Naif Arab University for Security Sciences
Journal of Information Security and Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR
JISCR

CMADS v1.0: Crowd Monitoring and Anomaly Detection System Using YOLOv8, ByteTrack, and Vision Transformer

Shouq Ali Alsubaie[1], Norah Mohammed Aldoohan[1], Afnan Falah Alqahtani[1], Taqwa Alhaj[2], and Jong Hyuk Kim[2]*

[1]Department of Cybersecurity and Digital Forensics, Naif Arab University for Security Sciences, Riyadh, Saudi Arabia
[2]Center of AI for Security, Naif Arab University for Security Sciences, Riyadh, Saudi Arabia

Received 29 Dec. 2025; accepted 16 Feb. 2026; available online 11 May 2026

Abstract

Managing crowds and noticing unusual behaviors are important for ensuring public safety at large events, particularly in areas with multiple gates where movement can be crowded and unpredictable. Traditional surveillance systems often fall short in accurately identifying individuals, gauging crowd density, or detecting abnormal activity. To tackle these issues, we propose an AI-based automated crowd monitoring and anomaly detection system, CMADS v1.0, that integrates YOLOv8 for people detection, ByteTrack for robust tracking, and Vision Transformers (ViT) for gender classification. This system can estimate the number of people present, including gender distribution, and detect unusual behaviors such as loitering, sudden surges, or slow movement—all of which are useful for monitoring gate zones at large events. The system achieves 13-15 frames per second and 89% mean Average Precision (mAP) in people detection, with 98.7% accuracy in gender classification. It also successfully identifies loitering behavior with 87% accuracy, abnormal speeds with 90% accuracy, and crowd surges with 92% accuracy. These preliminary results are promising, demonstrating that the system can perform the tasks effectively in real time. The system aims to support security teams by providing detailed information on crowd movements and possible hazards, finally enhancing crowd control efforts.

I. Introduction

Monitoring gate areas at big public events is essential for keeping crowds organized, ensuring everyone stays safe, and maintaining overall security. These gate zones are essential for controlling pedestrian flow, verifying identities, and detecting potential issues early [1]. However, traditional surveillance setups at these locations often lack the intelligence and automation needed to effectively manage unpredictable crowd behavior, particularly during busy periods or emergencies [2]. Relying solely on manual observation can limit response times and increase the risk of problems like overcrowding, bottlenecks, or unnoticed suspicious activity.

As demand for smarter event security grows, recent breakthroughs in computer vision and deep learning are opening new avenues to improve gate-

Keywords: Anomaly detection, ByteTrack, crowd behavior analysis, gender classification, multi-person tracking, vision transformer (ViT), you only look once version 8 (YOLOv8).

Production and hosting by NAUSS

* Corresponding author: Jong Hyuk Kim
Email: jkim@nauss.edu.sa
doi: 10.26735/URHY7443

level surveillance. Deep learning models, such as YOLOv8 (You Only Look Once version 8), are highly effective at accurately detecting and localizing individuals in crowded scenes [3]. Tracking tools such as ByteTrack enable continuous, frame-to-frame motion monitoring [4]. Additionally, Vision Transformers (ViT) offer efficient methods for analyzing demographic characteristics, such as gender, thereby providing a clearer picture of crowd composition at entry and exit points [5].

Traditional video surveillance systems are not equipped to analyze more complex crowd behaviors in real time, particularly at gate entry/exit points where large numbers of people converge. Large-scale event safety and efficiency may be threatened by security personnel's inability to respond swiftly and effectively to emerging threats or traffic surges due to a lack of intelligent systems for anomaly detection and demographic analysis. The aim of this work is to leverage recent advances in computer vision and deep learning to develop a prototype of an integrated, automated, and real-time system for tracking crowds at gates and identifying anomalous activity. The key contributions of this work are as follows:

- YOLOv8 and ByteTrack algorithms are integrated to detect and track people at the gate of an event. The human head and upper body are trained for reliable detection and tracking in crowded conditions.
- A pre-trained Vision Transformer (ViT) model is integrated to provide demographic insights in the crowds, such as gender. Crowd behaviors of loitering, abnormal speed, and crowd surges are analyzed.
- A real-time crowd monitoring and anomaly detection system, CMADS v1.0, is demonstrated, providing real-time displays on a dashboard as well as post-event summary reports to support security teams in event management.

The outline of this paper is as follows. Section two provides a comprehensive literature review on crowd monitoring, including AI-based detection and tracking models. Section three presents details of our crowd monitoring system, including the behavior-detection model. Section four presents the results using custom-collected datasets from an event, followed by the conclusion and discussion.

II. Literature Review

Current object detection and counting models, particularly YOLOv8, have significantly improved the ability to detect people in crowded scenes in real time. According to studies [6] and [7], YOLOv8's high accuracy and speed make it suitable for real-time surveillance applications. However, [8] suggests that, despite these advantages, problems such as occlusion and poor lighting persist in ship detection applications. Hence, accurate tracking is important for consistently identifying objects or individuals across video frames, even when occluded. Algorithms such as ByteTrack, DeepSORT, and OC-SORT have shown strong performance in managing identity consistency and handling occlusion. The potential of integrating YOLOv8 with contemporary tracking frameworks to achieve reliable performance in dynamic environments is emphasized by [9].

