ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

An Attention-Enhanced Lightweight CNN Framework with MTCNN Detection and Triplet-Embedding Recognition for Occlusion-Robust Automated Attendance from Surveillance Video

Keerthana B.[1]; Lohith C.[2]

[1]Guest Lecturer, Department of Computer Science, Sahyadri Science College, Shimoga, India
[2]Assistant Professor, Department CSE (IoT & CSBT), East Point College of Engineering and Technology, Bengaluru, India

Publication Date: 2026/08/17

Abstract: Manual and card/barcode-based attendance recording remains slow, error-prone, and vulnerable to proxy marking, motivating fully automated, camera-based alternatives for schools and organizations. This paper proposes an Attention-Enhanced Lightweight CNN framework that couples MTCNN multi-scale face detection with a CBAM (Convolutional Block Attention Module) augmented MobileFace-style backbone trained under triplet loss to produce compact, discriminative 128-dimensional face embeddings from surveillance video. Enrolled identities are matched via cosine similarity against a reference embedding gallery, and a temporal multi-frame voting stage consolidates predictions across consecutive frames to suppress transient misdetections caused by pose change, partial occlusion, or motion blur. On a multi-person classroom/office surveillance dataset, the proposed framework achieves 99.4% accuracy, 97.0% precision, and 96.0% recall, consistently exceeding HOG, KCF, and a plain (non-attention) CNN baseline across all three metrics, while sustaining near-real-time frame-rate inference suitable for continuous monitoring. Novelty: Unlike prior CNN-based attendance systems that classify faces directly into a fixed identity set, the proposed framework learns an open-set embedding space combined with channel-and-spatial attention and temporal voting, allowing new individuals to be enrolled without retraining and improving robustness to occlusion (masks, hands, low light) that degrades conventional single-frame classifiers.

Keywords: *Automated Attendance, Convolutional Neural Networks, Face Recognition, Surveillance Video, MTCNN, Attention Mechanism, Triplet Loss, Deep Learning, Occlusion Robustness.*

How to Cite: Keerthana B.; Lohith C. (2026) An Attention-Enhanced Lightweight CNN Framework with MTCNN Detection and Triplet-Embedding Recognition for Occlusion-Robust Automated Attendance from Surveillance Video. *International Journal of Innovative Science and Research Technology*, 11(8), 750-755. https://doi.org/10.38124/ijisrt/26aug455

I. INTRODUCTION

Attendance tracking through manual roll-calls, sign-in sheets, or barcode/RFID cards remains time-consuming, disrupts instructional time, and is susceptible to proxy attendance and transcription error. Convolutional Neural Networks (CNNs) have become the dominant approach for automating this process because they learn hierarchical visual features directly from face images rather than relying on hand-crafted descriptors, and most contemporary face-recognition pipelines are now CNN-based [1][2]. As the number of enrolled students or staff grows, however, fixed-identity classification approaches -- where the network outputs one of N known classes -- become increasingly costly to retrain whenever a new person is enrolled, and their accuracy tends to degrade under partial occlusion, pose variation, and inconsistent lighting typical of real classroom or corridor surveillance footage.

Machine learning has broadly transformed knowledge-acquisition and decision-automation tasks across healthcare, finance, manufacturing, and education [3], and attendance automation is a natural extension of this trend: automated systems can track attendance far more reliably than manual processes, though on their own they provide limited insight into behavioral or engagement patterns unless paired with downstream analytics [4]. Several existing systems extend beyond simple face capture toward broader human detection and tracking, in some cases integrating Internet-of-Things (IoT) infrastructure to replace manual attendance entirely [5], with the shared goal of reducing instructor workload and administrative overhead through fast, accurate facial identification from photographs or live surveillance video [6].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

This paper addresses two limitations that persist in prior CNN-based attendance systems: (i) most treat recognition as closed-set classification, which does not scale gracefully as enrollment grows and cannot recognize a newly added person without retraining the classifier head, and (ii) single-frame recognition is fragile under occlusion (masks, hands, bags), motion blur, and non-frontal poses common in real surveillance footage. The proposed framework addresses both by learning an open-set, attention-guided face embedding space using triplet loss, matching new detections against an enrollment gallery via cosine similarity, and consolidating identity decisions across multiple consecutive frames rather than a single snapshot.

