ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Real-Time Intrusion Detection in Cyber-Physical Systems Using YOLO v11 and Facial Recognition

Boonsong Srichand[1] and Nithizethe Mhuadthongon[2]

ABSTRACT

This research proposes the development of a Cyber-Physical System (CPS) for real-time monitoring and intrusion detection on edge computing architectures to mitigate high false alarm rates (FAR) of traditional systems while addressing VRAM constraints of edge processing hardware and supporting efficient data transmission using event-based communication methods. The prototype combines YOLO11, DeepSORT, and InsightFace models built under the 5C CPS architecture framework. To overcome VRAM constraints, a shared resource architecture with a global GPU lock is proposed. Additionally, mechanisms for identity stabilization and temporal reasoning were implemented to ensure reliable decision-making under changing conditions. The shared resource architecture used 71.23% less VRAM when processing six camera streams at once. This results in lower resource consumption for this type of application. The overall accuracy of the system is 96.25%, and the temporal voting mechanism improves identification accuracy with occlusion by 30.5%. In addition, no false-positive event was observed in the controlled access-control evaluation, corresponding to a 0.00% false-alarm rate under the evaluated conditions. This study shows that efficient hardware resource allocation allows the implementation of complex AI models without model compression. At the same time, context-aware decision-making mechanisms provide dependability in intelligent systems in CPS applications.

Article information:

Keywords: Cyber-Physical Systems, Edge Computing, YOLO11, DeepSORT, InsightFace, Dynamic Risk Scoring

Article history:
Received: April 29, 2026
Revised: August 1, 2026
Accepted: October 15, 2026
Published: October 31, 2026 (Online)

DOI: 10.37936/ecti-cit.2026204.267387

1. INTRODUCTION

The accelerated growth of digital infrastructures has led to a surge in demand for intelligent surveillance systems that can operate 24/7 within critical settings, including data centers, enterprise facilities, and restricted-access infrastructures [1-2]. These environments require prompt detection of abnormal activities to safeguard against unauthorized access and third-party interference. To fulfill these requirements, recent surveillance systems are increasingly being integrated with computer vision and deep learning techniques to automate detection and assist in real-time decision-making processes [3].

Contemporary surveillance methodologies typically consist of three functional pillars: object detection and tracking, identity recognition, and access control. The efficiency and accuracy of YOLO for real-time object detection led to the extensive adoption of this deep learning model [4-5]. DeepSORT is a tracking algorithm that enables consistent identity assignment across consecutive frames [6]. Moreover, InsightFace is a face recognition system that carries out identity verification using deep feature representations [7]. The development and assessment of these components, however, generally occur independently, while their unification in a surveillance architecture is limited.

When deployed in real-world environments, especially on edge computing devices, these systems face operational constraints. Using edge computing enables video streams to be processed locally, thereby reducing network latency and preventing sensitive data from being transmitted to remote servers [8-9]. Nonetheless, system effectiveness is constrained by hardware resources, particularly GPU memory capacity, as the number of processed camera streams

[1,2]The authors are with the Digital Technology Program, School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi, Thailand, Email: boonsong.s@coj.go.th and Nithizethe.Mhu@stou.ac.th
[2]Corresponding author: Nithizethe.Mhu@stou.ac.th

increases. Traditional designs setting aside distinct model instances per camera stream may quickly deplete video random access memory (VRAM), leading to erratic runtime behavior of the system or a total runtime failure [10].

In addition to computational limitations, current surveillance systems' decision-making mechanisms remain simple. Traditional access control mechanisms operate in binary mode and grant access solely based on absolute identity verification, treating individuals as either authorized or unauthorized. The model does not simulate real operational conditions, as access decisions are not based on contextual elements such as spatial permissions, temporal constraints, and behavioral characteristics [11]. Consequently, false alarms may be triggered by actual operators of interest working under certain situations, leading to an increase in the false alarm rate [12].

Another challenge is the instability of facial recognition systems under dynamic monitoring conditions. When an object or scene shifts its pose, moves, or gets blocked, any subsequent frames may be wrongly assigned an identity. Thus, there may be a case of identity switching or recognition error [13]. As a result, achieving stable identity tracking over time requires temporal consistency in continuously monitored environments [14].

The identified limits suggest that current studies mostly deal with detection, recognition, and computational optimization as -separate issues while ignoring their interaction within a system. As a result, limited research has examined integrated surveillance architectures that can jointly handle constraints on computational resources, context-aware decision-making, and stable identity tracking in real-world scenarios [15]. This underscores the importance of a system-level approach to integrate these aspects within a cyber-physical systems framework. Although YOLO, DeepSORT, and InsightFace are well-established, off-the-shelf algorithms, the primary scientific contribution of this study is the system-level architectural design that integrates these models under the 5C CPS framework. Specifically, the novelty resides in (1) the implementation of the Global GPU lock combined with Frame Skipping for resource serialization on constrained edge devices, (2) the Multi-dimensional Dynamic Risk Scoring engine that contextualizes binary access control, and (3) the 3-frame Temporal Voting mechanism designed to mitigate occlusion-induced identity switching. Accordingly, YOLO11, DeepSORT, and InsightFace serve as established perception and recognition components, whereas the original contribution of this study lies in the resource-sharing, temporal identity stabilization, and context-aware decision mechanisms that coordinate these components within the proposed CPS architecture.

To overcome these challenges, this study designs and develops an intelligent real-time surveillance framework based on the 5-level Cyber-Physical Systems (5C CPS) architecture for edge computing devices. The framework coordinates YOLO11 for person detection, DeepSORT for multi-object tracking, and InsightFace for facial recognition through shared computational resources to perform multi-camera processing, maintain stable identity tracking, and support context-aware access decisions.

The contributions of this study are summarized as follows:

1) A shared resource architecture is proposed to enable multiple surveillance cameras to share a single deep learning model instance on an edge computing device. A global GPU lock mechanism is introduced to regulate GPU access and significantly reduce VRAM consumption during multi-camera processing.

2) A multi-dimensional dynamic risk scoring mechanism is developed to extend traditional binary access control by incorporating identity, spatial, and temporal contextual factors, enabling context-aware decision-making in complex operational environments.

3) An identity stabilization mechanism is introduced based on name locking and a 3-frame temporal voting strategy to improve consistency in identity recognition under dynamic monitoring conditions, particularly in the presence of occlusion and motion variation.

The rest of the paper is arranged as follows. Section 2 reviews the related work. The system architecture and methodology are proposed in Section 3. The results and discussion of the experiment are listed in Section 4. Finally, Section 5 concludes the study, and future research directions are outlined.

2. RELATED WORKS

2.1 Introduction to Cyber-Physical Systems and Security

As the use of IoT (Internet of Things) grows exponentially, Cyber-Physical Systems (CPS) are becoming increasingly interconnected environments in which physical processes and computational intelligence design and control each other. Secured environments are no longer limited to static access control and post-facto analysis. They require continuous monitoring and contextual interpretation of events. Traditional approaches that rely on predefined rules or forensic investigation thus do not suffice, as dynamic CPS scenarios require real-time situational awareness.

Recent studies have explored integrating computer vision, deep learning, and edge computing to support automated surveillance in CPS environments. However, many of these efforts concentrate on individual technical components while giving less attention to how these components operate together within a unified CPS architecture. In addition, recent studies

have increasingly highlighted system-level challenges related to computational efficiency and scalability in multi-camera surveillance environments. Processing multiple concurrent video streams on edge devices has been shown to impose significant GPU memory constraints and may lead to performance degradation under real-time conditions [16–20]. Furthermore, conventional deployment strategies that allocate independent model instances for each camera stream are often inefficient in resource-constrained environments, limiting their practical applicability in large-scale surveillance systems [17–18]. To address this limitation, the related works are organized into three thematic areas: real-time object detection and tracking, edge-based facial recognition, and automated access control with anomaly detection.