Crowd management also requires additional information, such as crowd density and flow, as well as the ability to detect anomalous behavior. [10] emphasized the adaptability of YOLO-based models, though they were concerned about generalization across scenes and real-time efficiency. According to [11] and [12], methods such as trajectory-based analysis and weakly supervised learning yield promising results in detecting anomalous behavior, including loitering, rapid movement, and crowd surges. Despite advances, existing approaches lack integrated processing to support high-level decision-making and exhibit high false-positive rates and limited scalability.

Table I compares key related studies with respect to their data, methods, and findings, showing significant progress in object detection, crowd counting, multi-person tracking, and anomaly detection. However, it lacks an integrated study of these elements within a coherent, real-time system, particularly in situations with large numbers of people and frequent visual occlusions. This work addresses this gap by integrating state-of-the-art AI models, including YOLOv8, ByteTrack, and ViT, into a unified framework for gate-level surveillance

TABLE I
RELATED STUDIES CATEGORIZED BY METHODOLOGY AND KEY FINDINGS

Authors	Year	Datasets	Methodology	Findings	Advantage	Disadvantage	Identified Gaps
Abba, S. et al. [2]	2024	Multiple object datasets in security surveillance scenarios	Comprehensive framework for real-time detection and tracking using machine learning techniques	High effectiveness in real-time detection and tracking, with reduced error rates	Comprehensive framework that integrates detection and tracking, providing a complete solution for security applications	Complexity in implementation may pose challenges for non-specialist users	It is necessary to optimize system performance in low light levels and improve the user interface for simplicity of use
Chen, J. [10]	2024	Diverse crowd datasets (e.g., public spaces)	Overview of crowd counting techniques, including image processing and machine learning	Comprehensive analysis of crowd counting methods and their effectiveness in different environments	Provides a broad overview of existing methods and their applications	Limited focus on specific applications in real-time scenarios	Need for more studies on real-time implementation and performance in dynamic environments
Cheng et al. [8]	2024	OpenImages, COCO	YOLO-World (open-vocabulary detection)	uses few-shot learning to detect invisible object classes in real time	Expands YOLO's ability to detect unknown objects in dynamic environments	May struggle with high-density environments due to complexity	Expand to handle dense crowds while maintaining real-time performance
Ferreira and Basiri et al. [9]	2024	Dynamic UAV datasets	Multilevel target tracking using YOLOv8 and MOT algorithms	Improved tracking accuracy in dynamic environments, robust against occlusion	Effective in real-time, dynamic settings with multiple targets	Challenging in heavily occluded environments	Need for further optimization for complex scenarios and 3D modelling
Jayasingh et al. [15]	2024	Real-time crowd data	Utilized IoT sensors, OpenCV, and YOLO model for real-time people detection	The system demonstrated high accuracy in estimating people density in real-time environments	Accurate system leveraging sensors and deep learning techniques for person detection	Dependent on camera and sensor accuracy; may face challenges in low lighting	Need to improve accuracy in low-light conditions and develop models for diverse crowd patterns
Kwak, N. & Lee, B. [1]	2024	UCSD Ped1, Ped2, ShanghaiTech, UCF-Crime, UMN, Avenue, Subway	Survey of deep learning methods (supervised, unsupervised, partially supervised) for abnormal behavior detection	Comprehensive comparison of deep learning approaches with discussion of strengths, limitations, and research gaps	Covers a wide range of models and datasets, offers valuable insights for future research	Implementation complexity, sensitivity to environmental changes, and challenges in small object detection	Need to enhance detection in low-resolution videos, address noisy labels, and optimize models for real-time scenarios
Liu, Y., Kennedy, L., Amiri, H., & Züfle, A. [16]	2024	GeoLife, Agent-Based Simulation (ATL, NOLA, FVA, BJNG)	Unsupervised anomaly detection using Neural Collaborative Filtering (NCF) on user-POI matrices enriched with spatial-temporal features	NCF outperforms traditional and deep learning baselines in dense datasets, showing strong anomaly detection ability based on user behavior patterns	Effective in sparse and imbalanced data; captures both latent and explicit features; no need for semantic annotations	Performance drops on extremely sparse datasets like GeoLife; relies on sufficient user-POI overlap	Requires improved robustness in sparse real-world settings; future work includes enhancing spatial embeddings and handling noisy data