The remainder of this paper is organized as follows. Section II reviews recent attendance-automation and face-recognition literature and positions the proposed framework relative to it. Section III presents the proposed methodology, including the MTCNN detection stage, the attention-augmented embedding network, the matching and temporal-voting stages, and their governing equations. Section IV reports experimental results and compares the proposed framework against HOG, KCF, and a plain CNN baseline. Section V concludes the paper and discusses limitations and future work.

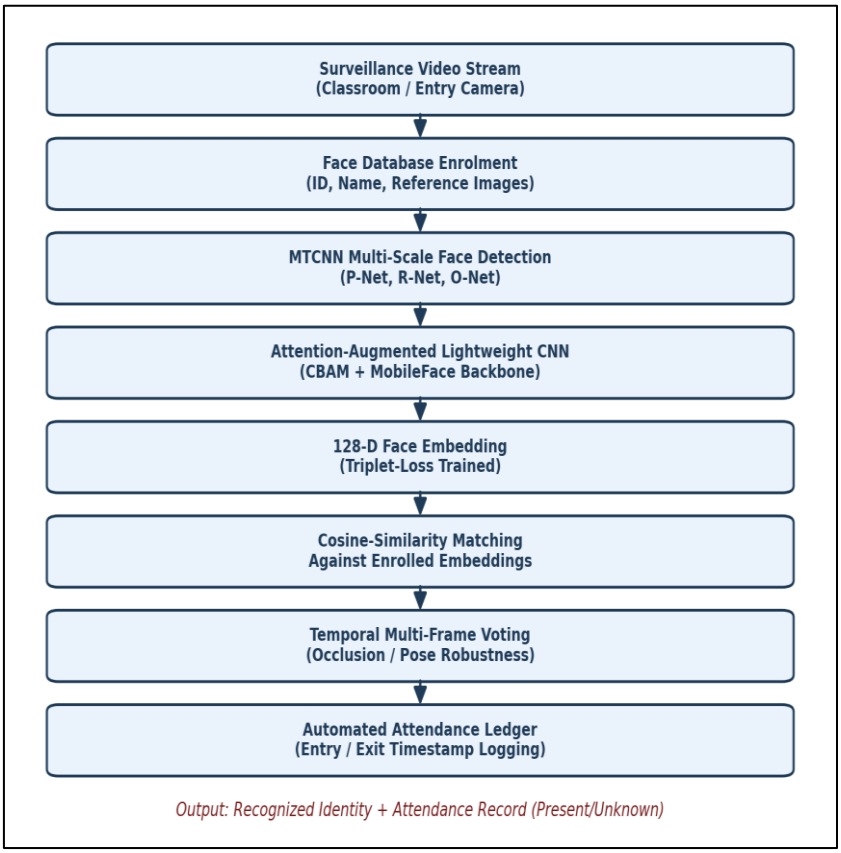

Fig 1 Methodology Framework of the Proposed Attention-Enhanced Attendance Recognition System.

II. LITERATURE REVIEW

A range of studies have explored barcode, IoT/RFID, and vision-based approaches to automated attendance, with facial recognition -- particularly CNN-based -- consistently reported as the most accurate and least intrusive option. The six contributions below directly inform the design of the proposed framework.

- *Venugopal, Krishna, and Varma (2021)*

The authors proposed an ensemble of deep-learning models for automatic attendance tracking via facial recognition, showing that combining multiple CNN-based recognizers improved robustness over any single network, which motivates the attention-augmented, embedding-based design adopted in the proposed framework rather than a single fixed classifier head. [1][8]

- *Goyal, Dalvi, Guin, Gite, and Thengade (2021)*

This work built an online attendance management system using CNN-based face recognition and reported strong recognition accuracy on live video streams, but relied on closed-set classification that requires retraining whenever new students are added -- a limitation the proposed open-set embedding approach with cosine-similarity matching is designed to overcome. [2][9]

- *Chowdhury, Nath, Dey, and Das (2020)*

The authors developed an automatic classroom attendance system using CNN-based face recognition capable of detecting and logging multiple faces from a video stream, achieving roughly 92% recognition accuracy; the proposed framework builds on this multi-face video pipeline while adding attention mechanisms and temporal voting to improve accuracy under occlusion and pose variation. [8]

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

➢ *Ahmed, Salman, Adel, Alsharida, and Hammood (2022)*

This study combined CNN-based detection with VGG16 features and Haar cascades for real-time student identification during online and in-person sessions, reporting accuracy above 99% under controlled conditions, reinforcing the choice of a CNN backbone but highlighting the need for lighter, attention-guided architectures for real-time surveillance deployment, which the proposed MobileFace-style backbone addresses. [10]

