary infrastructure and technical support to carry out this research work. We also thank our Guide and colleagues for their valuable guidance, constructive suggestions, and continuous encouragement throughout the development of this project. Their insights greatly contributed to refining the system design and experimental evaluation. Finally, we acknowledge the support of all individuals who indirectly contributed to the successful completion of this research study.

REFERENCES

1. W. Gazali, J. M. Kho, J. Santoso, and W. Ef-QuantFace: Streamlined Face Recognition with Small Data and Low-Bit Precision, arXiv preprint, Feb. 2024, highlighting efficient quantization for face recognition networks.
2. M. Fikry, “Performance Analysis of Smart Technology with Face Detection using YOLOv3 and InsightFace for Student Attendance Monitoring,” Int. J. Intelligent Syst. Appl. Eng., vol. 12, no. 4, pp. 3490–3497, Jun. 2024.
3. H. Al-Nayyef, “Advancing Attendance: A Facial Recognition System Empowered by Deep Learning Techniques,” J. Al-Qadisiyah Comp. Sci. Math., vol. 16, no. 1, pp. 61–71, Mar. 2024.
4. J. Biju et al., “Enhancing Attendance Management Systems Using Facial Recognition,” Int. J. Eng. Res. Technol., vol. 13, no. 1, Jan. 2024.
5. A. Agustiyar, R. R. Isnanto, and C. E. Widodo, “Face Recognition for Attendance Systems: A Bibliometric Review of Research Trends and Opportunities,” J. Sisfokom, 2024, providing a trends overview of attendance research.
6. M. E. Ali, A. Diwan, and D. Kumar, “Attendance System Optimization through Deep Learning Face Recognition,” Int. J. Computing & Digital Syst., vol. 15, no. 1, pp. 1527–1540, Apr. 2024.
7. A. Sharma et al., “A Review Paper On Attendance Tracking System Using Cloud Computing,” Recent Trends Parallel Comput., vol. 11, no. 02, pp. 30–35, 2024.
8. “Face Recognition based Attendance Management System,” IJRASET, 2025 — discusses automated attendance via facial biometrics.
9. Jan N. Kolf et al., “EFaR 2023: Efficient Face Recognition Competition,” arXiv preprint, Aug. 2023 — covers efficient quantized face recognition techniques.
10. “Face Recognition Smart Attendance System using Deep Transfer Learning,” Procedia Comp. Sci., vol. 192, pp. 4093–4102, 2021 — exploring deep learning for attendance log automation.
11. F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
12. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems (NIPS), 2012.

13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.
14. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
15. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in International Conference on Learning Representations (ICLR), 2016.
16. B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.
17. P. Gysel, J. Pimentel, M. Motamedi, and S. Ghiasi, "Ristretto: A framework for empirical study of resource-efficient inference in CNNs," IEEE Trans. Neural Networks and Learning Systems, vol. 29, no. 11, pp. 5784–5789, 2018.
18. A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, arXiv:1704.04861.
19. M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.
20. J. Deng et al., "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019.
21. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), 2019.
22. M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2016. [Online]. Available: https://www.tensorflow.org.