Authors	Year	Datasets	Methodology	Findings	Advantage	Disadvantage	Identified Gaps
Nguyen et al. [17]	2024	Multi-camera tracking data	Single-stage global association approach for multi-camera, multi-object tracking	The approach enhances tracking accuracy by associating objects across multiple moving cameras	Efficient multi-camera tracking, reducing the need for multiple stages	May struggle with occlusions and complex motion patterns	Needs further improvement in occlusion handling and adapting to complex scenes with high object density
Noghre, G. A., Pazho, A. D., & Tabkhi, H. [18]	2024	UCSD Ped1, UCSD Ped2, ShanghaiTech	Variational Autoencoders and trajectory prediction for anomaly detection	Models normal trajectories to detect deviations in movement patterns	Captures subtle behavior anomalies with focus on human-centric detection	Limited scalability for dense crowds; challenges in real-time performance	Need to enhance performance in crowded scenes and improve processing speed for real-time deployment
Solano-Carrillo, E. et al. [19]	2024	Multiple object datasets from various experiments (details not specified)	Multi-object tracking algorithm using uncertain object detection techniques	Significant improvement in multi-object tracking performance compared to traditional methods, with reduced errors	Ability to handle uncertain detections, leading to more accurate and reliable tracking	May require higher computational resources, affecting performance in real-time applications	Need to improve performance in highly congested environments and to deal with fast-moving objects
Vijayakumar & Vairavasundaram [7]	2024	COCO, VOC, CrowdHuman	YOLO-based object detection	YOLO has evolved from YOLOv1 to YOLOv8, improving speed and accuracy for real-time applications	Efficient for high-density environments, suitable for smart cities	Limited in handling extreme occlusions and low-light conditions	Further improvements in detection speed and accuracy in dense, low-light environments
Wu, P., Zhou, X., Pang, G., et al. [11]	2024	ShanghaiTech, UCSD Ped2, UCF-Crime	Weakly supervised anomaly detection using spatio-temporal prompts	Achieves effective anomaly detection and localization with limited annotations	Reduces labeling effort while maintaining high performance	May suffer in generalization to complex or unseen scenarios	Needs better adaptation for varied crowd behaviors and optimization for real-time large-scale applications
Cao et al. [20]	2023	MOT17, MOT20, DanceTrack	Observation-Centric tracking (OC-SORT) using Kalman filter and virtual observations	Improved tracking accuracy and robustness in occlusion scenarios	Effective in non-linear motion and occlusion	Requires careful parameter tuning	Need for more complex motion models
Fei and Han [21]	2023	Cityscapes, KITTI	Multi-camera deep learning tracking	Integrates detection across multiple camera feeds in transportation systems	Effective for urban areas with extensive surveillance	May not perform well in non-structured environments like events	Expand to non-structured environments and improve cross-camera identity matching

Authors	Year	Datasets	Methodology	Findings	Advantage	Disadvantage	Identified Gaps
Hamam Mokayed et al. [13]	2023	videos simulating a shopping mall entrance scenario.	YOLOv3, DeepSORT ,TensorFlow.	The system achieved an accuracy of 91.07% using the standard YOLOv3 model	Accuracy, Real-Time Processing: Using YOLOv3-tiny, Compatibility	Computational Demands, Occlusion Issues, Limited Hardware Support	Low Resolution Issues Limited Model Enhancements
Sudharson et al. [6]	2023	Headcount and activity images	YOLOv8 model for proactive headcount and suspicious activity detection	The model accurately detects headcount and identifies suspicious behaviors	Real-time detection and high accuracy for headcount and activity monitoring	Limited by challenging lighting and complex backgrounds	Need to improve model robustness under variable lighting and background conditions
Xie et al. [22]	2023	Deep learning-based detection using PVNet and ByteTrack	mAP@.5 of 95.2%, superior to YOLOv8s	High detection accuracy with fewer parameters	Limited testing in extreme weather conditions	Need for more robustness in adverse environments	Need for more robustness in adverse environments

in large-scale events, thereby providing a baseline for future performance evaluation.

III. Crowd Monitoring & Anomaly Detection System

The system is intended to observe how crowds move and gather during major public events such as festivals and conferences, using footage from surveillance cameras focused on entry and exit points. It aims to provide detailed reports on crowd behavior. The main functions of the system are to estimate crowd counts and densities, classify crowds by gender (male/female), and identify abnormal behaviors such as loitering, sudden changes in speed, and unusual group formations. The system is composed of three main components:

1) Detecting and tracking individuals is a fundamental component of this system, utilizing YOLOv8 for detecting objects, and ByteTrack algorithm to ensure accurate tracking through frame-to-frame. Individuals are identified by a unique identification number (ID), and their paths are monitored in real-time. The objective is to monitor every individual in real time, enabling reliable tracking across frames and understanding crowd dynamics.

2) Gender classification, the second component of the system, based on Vision Transformer-based models to process high-resolution pictures and classify the individual as males and females. It examines individual video frames to determine the gender of individuals in the crowd, providing valuable demographic information.

3) Behavioral anomaly analysis is the third component of the system, which monitors movement patterns to detect abnormal behaviors, such as loitering, sudden changes in speed, and unusual gatherings

A. System Architecture

It processes videos step by step: detecting individuals, determining their gender, tracking multiple individuals simultaneously, counting individuals, sending alerts if needed, and identifying anomalous patterns. All these components work together to accurately identify crowd movement and flag anomalous behaviors, such as individuals remaining in one spot for too long, sudden rushes, or moving much faster than usual.

Fig. 1 depicts the system's overall architecture and data flow. It features a modular architecture that enables in-depth analysis and intuitive visualizations of individual behaviors in crowded