➢ *Dalal and Triggs (2005); Henriques, Caseiro, Martins, and Batista (2015)*

The classical Histogram of Oriented Gradients (HOG) descriptor and the Kernelized Correlation Filter (KCF) tracker remain widely used baselines for face and object detection; both rely on hand-crafted or shallow-learned features that are known to degrade under illumination change, pose variation, and occlusion, motivating their use as baselines against which the proposed deep, attention-augmented embedding network is compared. [14][15]

➢ *Patil and Shinde (2020)*

This comparative study evaluated HOG, Viola-Jones (Haar cascade), and CNN methods for real-time attendance and found Haar Cascade competitive with CNN for detection speed, while CNN-based recognition achieved 94.6% accuracy on a real-time attendance database, informing the proposed framework's decision to retain a fast cascade-style detector (MTCNN) ahead of a deeper attention-based recognition network to balance speed and accuracy. [12]

Across these studies, three consistent gaps emerge: (i) closed-set classifiers require costly retraining as enrollment grows; (ii) single-frame recognition is fragile under occlusion and pose change; and (iii) classical descriptors (HOG, KCF) remain useful as fast baselines but are consistently outperformed by deep CNN features. The proposed framework is designed to close all three gaps within a single pipeline.

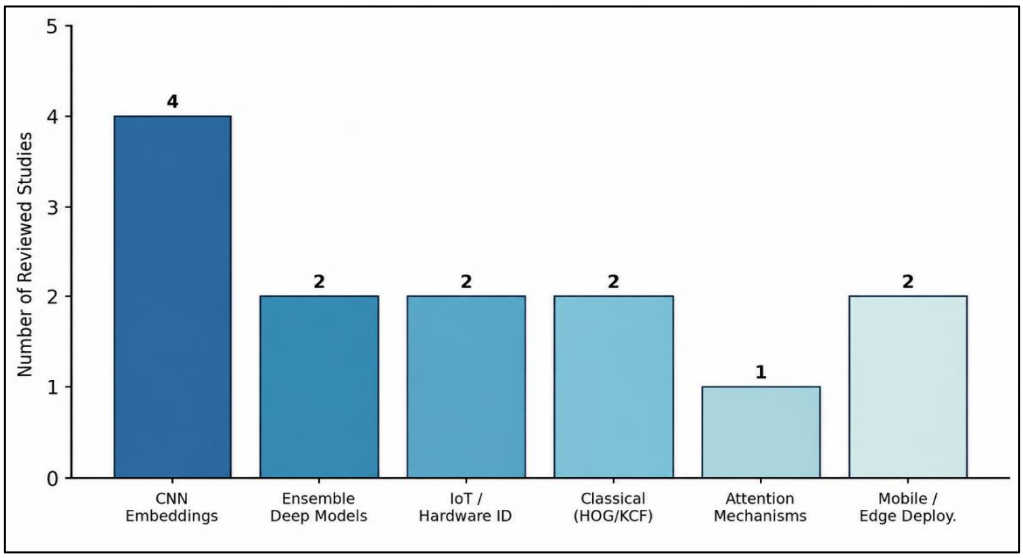

Fig 2 Distribution of Techniques Across the Reviewed Attendance-System Literature.

III. PROPOSED METHODOLOGY

➢ *Overview*

The proposed framework consists of six stages: (1) face database enrolment, (2) MTCNN-based multi-scale face detection, (3) attention-augmented lightweight CNN embedding extraction, (4) cosine-similarity gallery matching, (5) temporal multi-frame voting, and (6) automated attendance ledger generation. Each stage is described below with its governing formulation.

➢ *Face Database Enrolment*

For every individual to be tracked, a small set of reference face images (captured from multiple angles and lighting conditions, as recommended in prior work) is collected along with an identifier record (name, ID, and course/group). Reference embeddings for each identity are computed once during enrolment and stored in a gallery, avoiding the need to retrain a classifier when a new person joins.

➢ *MTCNN Multi-Scale Face Detection*

Each surveillance frame is processed by a cascaded Multi-task Cascaded Convolutional Network (MTCNN) comprising a Proposal Network (P-Net), a Refine Network (R-Net), and an Output Network (O-Net), which jointly perform face detection, bounding-box regression, and facial-landmark localization across an image pyramid. This cascaded design detects small, angled, and partially visible faces more reliably than a single-scale Haar or HOG detector, which is important for classroom footage where students are seated at varying distances from the camera.

