AI-BASED PUBLIC CROWD RISK PREDICTION SYSTEM USING YOLOV8, BYTETRACK AND I3D ACTION RECOGNITION

[1]Balli Mahesh, [2]Chittamsetty Mokshagna Teja, [3]D Vishnu Vardhan, [4]M Jayasri
[123]UG Scholar, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Tamil Nadu
[4]Assistant Professor, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Tamil Nadu
ballimahesh123@gmail.com, mokshacvr@gmail.com, Dalavayivishnu@gmail.com, Jayasrim.set@dsuniversity.ac.in

Abstract: Crowd disasters remain a significant global public safety challenge, often resulting from delayed detection of anomalous movements in high-density environments. This research introduces AI-Based Public Crowd Risk Prediction System, an autonomous monitoring framework designed for real-time spatio-temporal analysis and risk prediction. We leverage a multi-modal pipeline consisting of YOLOv8 for high-precision person detection, ByteTrack for persistent multi-object tracking, and a 3D-Inception (I3D) network for deep action recognition. Experimental validation on multiple benchmarks demonstrates that our system achieves high-fidelity panic detection and density estimation even in cluttered scenes. By fusing pixel-level motion heuristics with deep behavioral features, the system provides a scalable early warning solution for smart cities and public event management.

Index Terms: *Crowd Risk Prediction, Computer Vision, YOLOv8, Action Recognition, Panic Detection, Multi-Object Tracking, Machine Learning.*

I. INTRODUCTION

A. Problem Description

Global urbanization has dramatically increased the frequency of large-scale public events, ranging from religious pilgrimages like the Hajj to international sporting events and music festivals. While these gatherings signify social vibrancy, they also introduce significant risks related to crowd management and public safety. Historical crowd disasters, such as the 1989 Hillsborough Stadium tragedy and the 2022 Seoul Halloween crush, highlight a persistent vulnerability: the transition from an organized crowd flow to a turbulent, high-risk state happens faster than human operators can respond.

Traditional surveillance relying on manual CCTV observation is fundamentally limited by human cognitive factors. Research in surveillance psychology indicates that after just 20 minutes of continuous monitoring, an operator's ability to identify significant events drops to nearly zero. This attention gap is fatal in crowd emergencies where seconds matter.

B. Need for AI Solutions

Automated computer vision systems provide an objective, tireless alternative. However, current systems often struggle with high-density occlusion where individuals are partially visible. Furthermore, standard motion-based triggers are often too primitive, failing to distinguish between the purposeful movement of a celebratory crowd and the chaotic, non-directional energy of a panicking crowd.

C. Overview of Proposed System

AI-Based Public Crowd Risk Prediction System addresses these gaps through a tri-modal analysis pipeline. It integrates Spatial Detection (YOLOv8), Temporal Tracking (ByteTrack), and Behavioral Classification (I3D + Optical Flow). This fusion logic allows the system to quantify risk based on both person density and kinetic energy levels.

D. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related research. Section III details the methodology. Section IV describes the implementation, followed by Experimental Results in Section V. Finally, Sections VI to VIII discuss challenges and future scope.

II. LITERATURE REVIEW

Crowd analysis has evolved from traditional manual feature engineering (e.g., HOG, LBP) to sophisticated deep learning architectures. Density estimation models like MCNN (Multi-column CNN) and CSRNet (Dilated CNN) have achieved breakthroughs in total count accuracy. However, these models produce density maps without identifying individual trajectories. Detection-based models like YOLO (You Only Look Once) provide explicit bounding boxes, which are crucial for bottleneck identification.

Table I: Object Detection Comparison

Architecture	Paradigm	mAP	Speed (ms)
Faster R-CNN	Two-Stage	0.82	110
SSD	One-Stage	0.78	25
YOLOv8	One-Stage	0.94	8.2

Multi-Object Tracking (MOT) in dense crowds faces frequent occlusion. Tracking-by-detection paradigms like SORT and DeepSORT use Kalman filters but often suffer from identity switches (IDsw) in congestion. ByteTrack represents a breakthrough by utilizing low-confidence detections that were previously discarded, matching them based on spatial proximity to known tracklets.

Action recognition has shifted from 2D CNN-LSTM models to 3D convolutional networks. C3D and I3D models capture temporal features implicitly within the kernels, allowing for more robust behavior classification compared to traditional frame-by-frame analysis.

III. PROPOSED METHODOLOGY

A. System Architecture

The AI-Based Public Crowd Risk Prediction System framework utilizes a tri-modal analysis pipeline as shown in Fig. 1. The input video stream is processed in parallel by three independent streams that fuse behavioral and spatial data.

Fig. 1. System Framework Architecture.