2.2 Real-Time Object Detection and Spatial Tracking

Object detection and tracking are very important to spatial security in a CPS environment. The YOLO (You Only Look Once) family of models has seen wide acceptance due to its balance of accuracy and computational efficiency [5]. A number of recent studies have successfully integrated YOLO with tracking algorithms to maintain the temporal consistency of people. Jiang *et al.* [21] implemented a YOLOv5-DeepSORT framework for personnel tracking in surface mine settings and achieved robust detection performance under challenging lighting conditions.

To make tracking more robust, researchers have used algorithms and sensor fusion. Wang and Mariano [22] combined YOLOv8s and ByteTrack to tackle occlusion-induced issues, whilst Dai *et al.* [23] fused YOLOv8 with LiDAR data to enhance spatial perception in an autonomous system. Moksyakov *et al.* [24] enhanced tracking trajectories through a sliding innovation filter in YOLO-based tracking frameworks.

Although these studies show strong performance in real-time detection and tracking, they mainly treat persons as anonymous. Hence, though they offer precise localization and trajectory continuity, they do not promote identity-aware reasoning or verification of access rights in secure CPS environments. This limitation suggests that detection and tracking should not be the only techniques used for security applications that require both spatial awareness and identity-based decisions.

2.3 Edge-Optimized Facial Recognition Systems

According to recent studies, deep learning-based facial recognition can solve the dilemma of anonymous tracking versus identity verification. Nonetheless, it has proven challenging to deploy complex biometric models in IoT due to the restrictions on edge devices arising from computational, memory, and bandwidth constraints [25-26]. In [27], Boutros *et al.* proposed PocketNet, an efficient lightweight network relying on multistep knowledge distillation for accuracy enhancement. George and others [28] proposed EdgeFace, a hybrid CNN-Transformer model designed for edge deployment with recognition performance and computational efficiency.

Real-world applications have also been investigated. The study by Chang *et al.* [29] demonstrates the feasibility of ESP32-based edge computing for object detection, reducing cloud dependence and communication delay. Zhang *et al.* [30] presented a device-edge collaborative method to identify occluded faces through the fusion of cross-domain features, addressing the scenario of partial occlusion in surveillance.

Regardless of this advancement, a majority of edge-based facial recognition systems function as point-based authentication mechanisms rather than continuous identity tracking processes. Therefore, identity verification is usually executed at a single point in time and does not ensure identity consistency while moving through the monitored environment. The lack of persistent enforcement of identity verification could create a potential security loophole.

2.4 Automated Access Control and Anomaly Detection

The convergence of detection and recognition has enabled the development of automated access control and behavior analysis systems. Anjeana and Anusudha [6] presented a real-time framework integrating YOLO and InsightFace, demonstrating the feasibility of combining object detection with identity verification. Abdullahu *et al.* [31] proposed a decentralized hybrid multi-face recognition system for IoT environments. Sohail *et al.* [32] addressed the challenge of multi-pose face matching using deep learning, which is essential for unconstrained surveillance conditions [33-34].

At the same time, anomaly detection has progressed from simple motion-based methods toward more advanced spatiotemporal analysis. Vosta and Yow [35] utilized a CNN-RNN architecture for real-world violence detection. More recent approaches, including Violence-YOLO [36] and YOLO-ABD [37], adapted YOLO-based models to detect abnormal pedestrian behavior. Alsabei *et al.* [38] demonstrated the scalability of YOLOv9 in detecting abnormal crowd behavior in dense environments.

Duja *et al.* [39] mentioned that a majority of anomaly detection systems remain siloed from identity management in that they are able to detect abnormal events but not their access or identity. As a result, these systems distinguish abnormal behavior but cannot tell whether it is permitted or forbidden. The division between anomaly detection and access control based on identity demonstrates a shortcom-

ing of contemporary surveillance systems, where contextual decision-making regarding access is not calibrated with behavior.

2.5 Research Gap and Proposed Contribution

New research on surveillance systems based on CPS reveals a gap in the methodology.

Most methods tackle the surveillance-related task within a single operational domain: they either deliver robust continuous tracking without identity verification [21, 22], gainfully conduct efficient edge-based yet static facial recognition [31, 33], or detect behavioral anomalies without combining contextual access control [36, 38]. The distinction is consistent with the limitations highlighted in Section 1, where detection, recognition, and decision-making are often developed in isolation rather than as part of a fully integrated surveillance system.

In addition, current access control architectures still mainly depend on binary decision logic and do not have multi-dimensional evaluation paradigms where temporal parameters, role-based privileges, and dynamic risk assessment are considered at the same time. Nonetheless, current methods do not offer a unified framework for joint continuous tracking, identity continuity, and context decision-making under edge computing constraints. As a result, these systems cannot perform continuous identity tracking, context-aware decision-making, and efficient operation under edge computing constraints simultaneously. Moreover, it has been shown in past works that the use of multi-camera surveillance systems on edge platforms comes with significant resource management and scalability challenges, especially with limited GPU memory [16–18]. Nonetheless, these works mainly address computational efficiency in isolation, whereas they do not furnish a plan for resource optimization in coordination with identity tracking and context-aware decision-making in the scope of a unified CPS framework. Accordingly, the contribution of the proposed Shared Resource Architecture is positioned at the system-integration level for memory-constrained CPS deployment, rather than as a throughput-oriented replacement for established parallel batching frameworks.

To address these limitations, the present study proposes an integrated cyber-physical system based on the 5-level CPS (5C) architecture operating within an edge computing environment. The system integrates YOLO11 for person detection, DeepSORT for multi-object tracking, and facial recognition within a shared computational framework. Rather than emphasizing individual component performance, the proposed system focuses on the coordinated operation of detection, identity tracking, and decision-making processes. In addition, this study extends beyond binary access control by introducing a multi-dimensional evaluation mechanism that considers identity, spatial context, and temporal conditions. To support reliable monitoring in dynamic environments, the system also incorporates an identity stabilization strategy to maintain temporal consistency during tracking. Integrating detection, identity tracking, and decision-making within a unified CPS-oriented architecture directly addresses the fragmentation identified in existing studies. It enables more context-aware surveillance in real-world environments.

3. METHODOLOGY

This study adopted an experimental system development design to construct and evaluate a real-time intelligent surveillance system operating on edge computing devices. The system was developed within a controlled environment to simulate real-world surveillance scenarios involving continuous video streaming, identity recognition, and access control decision-making.

This design was selected because it supported the integration of multiple computational components, including object detection, tracking, facial recognition, and decision-making processes within a unified cyber-physical system (CPS). The implementation followed a structured software engineering approach by combining the Model–View–Controller (MVC) principle with the 5-level CPS (5C) architecture [40]. This combination allowed the separation of system control, data processing, and visualization while maintaining interaction between physical sensing devices and cyber intelligence.

Within this framework, surveillance data captured from physical sensors were progressively transformed into actionable security information through sequential processing stages. Video streams were acquired, processed using artificial intelligence models, and evaluated to identify potential security risks in real time. The overall architecture of the system is illustrated in Fig. 1. Within the proposed pipeline, YOLO11, DeepSORT, and InsightFace were employed as established components for person detection, multi-object tracking, and facial recognition, respectively. These models were not modified at the algorithmic level. The original development in this study focused on their system-level coordination through the Shared Resource Architecture with Global GPU Lock and Frame Skipping, the 3-frame Temporal Voting mechanism for identity stabilization, and the Multi-dimensional Dynamic Risk Scoring mechanism for context-aware access decisions. These mechanisms form the principal methodological contributions evaluated in the subsequent experiments.