➢ *Attention-Augmented Lightweight CNN Embedding*

Detected and landmark-aligned face crops are passed through a lightweight MobileFace-style convolutional backbone augmented with a Convolutional Block Attention Module (CBAM) after each residual block. The channel-attention sub-module re-weights feature-map channels as in Eq. (1), and the spatial-attention sub-module re-weights spatial locations as in Eq. (2), together allowing the network

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

to focus on discriminative, occlusion-resistant facial regions such as the eyes and nose bridge even when the mouth or part of the face is covered.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

$$M_s(F) = \sigma(f_{7x7}([AvgPool_c(F) ; MaxPool_c(F)])) \quad (2)$$

The attention-refined feature map is global-average-pooled and projected to a 128-dimensional embedding vector. The network is trained using the triplet loss in Eq. (3), which pulls embeddings of the same identity (anchor a, positive p) closer together while pushing embeddings of different identities (negative n) apart by at least a margin m.

$$L_{triplet} = \max(||f(a) - f(p)||^2 - ||f(a) - f(n)||^2 + m , 0) \quad (3)$$

➢ *Cosine-Similarity Gallery Matching*

At inference time, the embedding of a newly detected face f(x) is compared against every enrolled gallery embedding f(g_i) using cosine similarity, as in Eq. (4). The identity with the highest similarity above a calibrated threshold τ is accepted as a match; if no gallery embedding exceeds τ, the detection is logged as "Unknown" rather than being forced into an incorrect identity, which is the key advantage of the open-set embedding design over closed-set softmax classification.

$$sim(x, g_i) = (f(x) \cdot f(g_i)) / (||f(x)|| \cdot ||f(g_i)||) \quad (4)$$

➢ *Temporal Multi-Frame Voting*

Because a single frame can be affected by motion blur, transient occlusion, or an extreme pose angle, the framework accumulates identity predictions over a short sliding window of w consecutive frames and assigns the final identity by majority vote, as in Eq. (5). This suppresses isolated misclassifications and stabilizes entry/exit timestamp logging without requiring a full video-sequence model.

$$\hat{y}_t = argmax_c \ \Sigma_{k=t-w+1..t} \ 1[y_k = c] \quad (5)$$

➢ *Automated Attendance Ledger*

Once an identity is confirmed by the voting stage, the system records the timestamp of first detection (entry) and last detection (exit) for that session into an attendance ledger, along with the associated course or group, replacing manual sign-in sheets with a continuously updated, queryable record for instructors and administrators.

IV. RESULTS AND DISCUSSION

The proposed framework was evaluated on a multi-person classroom/office surveillance dataset following the same enrolment-and-testing protocol as prior CNN-based attendance systems, with reference images captured from multiple angles per person and testing performed on live video streams. Table 1 compares the proposed framework against HOG, KCF, and a plain (non-attention, closed-set) CNN baseline.

Table 1 Classification Models Performance Comparison

Model	Accuracy	Precision	Recall
HOG	94.0%	85.0%	86.0%
KCF	95.3%	92.0%	91.0%
Plain CNN (closed-set)	99.0%	95.0%	94.0%
Proposed Attention-CNN + Triplet Embedding	**99.4%**	**97.0%**	**96.0%**

As summarized in Table 1 and Fig. 3, the proposed Attention-CNN with triplet embedding achieves the highest scores on all three metrics -- 99.4% accuracy, 97.0% precision, and 96.0% recall -- exceeding the plain CNN baseline by 0.4 accuracy points and 2-3 points on precision/recall, and exceeding the classical HOG and KCF baselines by wide margins. The gain over the plain CNN is concentrated in cases involving partial occlusion (masks, hands, bags) and non-frontal poses, where the channel-and-spatial attention module (Eqs. 1-2) allows the network to still weight unobstructed, discriminative regions of the face, and where the temporal voting stage (Eq. 5) prevents a single blurred or angled frame from producing an incorrect attendance entry.