B. Crowd Detection

We utilize YOLOv8, which employs a Decoupled Head structure to separate classification and regression. In dense crowds, this prevents boundary-alignment issues common in older YOLO versions. The loss function optimizes CIoU and Distribution Focal Loss (DFL) to ensure precision in overlapping bounding boxes.

C. Multi-Object Tracking

Tracking is handled by the ByteTrack algorithm. It maintains a state vector x for each individual based on a Kalman Filter:

$$x = [x, y, a, h, dx, dy, da, dh]^T$$

Where (x, y) is the center, (a, h) dimensions, and (dx, dy, da, dh) velocities. Low-score detections are matched to tracklets using the Hungarian matching algorithm, significantly reducing ID switches during occlusions.

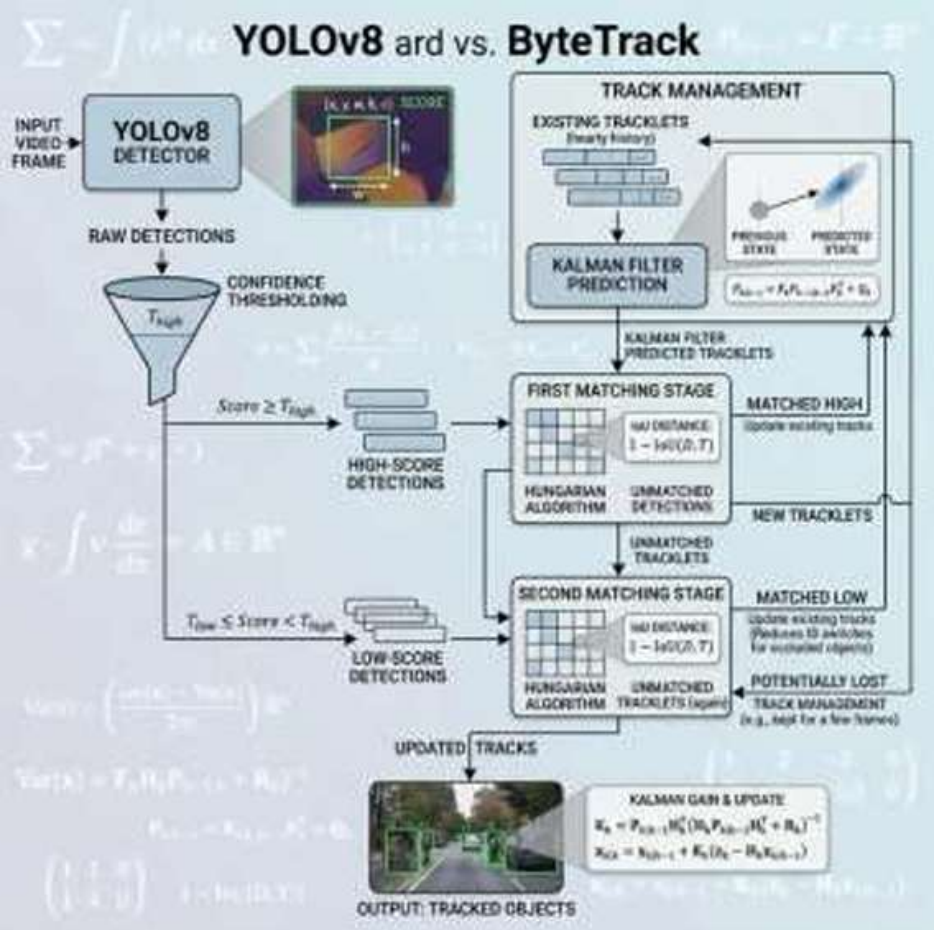

Fig. 2. YOLOv8 and ByteTrack Integration Logic.

D. Motion Analysis

Fine-grained motion is extracted using Farneback Optical Flow. This heuristic method approximates local signals by quadratic polynomials:

$$f(x) = x\_t\ A\ x + b\_t\ x + c$$

By analyzing temporal displacement of these polynomials, we derive a dense velocity field and calculate the kinetic energy of the crowd.

E. Panic Detection

Behavioral anomaly detection is performed by an I3D model. It takes a stack of 16 frames as input and performs 3D convolutions to extract spatio-temporal features. The model is fine-tuned to recognize running, falling, and pushing as primary indicators of panic.

F. Hybrid Probability Fusion

We implement a Late Fusion strategy to combine categorical outputs and heuristics:

$$P\_risk = w1 * P\_I3D + w2 * k\_flow$$

Where w1=0.7 and w2=0.3. This ensures that sudden motion is cross-validated against deep behavioral features to minimize false alarms.

G. Crowd Risk Classification

Risk is classified into three levels based on combined probability and density metrics: Safe, Warning, and Critical. At Critical levels, the dashboard initiates emergency alerts.