3.1 The 5C CPS Framework

The system was designed with five layers to manage data effectively and process it efficiently in cyber-

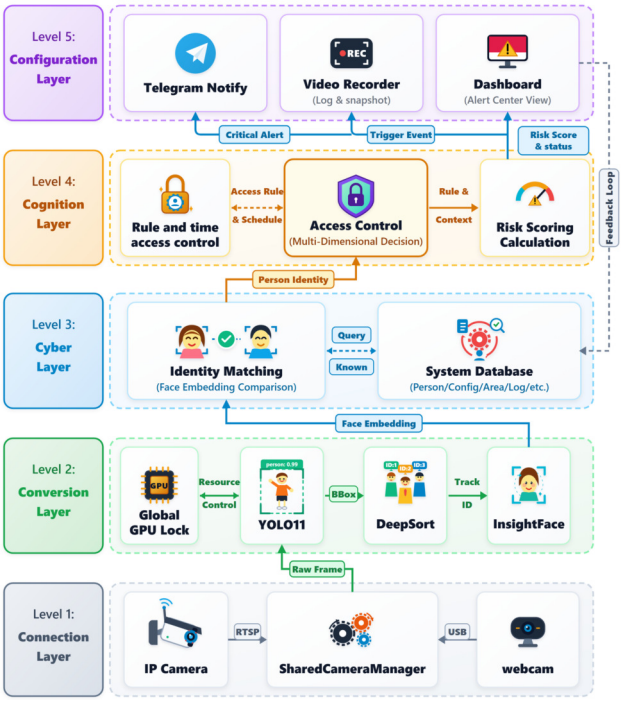


Fig.1: *The proposed 5C Cyber-Physical Systems architecture.*

physical surveillance. Through sensing, processing, analyzing, and responding, raw visual data was gradually converted into information by the 5C CPS structure. All layers of the architecture performed a specific and defined function while communicating continuously with the neighboring layers. Thus, as a whole, it worked as a seamless cyber-physical pipeline. The architecture with distinct layers defined the relation between raw data acquisition, computation, and decision-making for real-time operation under edge computing constraints.

(1) Connection Layer: The IP camera network captured continuous video streams at the connection layer. The infrastructure utilized Power over Ethernet (PoE, IEEE 802.3af/at). UTP cables provided the data and power connection for the project. Video streams were sent to the edge device using the Real-Time Streaming Protocol (RTSP) over either UDP or TCP to maintain low latency. The incoming streams were treated as sequential frames, which served as input for subsequent processing. The configuration chosen must ensure sufficient data throughput and low latency during real-time operation. Limiting the communications primarily to the local edge environment decreased dependencies on external networks, and hence, the operation remained uninterrupted even during network outages.

(2) Conversion Layer: The Conversion Layer processed incoming frames using the YOLO11 model to detect human subjects within the surveillance area. Each frame was analyzed to determine the spatial location of detected individuals. The resulting bounding box coordinates were forwarded to the DeepSORT algorithm, where appearance features were extracted and spatiotemporal track IDs were assigned. This stage was designed to maintain continuous tracking across frames, allowing the system to preserve identity continuity over time. As a result, detected individuals were associated with consistent tracking identifiers, forming the basis for subsequent identity recognition and contextual evaluation.

(3) Cyber Layer: The Cyber Layer functioned as the main computational component of the system. Facial regions extracted from detected individuals were processed using the InsightFace framework. RetinaFace was used for face detection, and ArcFace generated 512-dimensional feature vectors representing facial characteristics. These feature vectors were compared with a pre-enrolled facial database to determine individual identities. This process converted visual input into identity-related information, enabling the system to move from object-level detection to identity-aware analysis.

(4) Cognition Layer: The Cognition Layer evaluated outputs from the Cyber Layer using a dynamic risk scoring mechanism. Identity recognition results were interpreted together with contextual information, including spatial position and temporal behavior, to estimate potential security risks. This stage enabled context-aware interpretation of events, allowing the system to evaluate whether detected individuals met predefined access conditions within the surveillance environment.

(5) Configuration Layer: The Configuration Layer controlled system responses based on evaluated risk levels. When a critical condition was identified, alert notifications were generated and transmitted to authorized personnel through secure HTTPS communication using RESTful APIs. The system also initiated retrospective video recording to support further investigation. Through this mechanism, analytical results were translated into system actions, completing the operational cycle between sensing, analysis, and response.

3.2 Shared Resource Architecture

The real-time computer vision system was deployed on an edge computing device equipped with a GPU limited to 4 GB of VRAM. In this setting, multiple video streams had to be processed simultaneously in real-time. Given this restriction, assigning one copy of YOLO11 and one copy of the InsightFace model to each camera stream consumed too much memory and resulted in CUDA out-of-memory errors. Another mechanism was necessary to handle processing resources.

To address this issue, a shared resource architecture was introduced. Each deep learning model was loaded into GPU memory only once using the Singleton pattern, allowing all camera streams to access the

same model instance. This approach reduced redundant memory allocation and enabled stable operation under limited GPU capacity.

To control concurrent access to shared resources, a global GPU lock mechanism was applied. This mechanism applied the principle of mutual exclusion, ensuring that inference requests from multiple camera streams were processed sequentially. This design prevented simultaneous access conflicts and maintained stable GPU utilization during real-time operation. This design therefore prioritizes predictable GPU memory usage and operational stability under the 4 GB VRAM constraint, while accepting the throughput reduction associated with sequential inference. The proposed mechanism is intended as a lightweight shared-resource solution for the evaluated resource-constrained deployment rather than as a replacement for optimized parallel batching frameworks.

The operational workflow of this architecture is illustrated in Fig. 2, which shows how multiple video streams were scheduled and processed using a shared model instance. To further maintain real-time performance, the facial recognition module was configured to process a limited number of faces per frame. In this implementation, a maximum of five faces per frame were processed. This threshold was determined during system evaluation to balance computational workload with real-time processing requirements under constrained hardware conditions.

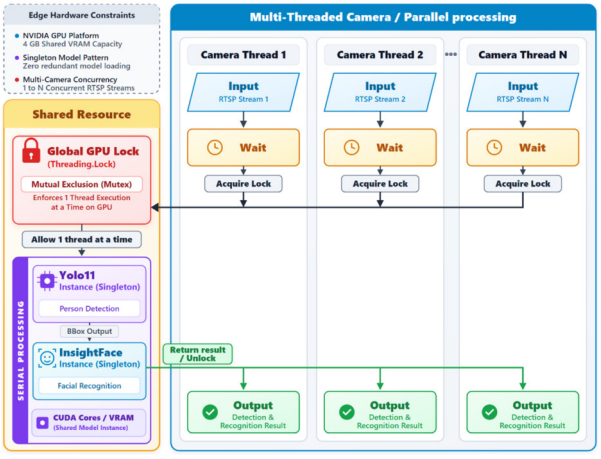


Fig.2: *The workflow of the Shared Resource Architecture and GPU Lock mechanism.*

3.3 Person Matching and Identity Conflict Management

Since the person detection (YOLO11) and facial recognition (InsightFace) models operate independently, the system required a mechanism to accurately associate facial bounding boxes with corresponding body bounding boxes. To quantify the association between detected faces and tracked bodies, a matching cost was defined based on spatial distance and bounding box overlap [41]. The matching probability was evaluated using the following cost function.

$$Cost = (W_d \times D_{norm}) + (W_o \times (1 - IoFA)) \quad (1)$$

where D_{norm} represents the normalized relative distance between the center points of the face and body bounding boxes, and IoFA (Intersection over Face Area) denotes the overlap ratio. W_d and W_o represent weighting parameters assigned to spatial displacement and overlap, respectively.

In this study, these parameters were empirically tuned to $W_d = 0.5$ and $W_o = 50.0$ based on preliminary system evaluation under typical surveillance conditions. The relative values of these parameters determine the balance between spatial proximity and bounding-box overlap during face-to-body association. Increasing W_d would place greater emphasis on the distance between face and body centers, whereas increasing W_o would make overlap consistency more influential in the matching decision. Consequently, inappropriate weighting may alter the association boundary and increase the likelihood of incorrect or missed face-to-body assignments, which can subsequently affect false-positive and false-negative decisions at the system level. The selected values were therefore treated as deployment-specific parameters rather than universal constants.

This formulation combined spatial proximity and overlap consistency to estimate the likelihood that a detected face corresponded to a specific tracked individual. Lower cost values were interpreted as higher matching confidence between face and body regions. To align the matching process with human anatomical structure, a corrective condition was applied. When the Y-coordinate of the face center was located above the body center, the total cost($F_{y_{center}} < T_{y_{center}}$) was reduced to favor anatomically consistent matches.

$$\begin{aligned} &\text{If } (F_{y_{center}} < T_{y_{center}}) \text{ Then} \\ &C_{total} = C_{total} \times P_{pos} \end{aligned} \quad (2)$$

This adjustment introduced a spatial constraint that prioritized anatomically plausible face-to-body associations during matching. In environments where multiple individuals overlapped within the same field of view, identity instability could occur due to occlusion and ambiguous detections. To mitigate this issue, a name locking mechanism was implemented. The system sorted facial recognition confidence scores within each frame and assigned identity labels to Track IDs based on the highest confidence values.