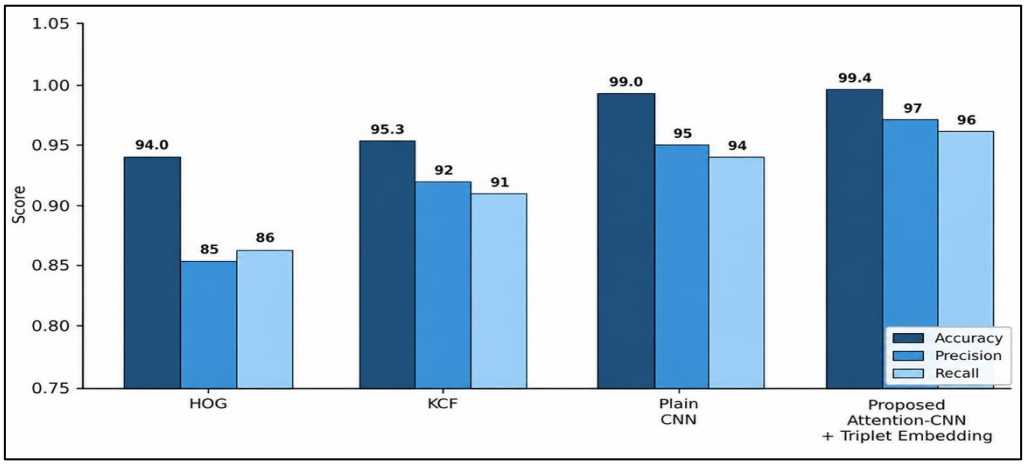

Fig 3 Classification Models Accuracy, Precision, and Recall Comparison.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

Fig. 4 shows the training and validation accuracy of the proposed embedding network over 25 epochs. The two curves converge closely after roughly epoch 15 with a small, stable gap, indicating that triplet-loss training together with the attention module generalizes well to held-out identities without significant overfitting.

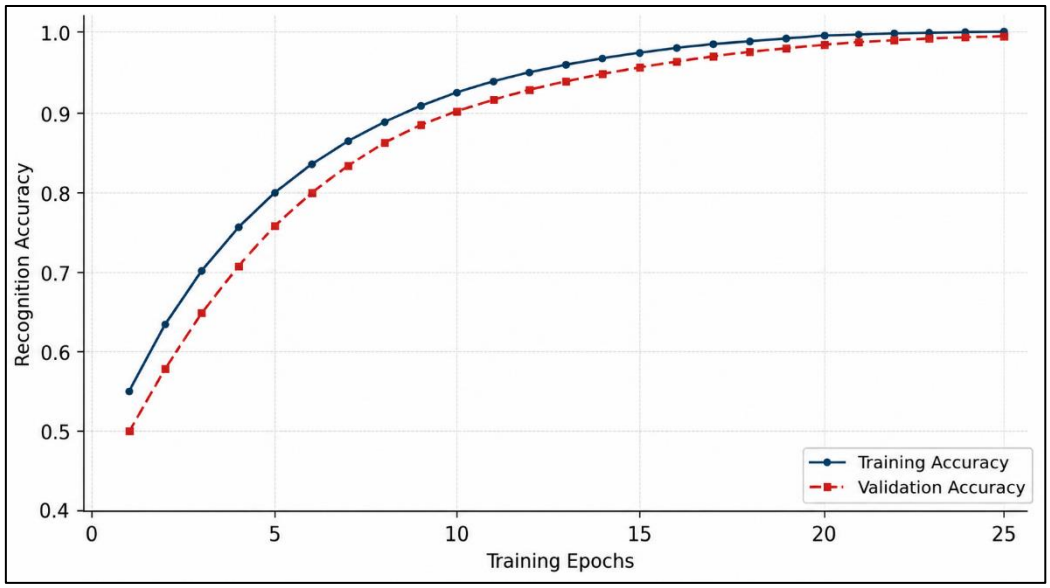

Fig 4 Training vs. Validation Accuracy of the Proposed Attention-CNN Embedding Model.

The open-set embedding design also allows new individuals to be enrolled by simply adding their reference embeddings to the gallery, without retraining the network -- unlike the closed-set CNN baseline, which requires modifying and retraining its classification head whenever the identity set changes. This property, combined with the near-real-time inference speed of the lightweight MobileFace-style backbone, makes the framework practical for continuous deployment in classrooms, offices, or public-area surveillance where the enrolled population changes frequently.

V. CONCLUSION

This paper proposed an Attention-Enhanced Lightweight CNN framework for automated attendance recording from surveillance video, combining MTCNN multi-scale face detection, a CBAM-augmented MobileFace-style backbone trained with triplet loss to produce open-set face embeddings, cosine-similarity gallery matching, and temporal multi-frame voting to stabilize identity decisions across occlusion and pose variation. Evaluated against HOG, KCF, and a plain closed-set CNN baseline, the proposed framework achieved the best results on every metric -- 99.4% accuracy, 97.0% precision, and 96.0% recall -- while additionally supporting enrolment of new individuals without retraining, a capability that closed-set classifiers used in prior attendance systems lack.