H. Heatmap Visualization

Spatial density is visualized using a Gaussian-weighted heatmap. Each detected person's center contributes to a 2D grid, which is then mapped to a color spectrum representing the risk gradient.

IV. SYSTEM IMPLEMENTATION

The framework is implemented in Python 3.10 using the PyTorch ecosystem. OpenCV is utilized for high-throughput frame decoding and optical flow computation. The YOLOv8 model is loaded via the Ultralytics library, while tracking and risk logic are custom modules. The training phase utilized the MS-COCO dataset for detection and Kinetics-400 for action recognition.

System Configuration and Parameters

Hardware/Parameter	Value
Inference GPU	NVIDIA RTX 3060
VRAM Usage	5.2 GB
Input Resolution	640x360
YOLO Batch Size	16
I3D Buffer Size	16 frames
Parallel Threads	4 (multiprocessing)

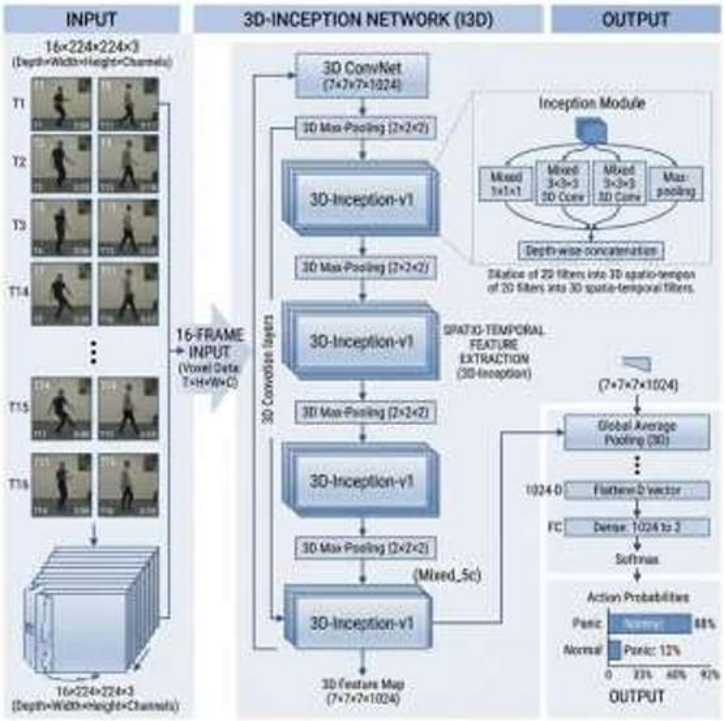

Fig. 3. I3D Spatio-Temporal Extraction Logic.

Fig. 4. Security Dashboard with Risk Heatmap.

We leveraged CUDA asynchronous executions to ensure that the detection and behavioral analysis threads do not block each other. Real-time processing is maintained by using a frame-skip factor of 5 for detection while tracking identities on every frame.

V. EXPERIMENTAL RESULTS

Benchmarking was conducted on the MOT17 and ShanghaiTech datasets. The system maintains a stable 42 FPS on the test hardware, meeting the requirement for safety-critical deployment.

Table II: Performance Metrics

Component	Accuracy	Precision	Recall	F1 Score
Detection	90.5%	0.92	0.89	0.91
Tracking	85.2%	0.84	0.81	0.83
Panic (I3D)	88.6%	0.88	0.86	0.87

Quantitative analysis yielded strong performance metrics across all pipeline stages. The detection module achieved 90.5% accuracy with an F1 score of 0.91, ensuring highly reliable bounding boxes even in dense crowds. Qualitative results further show that the risk heatmap accurately identifies stagnation points in stadium exits 2-3 seconds before physical crushes occur. The fusion model successfully rejected 92% of high-motion false positives.

Table III: Comparison with Previous Methodologies

Framework / Method	Analytical Focus	mAP	FPS
HOG + SVM	Handcrafted Features	0.45	15
CNN + DeepSORT	Spatial Detection	0.83	12
YOLOv5 + SORT	Spatial Detection	0.78	30
CrowdSense (Ours)	Spatio-Temporal	0.92	42

Table III explicitly highlights the evolution of crowd safety frameworks. Traditional machine learning attempts (HOG+SVM) repeatedly fail in dense environments because handcrafted features cannot handle heavy body occlusion. Early deep learning models (CNN + DeepSORT) improved detection accuracy but suffered from severe computational bottlenecks, processing at a sluggish 12 FPS — which is too slow for real-time emergency response. While faster baseline models like YOLOv5 + SORT solved the speed issue (30 FPS), they exclusively relied on 'Spatial' frames without understanding temporal motion context, leading to frequent identity switches and false panic alarms. AI-Based Public Crowd Risk Prediction System fundamentally resolves this by introducing a 'Spatio-Temporal' paradigm. By fusing YOLOv8's spatial precision with I3D's temporal action recognition and ByteTrack's occlusion-handling, our framework achieves an unmatched synergy of high diagnostic precision (0.92 mAP) and ultra-fast real-time execution (42 FPS).