Once an identity was assigned to a Track ID, the association was temporarily fixed within the same frame to prevent reassignment. This mechanism reduced identity switching and ensured consistency be-

tween detection, tracking, and recognition outputs during multi-person interactions.

3.4 Temporal Voting Mechanism

Single-shot facial recognition was found to be susceptible to inaccuracies caused by variations in head pose, partial occlusion, and temporary visibility changes. Such conditions introduced instability in identity recognition when decisions were based on a single video frame. To improve recognition stability, the system incorporated a temporal voting mechanism that evaluated facial recognition outcomes across multiple consecutive frames.

The mechanism maintained a short temporal history of recognition results for each track ID using a FIFO (First-In, First-Out) ring buffer data structure. This buffer stored recognition outputs from the most recent three frames for each tracked individual. By retaining a sequence of recent predictions, the system evaluated identity consistency over time rather than relying on a single recognition event.

The temporal voting mechanism was integrated with a multi-view enrollment strategy, where up to five facial images were registered for each individual in the system database. This configuration was designed to improve robustness by capturing variations in facial appearance under different viewing conditions.

During runtime operation, the system applied a majority consensus rule for identity verification. An identity was confirmed only when recognition results were consistent in at least two out of three buffered frames. This rule was introduced to reduce the influence of transient recognition errors caused by occlusion, pose variation, or temporary detection inconsistencies.

By evaluating identity decisions across a short temporal window, the Temporal Voting mechanism acted as a temporal smoothing process. This approach was designed to support more consistent identity verification in dynamic surveillance scenarios where individuals were in motion or partially occluded. The operational workflow of the 3-frame temporal voting mechanism utilizing the FIFO ring buffer is illustrated in Fig. 3.

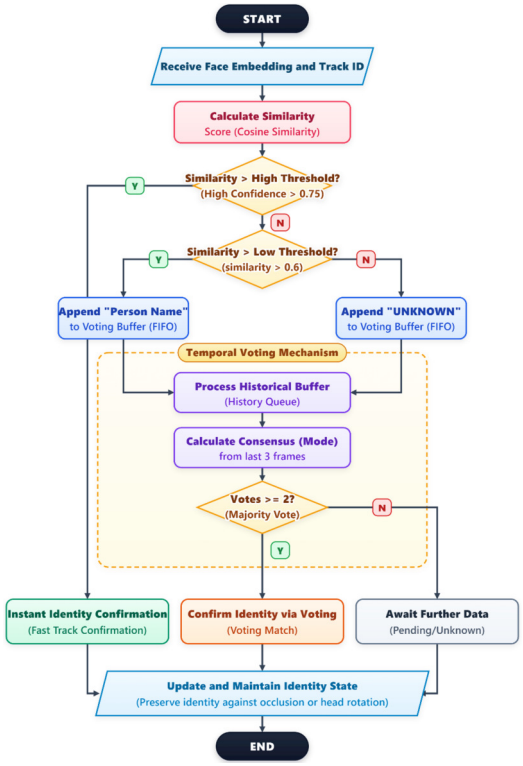


Fig.3: *The 3-frame Temporal Voting mechanism utilizing a FIFO Ring Buffer for identity verification.*

3.5 Multi-dimensional Dynamic Risk Scoring

To overcome the high false alarm rate common in binary allow/deny access control systems, a multi-dimensional dynamic risk scoring mechanism was developed. The system assessed risk levels on a scale of 0-100 using contextual information from the cyber layer as shown in Fig. 4. This method allowed for a more flexible approach to analyzing whether or not security conditions are met in the dynamic surveillance of various environments.

The decision process was structured into two evaluation paths. In the first path, defined as an override

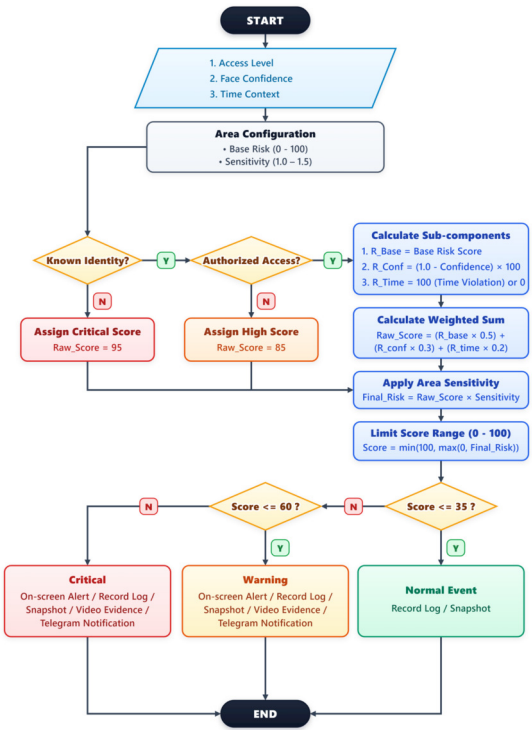


Fig.4: *The Multi-dimensional Dynamic Risk Scoring process.*

rule for maximum security enforcement, predefined high-risk scores were assigned when critical identity conditions were detected. When a detected individual had no corresponding facial record in the database (Unknown), the risk score was set to 95. When an individual was recognized but lacked spatial authorization and was therefore considered unauthorized, the score was set to 85. These conditions were treated as high-risk events to ensure an immediate response to critical security violations.

For individuals who were both recognized and authorized, the system applied a contextual weighted evaluation model to estimate risk levels. Instead of excluding authorized individuals from monitoring, contextual and behavioral conditions were continuously evaluated using the following linear scoring formulation [42].

$$\text{Risk}_{\text{total}} = \min\big(100,\ S \times (\alpha \times R_{base} + \beta \times R_{conf} + \gamma \times R_{time})\big) \tag{3}$$

Where S represents the access privilege status, R_{base} denotes the area significance level ($\alpha = 0.5$), R_{conf} represents the inverse of the model confidence score ($\beta = 0.3$), and R_{time} corresponds to temporal behavioral factors such as prolonged presence ($\gamma = 0.2$).

This formulation was designed to integrate spatial sensitivity, recognition reliability, and temporal behavioral indicators into a unified risk evaluation framework. A predefined threshold of 60 was used to determine escalation conditions. When the computed risk score exceeded this threshold, the system classified the situation as a critical security state.

When a critical state was triggered, the system activated a response sequence to support incident handling. The FIFO ring buffer retrieved the five seconds of video preceding the trigger event, which were stored as forensic evidence. The system then transmitted an alert notification to the monitoring interface for immediate response.

4. RESULTS AND DISCUSSIONS

The Cyber-Physical System's performance evaluation was carried out in a realistic environment on an edge server with an entry-level GPU (4 GB VRAM). The aim of the experimental trial was to evaluate the real-life deployment feasibility of the proposed architecture for surveillance purposes. The experiment was conducted to evaluate resource efficiency, decision-making accuracy, and environmental robustness. The functional goals of the system included optimal resource usage in edge computing systems, robust contextual decision-making, and stable recognition performance in real-world situations.

4.1 System implementation in real-world environments

The developed prototype was implemented as a comprehensive security monitoring solution in a real-world operational environment. The hardware and network topology were configured to connect the image acquisition nodes to the edge processing server, enabling stable real-time processing at the edge. This deployment configuration enabled the proposed cyber-physical system to continuously acquire visual data from distributed surveillance cameras and process the incoming streams directly on the edge device. The overall system implementation in the operational environment is illustrated in Fig. 5.