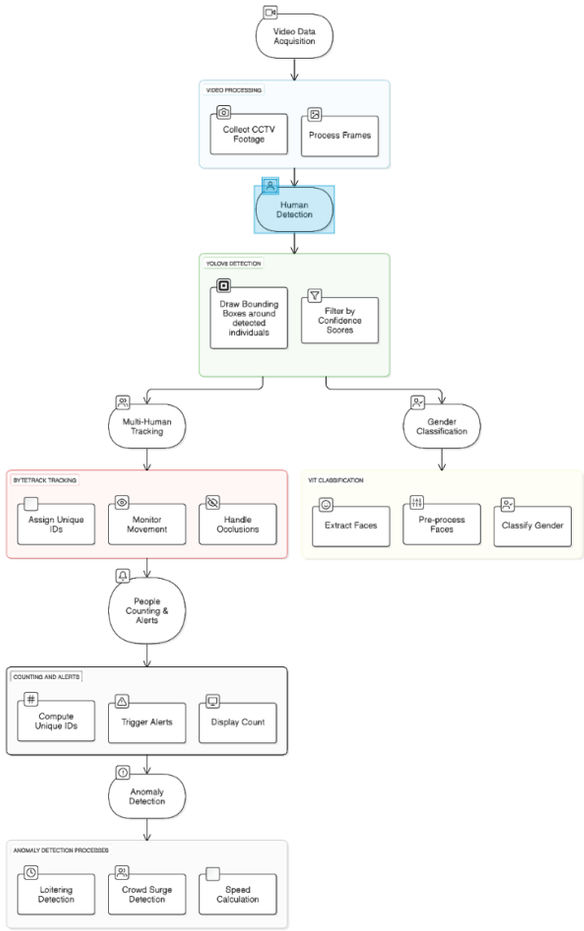

Fig. 1. The overall architecture and data flow of the gate-level crowd monitoring and anomaly detection system (CMADS).

environments. It integrates advanced computer vision and deep learning methods, using YOLOv8 for object detection, ByteTrack for tracking individuals' movements, and a Vision Transformer for gender classification. This combination provides a strong and detailed approach to understanding behavior in high-density settings. It produces videos that include annotations, interactive dashboards, and comprehensive reports. These tools help analyze data and inform better decisions for any intervention.

B. Data Processing

The system begins by capturing video footage from security cameras at the venue's main entrances and exits. These videos are the primary source of information for analysis. The footage is processed frame by frame using OpenCV's frame capture tools. The images are resized to be compatible with YOLOv8 and ViT pipelines. The timestamps for each frame are also recorded to analyze motion behaviors, such as how fast someone is moving or how long they linger.

The YOLOv8 model was trained on a custom dataset focused on head detection, including annotations of heads and upper bodies in highly crowded settings. These annotations are from crowded places where many people congregate. For the gender classification task using the ViT model, we compiled a balanced dataset of face images of men and women. The photos were taken from various angles and under different lighting conditions to reflect real-world variability. The system was tested using a recorded video simulating a busy, crowded environment, allowing us to evaluate its ability to detect, track, and accurately determine gender under such conditions. The facial data was not stored in the system, and only the necessary information was retained to comply with the General Data Protection Regulation (GDPR).

The faces and upper bodies were cropped from the images, then normalized with the ViT image processor to ensure accurate gender classification. Next, to monitor specific gate-zone areas, a map image was overlaid on the input video, displaying the number of people in each zone and updating their positions in real time based on object-detection boxes. This preprocessing converts the raw video into a structured format that's much easier to analyze. It enables the detection of movement patterns, the identification of issues such as crowding or loitering, and the groundwork for more advanced behavior analysis.

C. Human Detection Using YOLOv8

The crowd recognition phase is built on top of the YOLOv8 model, which is renowned for its fast, accurate, real-time object detection capabilities [6]. This system loads a custom-trained weight file and deploys the model using the Ultralytics library. It was trained in Google Colab on a dataset specifically annotated for head and upper-body detection, which are more visible and reliable in crowded environments such as entrances and

exits. The model generates bounding boxes for each recognized individual using the region's upper-left and lower-right coordinates. These detections are filtered by the model's confidence scores, retaining only detections with confidence scores at or above a specified threshold. Additionally, the Intersection over Union (IoU) metric quantifies the overlap between bounding boxes across frames, thereby ensuring temporal consistency in subsequent tracking stages. This is calculated using the formula.

$$LoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

This step produces critical information, notably the bounding box coordinates. generates the required track IDs for the following modules and predicts the class (in this example, "person") and confidence ratings. Overall, this stage serves as the foundation for all subsequent system components, since any detection error will directly affect tracking precision, gender categorization, and behavioral analysis.

D. Multi-Person Tracking Using ByteTrack

Once YOLOv8 detects individuals in each frame, ByteTrack module performs data association to maintain consistent tracking across frames. It links each detection to subsequent frames using spatial and temporal data, ensuring that even if an individual is briefly out of view, they can be re-identified when they reappear. The system stores a history of each tracked individual's movements. This history is presented as a collection of tuples, each including the individual's center coordinates and the matching timestamp. This monitoring history is critical for calculating metrics like speed and movement direction and even forecasting future movement.

$$H_i = \{(x_1^i, y_1^i, t_1), (x_2^i, y_2^i, t_2), \dots, (x_n^i, y_n^i, t_n)\}$$

where H_i denotes the history of person i, and each tuple contains the coordinates *(x,y)* of the person's center and the timestamp t when the detection occurred.

One of ByteTrack's key strengths is its ability to manage partial occlusions. In crowded environments, items may be obscured or overlap. ByteTrack tracks the same person even when they are temporarily obscured, ensuring no identity loss between frames. The system calculates each individual's instantaneous speed using their documented tracking history. This is done by calculating the Euclidean distance between consecutive center locations and dividing it by the time difference between the frames.

$$v_k^i = \frac{\sqrt{(x_k^i - x_{k-1}^i)^2 + (y_k^i - y_{k-1}^i)^2}}{t_k - t_{k-1}}$$

where v_k^i is the instantaneous speed at t_k of human i, and (x_k, y_k) and (x_{k-1}, y_{k-1}) are the center coordinates of the tracked individual in two consecutive frames with timestamps t_k and t_{k-1}.