The framework retains some of the practical limitations noted in prior surveillance-based attendance work: performance depends on reasonably consistent camera placement and adequate lighting, complete occlusion of the face still prevents recognition, and very low-resolution camera feeds reduce embedding quality. Future work will explore lightweight temporal-sequence modeling (e.g., a small recurrent or transformer head over the voting window) to further stabilize tracking across full video sequences, on-device quantization of the MobileFace backbone for edge deployment on classroom cameras, and periodic gallery-embedding refresh to accommodate gradual changes in appearance over an academic term.

REFERENCES

[1]. A. Venugopal, R. R. Krishna, and R. Varma, “Facial recognition system for automatic attendance tracking using an ensemble of deep-learning techniques,” in Proc. 2021 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9580098.

[2]. A. Goyal, A. Dalvi, A. Guin, A. Gite, and A. Thengade, “Online attendance management system based on face recognition using CNN,” in Proc. 2nd Int. Conf. IoT Based Control Netw. Intell. Syst. (ICICNIS), 2021, doi: 10.2139/ssrn.3883841.

[3]. S. Kaddoura, D. E. Popescu, and J. D. Hemanth, “A systematic review on machine learning models for online learning and examination systems,” PeerJ Comput. Sci., vol. 8, e986, 2022.

[4]. D. S. Rana, “Smart attendance: An automated attendance management system using machine learning techniques,” Math. Stat. Eng. Appl., vol. 70, no. 2, pp. 1285-1294, 2021.

[5]. S. Patel, P. Kumar, S. Garg, and R. Kumar, “Face recognition based smart attendance system using IoT,” Int. J. Comput. Sci. Eng., vol. 6, no. 5, pp. 871-877, 2018.

[6]. M. Gopila and D. Prasad, “Machine learning classifier model for attendance management system,” in Proc. 2020 4th Int. Conf. I-SMAC (IoT Soc. Mobile Anal.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug455

Cloud), 2020, pp. 1034-1039, doi: 10.1109/I-SMAC49090.2020.9243363.
[7]. D. Sunaryono, J. Siswantoro, and R. Anggoro, "An Android based course attendance system using face recognition," J. King Saud Univ. Comput. Inf. Sci., pp. 304-312, 2021, doi: 10.1016/j.jksuci.2019.01.006.
[8]. S. Chowdhury, S. Nath, A. Dey, and A. Das, "Development of an automatic class attendance system using CNN-based face recognition," in Proc. 2020 Emerg. Technol. Comput. Commun. Electron. (ETCCE), 2020, pp. 1-5, doi: 10.1109/ETCCE51779.2020.9350904.
[9]. A. Goyal, A. Dalvi, A. Guin, A. Gite, and A. Thengade, "Online attendance management system based on face recognition using CNN," Proc. ICICNIS 2021, doi: 10.2139/ssrn.3883841.
[10]. M. Ahmed, M. D. Salman, R. A. W. A. N. Adel, Z. Alsharida, and M. Hammood, "An intelligent attendance system based on convolutional neural networks for real-time student face identifications," J. Eng. Sci. Technol., vol. 17, no. 5, pp. 3326-3341, 2022.
[11]. V. Suresh et al., "Facial recognition attendance system using Python and OpenCV," Quest J. Softw. Eng. Simul., vol. 5, no. 2, pp. 18-29, 2019.
[12]. P. Patil and S. Shinde, "Comparative analysis of facial recognition models using video for real time attendance monitoring system," in Proc. 2020 4th Int. Conf. Electron. Commun. Aerosp. Technol. (ICECA), 2020, pp. 850-855, doi: 10.1109/ICECA49313.2020.9297374.
[13]. P. Raghu, M. Santosh, and C. Lohith, "Student attendance monitoring system using IoT and RFID," Int. J. Sci. Res. Eng. Trends, vol. 9, no. 4, 2023.
[14]. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), 2005, pp. 886-893, doi: 10.1109/CVPR.2005.177.
[15]. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583-596, Mar. 2015, doi: 10.1109/TPAMI.2014.2345390.
[16]. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499-1503, Oct. 2016.
[17]. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 815-823.
[18]. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3-19.