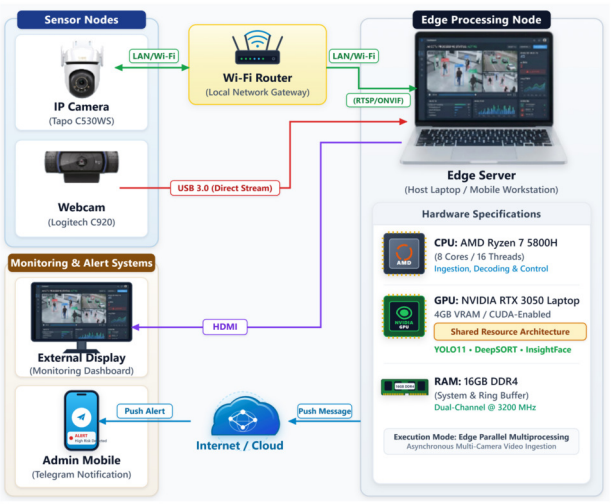
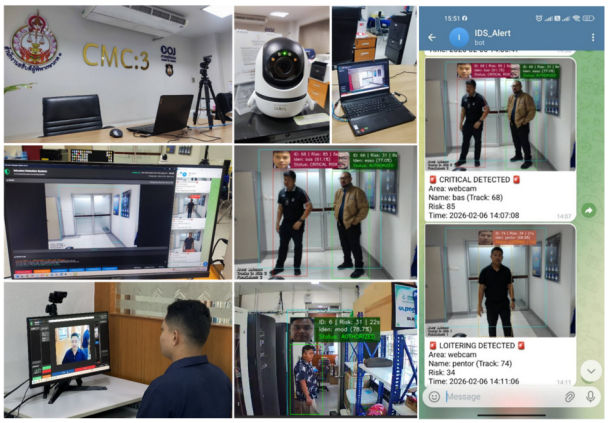


Fig.5: *System implementation in real-world environments.*

Building upon this infrastructure, the system processed live video streams and presented the outputs through an integrated monitoring dashboard. The monitoring interface displayed real-time detection results, including bounding boxes of detected individuals, associated identity labels, and dynamically calculated risk levels. This allowed security personnel to monitor system outputs alongside the physical environment during operation, as shown in Fig. 6a.

The response was triggered if there was a violation, such as an unknown person entering the monitored area or a known person entering a restricted zone during prohibited periods. The local forensic video was recorded, and a high-priority alert notification with an event snapshot was generated. The alerts promptly reached the security personnel via the Telegram Messenger application, as demonstrated in Fig. 6b. The deployment results show that the system continuously operated in the target environment and, under real application conditions, supported the real-time monitoring task.



(a) (b)

Fig.6: *System deployment and operational workflow: (a) Experimental setup and real-time monitoring interface, (b) Automated critical alerts via Telegram.*

4.2 Processing Performance and Scalability

The capability of the Shared Resource Architecture to handle concurrent multi-camera streams (1, 2, 4, and 6 cameras) was evaluated to quantify the effect of GPU resource serialization, as summarized in Table 1. For system evaluation, real-time operation was defined as a processing rate of at least 12 FPS, while soft real-time performance corresponded to 8–12 FPS. Experimental results show that the 1-camera and 2-camera configurations satisfied the real-time requirement at 14.1 FPS and 12.3 FPS, respectively. Increasing the workload to 4 concurrent cameras reduced the rate to 10.4 FPS, operating within the soft real-time range. In contrast, the 6-camera configuration achieved 7.8 FPS, falling slightly below the soft real-time threshold and reflecting the throughput-latency trade-off introduced by sequential GPU access.

This throughput reduction highlights the latency trade-off of the Global GPU Lock. To mitigate the latency introduced by sequential queuing, the system architecture incorporates a strategic Frame Skipping mechanism. By applying Frame Skipping to reduce the number of frames submitted for inference from the 30 FPS input streams, the system maintains an effective AI processing rate of 7.8 FPS (approximately 128 ms per processed frame) at 6 cameras. Although this processing rate is below the predefined soft real-time threshold of 8 FPS, the system remained operational under the evaluated 6-camera workload. Although fast-moving objects or short-lived events may appear in fewer processed frames as the camera load increases, potentially reducing the opportunity for detection, the integration of DeepSORT utilizing Kalman filtering helps maintain trajectory continuity over short temporal gaps. However, this mechanism cannot eliminate the possibility of missing events that occur entirely between processed frames.

Consequently, 4 cameras represent the highest experimentally evaluated configuration that maintains soft real-time performance. In comparison, 6 cameras represent the first evaluated configuration that falls below the soft real-time threshold on the evaluated 4 GB VRAM hardware configuration. Furthermore, based on the degradation trend observed during GPU queueing, scaling beyond 6 concurrent cameras would be expected to reduce the processing rate further below the predefined soft real-time threshold.

In addition to processing throughput, the experiment evaluated computational resource utilization, including GPU memory consumption and CPU usage. The results indicate that CPU usage increased from 12% to 39% as the number of camera streams increased, reflecting the increased computational workload under multi-camera operation. Attempting a natively configured multi-stream baseline (instantiating independent models per camera) resulted in immediate Out-Of-Memory (OOM) crashes when initializing the second camera stream on the evaluated 4 GB edge device, necessitating the shared resource architecture. This empirical baseline demonstrates that independent per-camera model instantiation was not feasible under the evaluated 4 GB VRAM configuration. The failure to initialize the second stream under the native per-camera configuration prevented the collection of stable two-camera FPS or VRAM measurements for this baseline; therefore, the observed OOM event itself represents the empirical scalability limit of the native configuration on the evaluated hardware.

Industry-standard edge-streaming frameworks, such as NVIDIA DeepStream, employ optimized parallel processing, batching, and memory-management mechanisms to improve multi-stream processing efficiency [43]. In contrast, the proposed Shared Resource Architecture serializes access to a single shared model instance through the Global GPU Lock and therefore represents a different resource-management trade-off. In the proposed pipeline, dynamic face cropping generates variable-sized inputs. Applying dynamic batching to these inputs would require additional pre-processing, such as resizing or padding, together with additional memory-management operations before batch inference. Given the strict 4 GB constraint, sequential locking was selected as a lightweight and framework-agnostic approach that provides stable VRAM usage and isolation of inference streams under the evaluated hardware configuration.

A direct empirical benchmark against NVIDIA DeepStream or another industry-standard parallel batching framework was not conducted in this study. Therefore, the comparison is limited to the experimentally observed native multi-stream baseline and the architectural trade-offs between parallel batching and sequential resource sharing. Accordingly,

the proposed architecture is not claimed to outperform optimized batching frameworks in throughput; rather, its contribution lies in enabling stable multi-camera operation under the specific memory-constrained edge environment evaluated in this study.

In terms of Video Random Access Memory (VRAM) usage, the system demonstrated a substantial reduction in memory consumption compared with a conventional per-camera model instantiation approach. Processing 6 cameras under a traditional architecture would require a theoretical VRAM of 6,355 MB, exceeding hardware limits and potentially causing system instability. This estimation was based on the assumption that each camera stream loads an independent instance of the deep learning models into GPU memory.

Conversely, the developed Global GPU Lock architecture consumed only 1,828 MB of VRAM while processing 6 concurrent video streams. This shared-resource mechanism reduces redundant model loading and enables efficient multi-camera processing within the limited hardware capacity of edge computing devices, as illustrated in Fig. 7. The results indicate that the proposed architecture reduced VRAM consumption by approximately 71.23%, allowing the system to operate within the constraints of a 4 GB GPU environment.

Table 1: *Results of performance measurements for processing and resource utilization.*

Number of Cameras	Average FPS per Camera	Total VRAM Usage (MB)	CPU Usage (%)	Operational Status
1	14.1	1,355	12%	Normal (Real-time)
2	12.3	1,475	18%	Normal (Real-time)
4	10.4	1,576	26%	Normal (Soft Real-time)
6	7.8	1,828	39%	Operational (Below soft real-time)

Furthermore, the system was designed to operate within an edge-based processing paradigm, where video data were processed locally, and only event-driven outputs were transmitted. Instead of continuously streaming full video data, the system transmitted lightweight alert information, including textual notifications and snapshot images, when predefined conditions were triggered. This configuration reduced continuous network transmission requirements and supported stable operation under varying network conditions.