ByteTrack is designed to be computationally efficient, enabling real-time tracking even when tracking a large number of objects concurrently. Its form makes it ideal for applications that demand rapid processing, such as surveillance, crowd monitoring, and self-driving cars. This accurate tracking and motion analysis allows the system to not only monitor people's movements over time but also analyze their behavior, detect irregularities, and forecast outcomes, which is essential for applications such as crowd control, security monitoring, and autonomous systems.

E. Gender Classification Using ViT

In combination with the tracking technique, the system extracts the facial area of each recognized and tracked individual to determine gender. This stage begins with obtaining the bounding box coordinates generated by YOLOv8 for each participant. The region within the box, assumed to contain the facial area, is cropped from the original frame. The cropped image is first converted from BGR to RGB using OpenCV, then converted to PIL format, and finally processed by the image processor associated with the Vision Transformer (ViT) model [5]. A trained ViT architecture designed to classify facial images as male or female. Because of its Transformer-based design, which allows it to identify fine-grained patterns in image attributes, ViT is particularly well-suited for this function.

ViT architecture consists of several key components. Following that, the picture is separated into non-overlapping patches and flattened into a token-like 1D vector. Typically, these patches are 16 by 16 pixels. Transformers employ a linear embedding layer to convert these flattened patches into a high-dimensional space. Transformers process tokens in parallel and do not record spatial connections by default; thus, position encoding is applied to the patch embeddings to maintain spatial information. After being enhanced with position encoding, the patch embeddings are passed through several Transformer encoder layers. These layers employ self-attention techniques to learn relationships between patches, which help the model understand the image's broader context and identify complex relationships. The output from the Transformer layers is then sent to a classification head. The final class prediction is often made using a fully connected layer. The model first determines whether a male or female face is present in the input image before classifying the gender.

ViT's capacity to recognize minute details in facial pictures makes it especially well-suited for gender categorization. The model can focus on key facial features such as the eyes, nose, and jawline, and understand their spatial relationships and context thanks to ViT's self-attention mechanism. This is essential for differentiating between male and female traits, as even minor variations in facial structure must be observed. Because ViT can display both local and global dependencies in a single image, it is helpful for this purpose. Additionally, employing pre-trained ViT models eliminates the need for intensive training while allowing for more accurate and effective classification.

Once the input passes through the model, it returns a vector of raw logits for each class (male and female). These logits are then transformed into probabilities using the SoftMax function:

$$\mathrm{P_i} = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where z_i is the raw logit for class, and Pi is the corresponding probability for human (i). A person is classified as male if $Pi > 0.5$ or female otherwise. This classification result is immediately added to the system's real-time statistics, updating the counts of males and females in each monitored area. These counts are later used to compute demographic distributions. The clarity of facial features within the image bounds and the quality of the recorded face region determine this module's accuracy. By providing demographic data that facilitates decision-making, particularly for determining gender distributions in densely populated areas, gender categorization greatly increases the system's analytical utility. This step is essential for generating final reports that include gender ratios for each surveillance area.

F. Crowd Counting and Alert Generation

After individuals are identified, tracked, and classified, the system proceeds to the people-counting and alert-generation process. This module is used to calculate occupancy and crowd density in the areas under observation. The unique IDs collected during the tracking phase form the basis for the counting method. By tracking previously registered IDs, the system ensures that no one is counted more than once. Each ID represents a separate individual. When a new unregistered ID is detected, the system updates the general demographic data and increases the population in that region (e.g., Region A or Area B). This strategy reduces duplication and ensures accuracy, particularly in dynamic situations in which individuals may leave and later return to the stage.

The count data are used to compute the occupancy ratio, defined as the current count divided by the zone's maximum capacity. This percentage is used to track crowd density in real time and assess whether a given area is approaching or exceeding safe capacity limits. Every system component has a preset crowd threshold, such as fifty people. It immediately triggers a visual alarm when the population in a specified region exceeds a predefined threshold. Using OpenCV, this notice is directly superimposed on the video feed, instructing viewers to move to a less congested area (e.g., from region A to

location B). Additionally, a time variable is used to implement a flashing system that emphasizes the alarm at regular intervals.

This module also contributes to generating summary reports by recording the total counts of males and females, as well as the combined count, for each area. These statistics are continuously collected for operational and safety purposes, and they may be evaluated during or after an occurrence. This component transforms raw detection and tracking data into actionable insights, supporting security officers and event planners by enabling crowd management, identifying potential danger zones, and sending timely messages to mitigate traffic congestion.

G. Behavioral Anomaly Detection

This module analyzes mobility behavior and identifies aberrant patterns in monitored zones using data from detection and tracking components. It employs three primary mechanisms: loitering detection, anomalous speed detection, and crowd-surge detection. Each mechanism is based on temporal and geographical data derived from the history of movement of each monitored person.

First, loitering is defined as a person remaining in a place for a period exceeding the permitted time, for example, 300 seconds. The time is calculated using the difference between the current time of stay and the time of entry into the place.

$$T_{dwell} = t_{current} - t_{entry}$$

A person is flagged for loitering if they exceed the configured threshold. Second, abnormal speed behavior is detected by calculating each individual's movement speed and comparing it with the group average. If a person's speed exceeds the group average speed by more than 20%, the behavior is considered anomalous. Lastly, crowd surges are detected when a large number of people enter a specific zone within a short time frame. This is evaluated by analyzing the number of new entries within a defined time window (e.g., one second)

This module is one of the most difficult and critical components of the system. It is integrated into all previous stages and is transformed into self-control indicators. All abnormal behaviors are color-coded by changing the color of their bounding boxes in the video: green for normal activities, blue for high-speed anomalies, and red for loitering. This color-coded visual tool helps operators identify potential threats and respond promptly.

IV. Results

The system's performance is assessed using both a publicly available dataset and a custom-built dataset.

A. Datasets and Experimental Setup

The first dataset is a publicly available dataset [13] recorded in a shopping mall environment, consisting of surveillance-style videos of entrance and exit gates with moderate to high crowd density. Crowd density ranges from sparse to highly congested, enabling evaluation under different flow conditions. As the AI detection model is pretrained on the dataset, 8,000 frames are used to test human detection and gender classification (3,800 males and 4,200 females). It includes challenging visual characteristics such as partial and full occlusions, variations in lighting conditions (indoor/outdoor illumination and shadows), and different camera viewpoints typical of fixed CCTV installations. These characteristics were intentionally included to reflect real-world deployment environments. The second dataset is custom-collected at an outdoor event to simulate a realistic environment and comprises 12,000 frames (9,600 for training and 2,400 for testing). It also includes 5,500 males and 6,000 females.

The indoor mall dataset was used exclusively for testing to assess generalization, while the custom outdoor dataset was used for both training and testing. Gender statistics and crowd density levels are reported as approximate values derived from labeled samples and observed scene conditions. This setup accurately reflects the experimental conditions under which all results in the manuscript were obtained. For human detection, a custom head/upper-body detection dataset was used to train the YOLOv8 model. This choice was made

because head and upper-body features are more visible in dense crowds than full-body features. The dataset was split into training and testing sets following standard practice, with the majority used for training and a held-out portion used for validation and testing. For gender classification, a balanced face image dataset was prepared, containing male and female samples captured at different angles and under varying lighting conditions. This dataset was used to fine-tune and evaluate the Vision Transformer model. All experiments were conducted on a standard workstation equipped with a dedicated GPU (NVIDIA Quadro 16GB) to support real-time deep learning inference. GPU acceleration was used during both model training and testing phases to ensure efficient processing. The system was implemented in Python, using frameworks such as Ultralytics YOLOv8, OpenCV, PyTorch, and the Vision Transformer model. The evaluation was primarily conducted on the described dataset to ensure controlled and consistent benchmarking of detection, tracking, gender classification, and anomaly detection performance. In addition, the system was tested on recorded CCTV-style video streams that simulate real gate surveillance environments. While direct deployment on live NAUSS CCTV infrastructure was not conducted due to access and privacy constraints, the experimental videos closely resemble NAUSS gate layouts and camera perspectives, providing indicative performance results.

B. Crowd Detection and Tracking Performance

We used mean average precision to assess system performance, as it accounts for both precision and recall across multiple thresholds, providing a fair and comprehensive estimate of detection accuracy [14]. The developed system reliably detects humans for the shopping mall dataset [13] as shown in Fig. 2. On the custom-collected dataset, it achieved 89% mAP, demonstrating strong accuracy in identifying individuals under crowded conditions, as shown in Fig. 3. As the lighting darkened, the mAP decreased to 83%, indicating reduced sensitivity in low-light conditions as summarized in Table II.

For tracking, the ByteTrack module assigns a unique identity (ID) to every detected person and updates their trajectory across frames. This consistent ID labelling enables the system to maintain the continuity of each individual's movements, even when visual interruptions happen due to occlusion. The system preserves a tracking history for each ID as a list of tuples containing center coordinates and timestamps. This historical data is used not only for tracking visualization but

Fig. 2. Real-time head detection example using a public dataset in a shopping mall environment

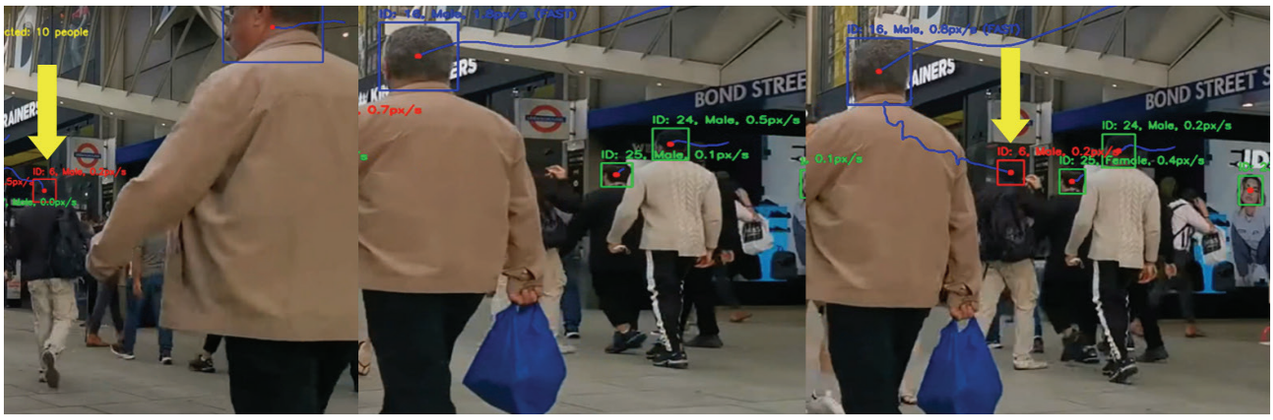

Fig. 3. Demonstration of continuous tracking under occlusion. A man (shown under a yellow arrow, left) is occluded (middle) and then reappears (right), with the same ID number.