4.3 Accuracy of the Multi-dimensional Decision System

To evaluate the system's ability to classify access conditions and detect potential security threats,

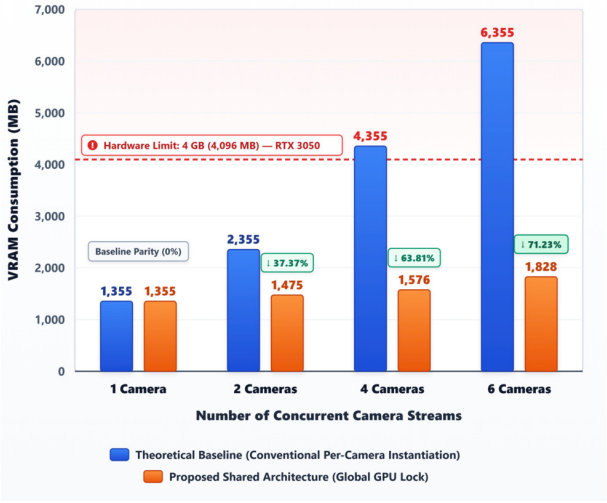


Fig. 7: *Chart comparing VRAM usage between the theoretical and proposed architectures.*

four access scenarios (Normal, Time-violation, Insider Intrusion, and Outsider Intrusion) were simulated across 80 experimental trials. The evaluation dataset (N = 80) was constructed using scenario-based testing to represent a range of surveillance conditions, including variations in viewing angles, partial occlusions, and temporal and spatial access constraints. The evaluation was therefore conducted as a controlled scenario-based experiment rather than using an unconstrained "in-the-wild" benchmark dataset. The classification results are summarized in the confusion matrix presented in Table 2.

The experimental results indicate that the proposed system overall attained 96.25% classification accuracy, successfully classifying both anomalous and authorized access events. Based on the confusion matrix, it generated 57 true-positive (TP) detections regarding abnormal scenarios and 20 true-negative (TN) detections regarding normal scenarios. Another three were false negatives (FN), which were failed detections of abnormal events identified as normal. There were no false-positive (FP) events observed, which means no normal scenarios were misclassified as abnormal.

Based on the observed confusion matrix (TP = 57, TN = 20, FP = 0, and FN = 3), the system achieved an accuracy of 96.25%, precision of 100.00%, recall (sensitivity) of 95.00%, specificity of 100.00%, and an F1-score of 97.44%. The observed false-positive (false-alarm) rate was 0.00%, as no false-positive event occurred among the 20 normal trials. The results show that the system maintained strong detection performance under the evaluated conditions. The observed 0.00% false-alarm rate applies specifically to the present controlled experimental setting. It should not be interpreted as evidence of zero false alarms in fully unconstrained or large-scale real-world deployment. The presence of three false negative

cases shows detections missed under limited circumstances. The results indicate that the system is capable of accurately distinguishing between abnormal and normal access conditions under the experimental conditions.

It is important to note that no false-positive event was observed within the 80 controlled experimental trials, which included low-light environments, partial occlusions (sunglasses, masks), and extreme facial angles. This robust performance is not solely the result of flawless visual classification, but rather the synergistic effect of a three-tiered defense: (1) multi-sample enrollment providing 5 diverse reference images, (2) the 3-frame Temporal Voting mechanism stabilizing intermittent occlusions, and (3) the contextual gating of the Dynamic Risk Scoring.

For Dynamic Risk Scoring, the coefficients $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$ control the relative contributions of spatial risk, recognition confidence, and temporal behavior, respectively. Increasing a coefficient increases the influence of its corresponding component on the final risk score and therefore changes how readily the predefined decision threshold of 60 is reached. For example, increasing α would make spatial-risk conditions contribute more strongly to escalation, whereas increasing β would make the decision more sensitive to recognition uncertainty, and increasing γ would increase the contribution of temporal behavior. These changes may alter the balance between false alarms and missed detections. However, the present study did not perform a systematic parameter sweep; therefore, quantitative changes in FAR and FNR for alternative coefficient settings cannot be claimed from the current experiments. The selected coefficients should therefore be regarded as empirically tuned settings for the evaluated environment and may require recalibration when the system is deployed under different operational conditions.

Table 2: *Confusion matrix of the proposed system (N=80).*

Actual Scenario (Ground Truth)	**Predicted: Abnormal** *(Status: WARNING / CRITICAL)*	**Predicted: Normal (Status: NORMAL)**	**Total**
Abnormal	**True Positive (TP) = 57** *(Correct Alert)*	**False Negative (FN) = 3** *(Missed Detection)*	**60**
Normal	**False Positive (FP) = 0** *(False Alarm)*	**True Negative (TN) = 20** *(Correct Non-Alert)*	**20**
Total	**57**	**23**	**80**

4.4 Robustness of Facial Recognition

To evaluate the robustness of the proposed facial recognition mechanism under realistic surveillance conditions, comparative experiments were conducted between a traditional single-shot recognition approach and the proposed 3-frame temporal voting mechanism. The evaluation was performed under three operational conditions: normal conditions, challenging pose conditions, and severe occlusion conditions, as illustrated in Fig. 8, with detailed results summarized in Table 3.

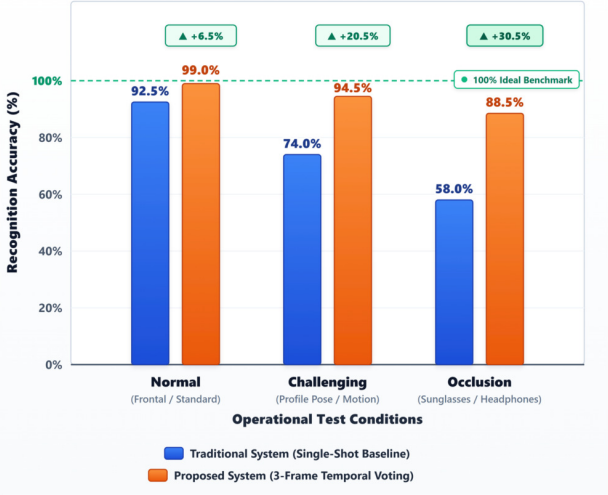


Fig.8: *Chart comparing accuracy under 3 test conditions.*

Table 3: *Comparison of facial recognition accuracy between the traditional baseline and the proposed system.*

Test Condition	**Baseline (Single-Shot)**	**Proposed (Temporal Voting)**	**Improvement**
Normal *(Frontal face, standard lighting)*	92.5%	99.0%	+6.5%
Challenging *(Profile view, downward gaze, fast movement)*	74.0%	94.5%	+20.5%
Occlusion *(Wearing sunglasses, over-ear headphones)*	58.0%	88.5%	+30.5%
Average	**74.8%**	**94.0%**	**+19.2%**

Under normal environmental conditions, where subjects directly faced the camera without obstruction, both systems achieved high recognition accuracy. The traditional baseline achieved 92.5% accuracy, while the proposed system achieved 99.0%. Under challenging conditions, including profile views, downward gaze, and motion, the performance of the traditional single-shot recognition approach decreased to 74.0%. In contrast, the proposed system maintained an accuracy of 94.5%, corresponding to an improvement of 20.5%. Under severe occlusion conditions, such as the presence of sunglasses or over-ear headphones, the accuracy of the traditional approach decreased further to 58.0%. The proposed method achieved an accuracy of 88.5% under the same conditions, representing an improvement of 30.5%.

Across all conditions, the average accuracy of the proposed system was 94.0%, while the base-

line accuracy was 74.8%, with an overall improvement of 19.2%. The findings show that the proposed mechanism maintained stable recognition performance across visual changes. Under difficult and occluded conditions, the gains were greater since single-frame recognition was less accurate under these conditions. The experimental conditions demonstrate the system's ability to perform consistently on the task of identity recognition in varying environments.

4.5 Discussion and Research Implications

The studies indicate that multi-camera edge-based surveillance systems are mainly limited by GPU memory [44], as each model is instantiated independently across parallel streams, leading to redundant allocation of memory and rapid exhaustion of VRAM. Utilization of an independent per-camera model led to resource misuse and instability. The proposed shared resource architecture addresses this issue by enabling centralized model sharing with limited GPU access. This matches the previous investigations that observed scalability and memory problems of multi-camera systems [16-19] and limitations of standard deployment strategies under constrained conditions [19-20]. Unlike earlier strategies that involved model compression or simplification of architectures, the proposed system relies on models of full capacity but enables higher resource efficiency, suggesting a practically deployable solution for scalable edge deployment.

The reduction in processing rate under increasing camera loads also introduces an important security trade-off. At the evaluated 6-camera configuration, the processing rate decreased to 7.8 FPS, which is below the predefined soft real-time threshold of 8 FPS. Under this condition, fast-moving individuals and short-duration events may be represented in fewer processed frames, increasing the possibility that brief visual events are not captured by the inference pipeline. The present experiments did not quantify the probability of missed detections as a function of object speed or event duration; therefore, this limitation should be interpreted qualitatively rather than as a measured miss rate. Nevertheless, the use of DeepSORT with Kalman filtering supports trajectory continuity across short temporal gaps, while the Frame Skipping mechanism prevents excessive inference queue accumulation. Accordingly, the current architecture prioritizes continuous operation and predictable GPU memory usage under the 4 GB VRAM constraint, accepting reduced temporal resolution as the number of concurrent camera streams increases.

From a decision-making perspective, the results highlight the limitations of binary identity-based systems, which often produce high false alarm rates due to the absence of contextual reasoning [11-12]. By integrating identity, spatial, and temporal factors, the Dynamic Risk Scoring mechanism supported a 0.00% observed false-alarm rate within the controlled 80-trial evaluation, as access decisions are evaluated based on contextual conditions rather than static identity labels, supporting the role of context-aware reasoning in CPS environments [15]. The presence of a small number of false negatives (FN = 3) suggests a conservative decision boundary, indicating a controlled trade-off between detection sensitivity and operational reliability in security-critical environments. The facial recognition results further indicate that single-frame approaches are highly sensitive to pose variation and occlusion [32], [45], since recognition decisions rely on instantaneous visual features that are unstable under dynamic conditions. The proposed 3-frame temporal voting mechanism mitigates these limitations by leveraging temporal consistency, resulting in stable recognition performance across all conditions. This supports theoretical perspectives on temporal aggregation, where short-term temporal smoothing improves prediction reliability in dynamic environments [14].

In addition to enhancing accuracy, identity stabilization mitigates system-level inefficiencies. Conventional systems often switch identities, causing false alerts and a heavy communication burden. The proposed method curtails this instability, ensuring the events maintain more consistency, which limits overlapping alerts and unnecessary data transmissions, an important feature, especially for scalable edge-based architectures. From a theoretical viewpoint, this study suggests that the perception, reasoning, and control of any CPS must be integrated at the system level for the effective operation of the system as an integrated whole. Isolated optimization of each component at the individual level does not address inefficiencies at the interaction level. Although these components have been studied in isolation in previous work [21], [28], [36], the present work shows that system-level integration improves performance in a number of measurable ways. In practice, the system can also be deployed on resource-constrained edge devices by lowering hardware demands and improving robustness. The event-driven means of communication helps in further reducing the continuous usage of the bandwidth, as information is sent only when there is an event. Hence, the system can be used even when network capacity is low. To sum up, implementing the management of shared resources, context-aware decision-making, and temporal identity stabilization in the framework of a CPS significantly enhances efficiency, accuracy, and robustness.

5. CONCLUSION

This study designed and implemented a real-time intelligent surveillance system based on the 5-level Cyber-Physical Systems (5C CPS) architecture for resource-constrained edge devices. By coordinating object detection (YOLO11), multi-object track-

ing (DeepSORT), and facial recognition (InsightFace) within a unified framework, the proposed system addresses key limitations of fragmented surveillance approaches. Experimental results confirm that coordinating shared resource management (achieving a 71.23% VRAM reduction across six cameras), 3-frame temporal voting (improving occluded recognition accuracy by 30.5%), and multi-dimensional dynamic risk scoring (attaining 96.25% classification accuracy with a 0.00% false-alarm rate in controlled evaluations) ensures dependable operation across varying workloads and environmental conditions.

Theoretically and practically, this work demonstrates that holistic system-level integration among perception, reasoning, and control is essential in advancing CPS surveillance beyond isolated component optimization. Prioritizing contextual awareness and identity stability over binary decision mechanisms reduces unnecessary alerts and bandwidth consumption, enabling complex AI-powered monitoring to operate reliably within restricted hardware and network infrastructures.

Nevertheless, this study has certain limitations, as the evaluation was conducted under controlled experimental conditions on a single hardware configuration, and sequential GPU locking imposes processing density limits at higher camera counts. Future research will extend this framework to large-scale, heterogeneous environments by incorporating adaptive behavior-level analysis, context-specific temporal adjustments, and distributed edge-cloud collaborative architectures.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Administrative Office of the Court of Justice Region 3 and the Nakhon Ratchasima Public Health Savings Cooperative, Limited, Thailand, for their continued support and assistance throughout the development and evaluation of this study.

AUTHOR CONTRIBUTIONS

Conceptualization, B.S. and N.M.; methodology, B.S. and N.M.; software, B.S.; validation, B.S. and N.M.; formal analysis, B.S. and N.M.; investigation, B.S. and N.M.; data curation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, N.M.; visualization, B.S.; supervision, N.M.; project administration, N.M. All authors have read and agreed to the published version of the manuscript.

References

[1] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, pp. 16979–16995, 2021.

[2] C. Chimpimol, N. Mhuadthongon and P. Krachodnok, "Development of a real-time backup battery voltage and environmental monitoring system for a server room using IoT technology," *EUREKA: Physics and Engineering*, no. 2, pp. 108–120, 2026.

[3] K. Khurshid, K. Khurshid, M. Usman Hadi, M. Al Bataineh and N. Saeed, "Securing AIoT Surveillance: Techniques, Challenges, and Solutions," in *IEEE Open Journal of the Communications Society*, vol. 6, pp. 6517-6550, 2025.

[4] M. L. Ali and Z. Zhang, "The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection," *Computers*, vol. 13, no. 12, p. 336, 2024.

[5] S. Kang, Z. Hu, L. Liu, K. Zhang and Z. Cao, "Object detection YOLO algorithms and their industrial applications: Overview and comparative analysis," *Electronics*, vol. 14, no. 6, p. 1104, 2025.

[6] R. Pereira, G. Carvalho, L. Garrote and U. J. Nunes, "SORT and Deep-SORT based multi-object tracking for mobile robotics: Evaluation with new data association metrics," *Applied Sciences*, vol. 12, no. 3, p. 1319, 2022.

[7] N. Anjeana and K. Anusudha, "Real-time face recognition system based on YOLO and InsightFace," *Multimedia Tools and Applications*, vol. 83, pp. 31893–31910, 2024.

[8] R. Xu, S. Razavi and R. Zheng, "Edge Video Analytics: A Survey on Applications, Systems and Enabling Techniques," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2951-2982, 2023.

[9] S. Wan, S. Ding and C. Chen, "Edge computing-enabled video segmentation for real-time traffic monitoring in the Internet of Vehicles," *Pattern Recognition*, vol. 121, p. 108146, 2022.

[10] H. -T. Chen, Y. Chiang and H. -Y. Wei, "Edge Computing Resource Management for Cross-Camera Video Analytics: Workload and Model Adaptation," in *IEEE Access*, vol. 12, pp. 12098-12109, 2024.

[11] R. Kalaria, A. S. M. Kayes, W. Rahayu, E. Pardede and A. S. Shahraki, "Adaptive context-aware access control for IoT environments leveraging fog computing," *International Journal of Information Security*, vol. 23, no. 4, pp. 3089–3107, 2024.

[12] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri and S. S. Ullah, "Anomaly detection in traffic surveillance videos using deep learning," *Sensors*, vol. 22, no. 17, p. 6563, 2022.