TABLE II
DETECTION PERFORMANCE USING THE CUSTOM DATASET

Metric	Value
Precision	90%
Recall	88%
Mean Average Precision (mAP)	89% (good lighting), 83% (low light)

TABLE III
TRACKING PERFORMANCE

Metric	Value
ID switch rate (Moderate density)	2.1 per 100 frames
ID switches rate (High-density)	4.8 per 100 frames
Tracking continuity (Partial occlusion)	91%
Tracking continuity (Complete occlusions over 2secs)	76%
Re-identification accuracy post-occlusion	85%

also for calculating movement-related metrics such as velocity, direction, and behavioral patterns.

The integration of YOLOv8 and ByteTrack in the proposed surveillance system enables robust multi-person tracking in dynamic and crowded environments, as shown in Fig. 3, in which a man shown under a yellow arrow disappears after being obstructed by another individual moving in front of him. The man reappears after being obstructed. Thanks to the robustness of the tracking algorithm, the man was tracked continuously rather than being counted as a new person, demonstrating the continuity across frames, even in the presence of visual interruptions such as partial or full occlusions. Each individual's trajectory is stored as a list of tuples containing center coordinates and timestamps, forming a tracking history that supports both real-time visualization and downstream analysis of movement patterns, including velocity, direction, and behavior.

The multi-person tracking performance is summarized in Table III. In moderately dense environments with clear visibility, the system performed reliably, achieving 2.1 ID switches per 100 frames, indicating strong identity preservation. However, in high-density settings, where visual overlaps are frequent, the ID switch rate rose to 4.8, highlighting the system's sensitivity to complex visual clutter. Despite these challenges, tracking continuity remained strong: during partial occlusions, the system preserved continuity in 91% of cases, aided by the Vision Transformer's spatial reasoning. In scenarios involving complete occlusions lasting over 2 seconds, continuity dropped to 76%, while post-occlusion re-identification accuracy remained at 85%, demonstrating the system's resilience and ability to recover from temporary visual loss.

C. Gender Classification Performance

The system offers more than simple detection by providing deeper insights into crowd demographics, including gender, which is important

for effective crowd management in the Kingdom of Saudi Arabia. The system showed an impressive 98.7% accuracy, thanks to the Vision Transformer algorithm, enabling demographic analysis for improved crowd understanding, as illustrated in Fig. 4, which annotates the detected bounding boxes with gender, ID number, and speed.

Fig. 4. Real-time gender classification (in green label over the bounding boxes) using custom-collected datasets.

D. Abnormal Behavior Detection Performance

The behavior and anomaly detection module in the proposed gate-level monitoring system plays a vital role in identifying potential security threats and managing crowd dynamics in real-time. This module operates on temporal and spatial patterns captured in individuals' tracking histories.

Loitering detection is implemented by maintaining a per-person track history consisting of center coordinates and timestamps. For each detected person, the system calculates the dwell time as the difference between the current timestamp and the timestamp of their first recorded appearance. If this dwell time exceeds the configured threshold (300 seconds in this work, but this value can be adjusted for specific scenarios), the individual is flagged as loitering and highlighted with red bounding boxes in the visual output. The experimental results on the custom dataset show an accuracy of 87%, sufficient to help security teams spot individuals lingering in sensitive areas, as summarized in Table IV.

TABLE IV
BEHAVIOR RECOGNITION AND DEMOGRAPHIC ANALYSIS

Task	Accuracy
Gender classification	98.7%
Behaviour detection (walking, stopping, loitering)	88%
Loitering detection	87%
Abnormal speed detection	90%
Crowd surge detection	92%
Heatmap crowd density	94%

Fig. 5. Detection of abnormal speed (left) and surge detection near an event gate (right).

The abnormal speed detection module achieved 90% accuracy by using the ByteTrack-based trajectory history to compute each person's instantaneous speed and then compare it with the group's average speed. Fig. 5 (left) shows an example of abnormal speed detection. If a person's speed exceeds the average by at least 20%, they are classified as exhibiting abnormal behavior. This enables the system to detect running, fleeing, or erratic motion, which is useful for identifying panic or security threats.

Crowd surge detection tracks the number of new individuals entering a zone over a sliding time window. It maintains a list of individuals counted over time and computes the change within the surge window. A surge is flagged if the count difference exceeds the surge threshold, alerting to a sudden influx of people that could indicate bottlenecks or emergencies. The experimental

results showed 92% accuracy in identifying crowd congestion using movement tracking and spatial analysis, which is promising. Fig. 5 (right) illustrates this surge detection, in which the system uses color-coded bounding boxes (in the top-right corner of the image) and real-time alerts to help operators monitor and respond to these behaviors. Red is used for loitering, blue for speed anomalies, and general crowd surge warnings are displayed as overlay messages on the screen. These features enable proactive interventions to prevent bottlenecks or to address suspicious behavior before it escalates.

Fig. 5 (right) shows two heatmaps in the upper-left corner, visualizing crowd counts as colored boxes across gate entrances. The system achieved 94% accuracy, which is also quite promising. This capability provides insight into how people move through and gather around entrances, which is critical for optimizing space utilization and ensuring smooth flow. Consequently, the heatmaps transform raw data into actionable insights, enabling event organizers or security teams to make more informed decisions about crowd management.

V. Conclusions

A real-time crowd-monitoring system was implemented and demonstrated to enable fully data-driven event management at the gate, which is typically the most congested area at events. The system integrates YOLOv8 for accurate human detection, a Vision Transformer for gender categorization, and ByteTrack for robust tracking in cluttered environments, analyzing crowd gender and potential abnormal behaviors, including loitering, speed, and surge. Experimental results on custom datasets showed promising performance, with 90% detection precision, 91% tracking continuity under partial occlusion, 98.7% gender classification accuracy, and 88% average behavior detection (loitering, speeding, and crowd surges). Based on these, the system provides real-time visual alarms on the dashboard, which security teams can use to respond. The system still shows some limitations under extreme occlusion and in low-light settings, requiring further investigation, including the integration of thermal imaging, improved re-identification techniques, adaptive learning models, and thorough performance evaluation. In addition, the system requires field evaluation by security teams to assess its suitability for real-time, large-scale event management.

Acknowledgment

The work was conducted as part of the graduation project of the Master of Cybercrimes and Digital Forensic Investigation program at Naif Arab University for Security Sciences.

Funding

This article received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflicts of interest.

References

[1] N. Kwak and B. Lee, "Deep learning applied abnormal human behavior detection in video surveillance systems – A survey," *Int. J. Contents, vol.* 20, no. 4, pp. 84–95, 2024.

[2] S. Abba, A. M. Bizi, J. A. Lee, S. Bakouri, and M. L. Crespo, "Real-time object detection, tracking, and monitoring framework for security surveillance systems," *Heliyon,* vol. 10, no. 15, 2024.

[3] Ultralytics, "YOLOv8," GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics/blob/main/docs/en/models/yolov8.md

[4] I. Zhang, "ByteTrack: Multi-object tracking by associating every detection box," GitHub repository, 2022. [Online]. Available: https://github.com/ifzhang/ByteTrack

[5] Google, "Vision transformer (ViT) model," Hugging Face, 2023. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224-in21k

[6] D. Sudharson et al., "Proactive Headcount and Suspicious Activity Detection using YOLOv8," *Procedia Comput. Sci.*, vol. 230, pp. 61–69, 2023.

[7] A. Vijayakumar and S. Vairavasundaram, "YOLO-based object detection models: A review and its applications," *Multimedia Tools Appl.,* pp. 1–40, 2024.

[8] X. Cheng et al., "Ship imaging trajectory extraction via an aggregated YOLO model," *Eng. Appl. Artif. Intell.,* vol. 130, p. 107742, 2024.

[9] D. Ferreira and M. Basiri, "Dynamic Target Tracking and Following with UAVs Using Multi-Target Information: Leveraging YOLOv8 and MOT Algorithms," *Drones, vol.* 8, p. 488, 2024.

[10] J. Chen, "Crowd Counting and People Density Detection: An Overview," in Proc. *3rd Int. Conf. Eng. Manage. Inf. Sci. (EMIS 2024)*, Springer Nature, 2024, p. 434.

[11] P. Wu et al., "Weakly supervised video anomaly detection and localization with spatio-temporal prompts," in *Proc. 32nd ACM Int. Conf. Multimedia,* Oct. 2024, pp. 9301–9310.

[12] R. Nasir, Z. Jalil, M. Nasir, T. Alsubait, M. Ashraf, and S. Saleem, "An Enhanced Framework for Real-Time Dense Crowd Abnormal Behavior Detection Using YOLOv8," *Artificial Intelligence Review,* Vol. 58, No. 202, 2025.

[13] Mokayed, Hamam, et al. "Real-time human detection and counting system using deep learning computer vision techniques." *Artificial Intelligence and Applications*. Vol. 1. No. 4. 2023.

[14] R. Padilla et al., "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics,* vol. 10, no. 3, p. 279, 2021. doi: 10.3390/electronics10030279.

[15] S. K. Jayasingh et al., "Integrated Crowd Counting System Utilizing IoT Sensors, OpenCV and YOLO Models for Accurate People Density Estimation in Real-Time Environments," in *Proc. 1st Int. Conf. Cognit., Green and Ubiquitous Comput. (IC-CGU),* IEEE, Mar. 2024, pp. 1–6.

[16] Y. Liu, L. Kennedy, H. Amiri, and A. Züfle, "Neural collaborative filtering to detect anomalies in human semantic trajectories," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Geospatial Anomaly Detection (GeoAnomalies'24),* Atlanta, GA, USA, 2024.

[17] P. Nguyen et al., "Multi-camera multi-object tracking on the move via single-stage global association approach," *Pattern Recognition,* vol. 152, p. 110457, 2024.

[18] G. A. Noghre, A. D. Pazho, and H. Tabkhi, "An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*, 2024, pp. 995–1004.

[19] E. Solano-Carrillo et al., "UTrack: Multi-Object Tracking with Uncertain Detections," *arXiv preprint*, arXiv:2408.17098, 2024.

[20] J. Cao et al., "Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking," arXiv preprint, 2023.

[21] J. Fei and M. Han, "Real-time object detection methods: An evaluation of CMCA-YOLO for surveillance," *Electronics,* vol. 12, no. 4, pp. 222–238, 2023.

[22] H. Xie, Z. Xiao, W. Liu, and Z. Ye, "PVNet: A Used Vehicle Pedestrian Detection, Tracking, and Counting Method," *Sustainability*, vol. 15, pg. 14326, 2023.