[13] M. Wang and W. Deng, "Deep face recognition:

A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[14] G. Zhang, J. Yin, P. Deng, Y. Sun, L. Zhou and K. Zhang, "Achieving adaptive visual multi-object tracking with unscented Kalman filter," *Sensors*, vol. 22, no. 23, p. 9106, 2022.

[15] E. Badidi, K. Moumane and F. E. Ghazi, "Opportunities, Applications, and Challenges of Edge-AI Enabled Video Analytics in Smart Cities: A Systematic Review," in *IEEE Access*, vol. 11, pp. 80543-80572, 2023.

[16] S. Kim, C. Kim and S. Kim, "Improving Performance of Real-Time Object Detection in Edge Device Through Concurrent Multi-Frame Processing," in *IEEE Access*, vol. 13, pp. 1522-1533, 2025.

[17] X. Kong, Y. Wu, H. Wang and F. Xia, "Edge Computing for Internet of Everything: A Survey," in *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23472-23485, Dec. 1, 2022.

[18] W. Wu and J. Lai, "Multi Camera Localization Handover Based on YOLO Object Detection Algorithm in Complex Environments," in *IEEE Access*, vol. 12, pp. 15236-15250, 2024.

[19] T. Kim, H. Park, Y. Jin, S. -S. Lee and S. Lee, "Partition Placement and Resource Allocation for Multiple DNN-Based Applications in Heterogeneous IoT Environments," in *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9836-9848, 2023.

[20] X. Hou, J. Zhou, L. Li, P. Cong, Z. Wu and M. Chen, "Latency and Reliability-Aware Dynamic Task Offloading and Scheduling for Energy-Harvesting Systems in Mobile Edge Computing," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 45, no. 8, pp. 3942-3956, Aug. 2026.

[21] J. Jiang, G. Xie, J. Cui and M. Guo, "Surface mine personnel object video tracking method based on YOLOv5-DeepSORT algorithm," *Scientific Reports*, vol. 15, no. 1, p. 17123, 2025.

[22] Y. Wang and V. Y. Mariano, "A Multi Object Tracking Framework Based on YOLOv8s and Bytetrack Algorithm," in *IEEE Access*, vol. 12, pp. 120711-120719, 2024.

[23] Y. Dai, D. Kim and K. Lee, "An advanced approach to object detection and tracking in robotics and autonomous vehicles using YOLOv8 and LiDAR data fusion," *Electronics*, vol. 13, no. 12, p. 2250, 2024.

[24] A. Moksyakov, Y. Wu, S. A. Gadsden, J. Yawney and M. AlShabi, "Object detection and tracking with YOLO and the sliding innovation filter," *Sensors*, vol. 24, no. 7, p. 2107, 2024.

[25] Z. Chang, S. Liu, X. Xiong, Z. Cai and G. Tu, "A Survey of Recent Advances in Edge-Computing-Powered Artificial Intelligence of Things," in *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13849-13875, 2021.

[26] M. Jiang and W. Gong, "Bidirectional Bluetooth Backscatter With Edges," in *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1601-1612, 2024.

[27] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner and A. Kuijper, "PocketNet: Extreme Lightweight Face Recognition Network Using Neural Architecture Search and Multistep Knowledge Distillation," in *IEEE Access*, vol. 10, pp. 46823-46833, 2022.

[28] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal and S. Marcel, "EdgeFace: Efficient Face Recognition Model for Edge Devices," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 2, pp. 158-168, 2024.

[29] Y.-H. Chang, F.-C. Wu, and H.-W. Lin, "Design and implementation of ESP32-based edge computing for object detection," *Sensors*, vol. 25, no. 6, p. 1656, 2025.

[30] P. Zhang, L. Tan, Z. Yang, F. Huang, L. Sun and H. Peng, "Device-edge collaborative occluded face recognition method based on cross-domain feature fusion," *Digital Communications and Networks*, vol. 11, no. 2, pp. 482–492, 2025.

[31] E. Abdullahu, H. Wache, and M. Piangerelli, "Secure and decentralized hybrid multi-face recognition for IoT applications," *Sensors*, vol. 25, no. 18, p. 5880, 2025.

[32] M. Sohail *et al.*, "Deep Learning Based Multi Pose Human Face Matching System," in *IEEE Access*, vol. 12, pp. 26046-26061, 2024.

[33] H. Lee, S.-H. Park, J.-H. Yoo, S.-H. Jung and J.-H. Huh, "Face recognition at a distance for a stand-alone access control system," *Sensors*, vol. 20, no. 3, p. 785, 2020.

[34] F. Hidayat, U. Elviani, G. B. G. Situmorang, M. Z. Ramadhan, F. A. Alunjati and R. F. Sucipto, "Face Recognition for Automatic Border Control: A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 37288-37309, 2024.

[35] S. Vosta and K.-C. Yow, "A CNN-RNN combined structure for real-world violence detection in surveillance cameras," *Applied Sciences*, vol. 12, no. 3, p. 1021, 2022.

[36] W. Xu, D. Zhu, R. Deng, K. Yung and A. W. H. Ip, "Violence-YOLO: Enhanced GELAN algorithm for violence detection," *Applied Sciences*, vol. 14, no. 15, p. 6712, 2024.

[37] C. Hua, K. Luo, Y. Wu, and R. Shi, "YOLO-ABD: A multi-scale detection model for pedestrian anomaly behavior detection," *Symmetry*, vol. 16, no. 8, p. 1003, 2024.

[38] A. A. Alsabei, T. M. Alsubait and H. H. Alhakami, "Enhancing Crowd Safety at Hajj: Real-Time Detection of Abnormal Behavior Us-

ing YOLOv9," in *IEEE Access*, vol. 13, pp. 37748-37761, 2025.

[39] K. U. Duja, I. A. Khan and M. Alsuhaibani, "Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks," in *IEEE Access*, vol. 12, pp. 164811-164842, 2024.

[40] J. Lee, B. Bagheri and H.-A. Kao, "A cyber-physical systems architecture for Industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, pp. 18–23, 2015.

[41] Z. Weng, H. Zhuang, H. Li, B. Ramalingam, R. E. Mohan and Z. Lin, "Online Multi-Face Tracking With Multi-Modality Cascaded Matching," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2738-2752, 2023.

[42] B. Li, F. Yang and S. Zhang, "Context-aware risk attribute access control," *Mathematics*, vol. 12, no. 16, p. 2541, 2024.

[43] NVIDIA, "DeepStream SDK," *NVIDIA Developer*, 2023. [Online]. Available: `https://developer.nvidia.com/deepstream-sdk`. Accessed: Oct. 25, 2026.

[44] T. Li *et al.*, "ESMO: Joint Frame Scheduling and Model Caching for Edge Video Analytics," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 8, pp. 2295-2310, 2023.

[45] A. Masoudi and M. Ahmadi, "Auto-Encoders Derivatives on Different Occluded Face Images: Comprehensive Review and New Results," in *IEEE Access*, vol. 13, pp. 195080-195103, 2025.



Boonsong Srichand received his B.Sc. degree in Computer Science from Nakhon Ratchasima Rajabhat University and his M.Sc. degree in Digital Technology from Sukhothai Thammathirat Open University, Thailand. He is currently a Computer Technical Officer at the Administrative Office of the Court of Justice Region 3, Office of the Judiciary, Thailand, where his responsibilities include IT infrastructure management, network administration, and software development. He holds professional certifications, including CCNA, MTCNA, and AWS. His research interests include computer networking, cybersecurity, the Internet of Things (IoT), cloud computing, system programming, and artificial intelligence applications.



Nithizethe Mhuadthongon received his B.Eng., M.Eng., and D.Eng. degrees in Electrical Engineering from King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand, in 2008, 2010, and 2016, respectively. He is currently an Assistant Professor at the School of Science and Technology, Sukhothai Thammathirat Open University (STOU), Nonthaburi, Thailand. His research interests include antenna and wireless communication systems, Cyber-Physical Systems (CPS), the Internet of Things (IoT), and intelligent automation. He has extensive experience in both teaching and research in these areas.