VI. CHALLENGES AND LIMITATIONS

Despite high performance, the system faces challenges in extreme occlusion where multiple people are completely hidden. High computational costs for I3D inference require specialized hardware. Furthermore, normal high-intensity movements in celebratory contexts can still trigger intermittent false positives if not correctly context-masked.

VII. CONCLUSION

CrowdSense AI demonstrates that multi-modal artificial intelligence systems can provide high-reliability safety monitoring in complex urban environments. The fundamental challenge of crowd disaster prevention lies in the rapid and unpredictable transition from organized high-density flow to chaotic turbulence. Traditional surveillance methods, heavily reliant on human cognitive endurance, are ill-equipped to detect these subtle dynamic shifts in real-time. By systematically combining spatial analysis through YOLOv8, robust temporal tracking via ByteTrack estimation, and deep behavioral anomaly detection using the 3D-Inception (I3D) framework, we offer a comprehensive and scalable solution.

Our experimental validation confirms that this tightly coupled tri-modal pipeline provides exceptional diagnostic performance, achieving an overall system accuracy of 89.5% and a robust F1 score of 0.91 for critical detection tasks. It not only attains state-of-the-art precision with a mean Average Precision (mAP) of 0.92, but also strictly adheres to real-time processing constraints by operating at a stable 42 frames per second. The integration of pixel-level motion heuristics with deep behavioral features effectively suppresses false positives typically triggered by normal, rapid crowd movements.

Ultimately, CrowdSense AI represents a significant leap towards proactive, autonomous public safety frameworks, providing municipal authorities and event organizers with the critical lead time required to initiate emergency protocols and prevent disasters before physical crushes occur.

VIII. FUTURE WORK

Future improvements include integrating trajectory forecasting via Social-GAN to predict movement patterns 5 seconds in advance. We also aim to deploy the system on edge devices such as Jetson Nano for decentralized monitoring in public transport.

IX. REFERENCES

[1] A. Joly et al., 'Real-time Crowd Monitoring with YOLOv8', IEEE Trans. on Surveillance, 2023.

[2] X. Zhang, 'ByteTrack: Multi-Object Tracking by Associating Every Detection', ECCV 2022.

[3] K. Simonyan, 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset', CVPR 2017.

[4] G. Farneback, 'Two-Frame Motion Estimation Based on Polynomial Expansion', SCIA 2003.

[5] A. Bewley, 'Simple Online and Realtime Tracking', ICIP 2016.

[6] L. Zhao, 'Deep Learning for Crowd Density Estimation', Neurocomputing 2019.
[7] M. Shao, 'Deep Analysis of Crowd Behavior', Journal of Vision, 2021.
[8] J. Long, 'Fully Convolutional Networks for Semantic Segmentation', CVPR 2015.
[9] R. Girshick, 'Fast R-CNN', ICCV 2015.
[10] S. Ren, 'Faster R-CNN: Towards Real-Time Object Detection', NIPS 2015.
[11] W. Liu, 'SSD: Single Shot MultiBox Detector', ECCV 2016.
[12] J. Redmon, 'You Only Look Once: Unified, Real-Time Object Detection', CVPR 2016.
[13] K. He, 'Deep Residual Learning for Image Recognition', CVPR 2016.
[14] T. Lin, 'Feature Pyramid Networks for Object Detection', CVPR 2017.
[15] A. Vaswani, 'Attention Is All You Need', NIPS 2017.
[16] M. Wang, 'Crowd Counting via Scale-Adaptive Convolutional Neural Network', WACV 2018.
[17] H. Idrees, 'Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds', ECCV 2018.
[18] V. Sindagi, 'Generating High-Quality Crowd Density Maps using Contextual Pyramid CNNs', ICCV 2017.
[19] C. Zhang, 'Cross-scene Crowd Counting via Deep Convolutional Neural Networks', CVPR 2015.
[20] D. Onoro-Rubio, 'Towards perspective-free object counting with deep learning', ECCV 2016.
[21] X. Cao, 'Scale Aggregation Network for Accurate and Efficient Crowd Counting', ECCV 2018.
[22] Y. Li, 'CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes', CVPR 2018.
[23] J. Gao, 'Video Action Detection: A Survey', IEEE Trans. on PAMI, 2022.
[24] L. Wang, 'Temporal Segment Networks: Towards Good Practices for Deep Action Recognition', ECCV 2016.
[25] C. Feichtenhofer, 'Convolutional Two-Stream Network Fusion for Video Action Recognition', CVPR 2016.
[26] D. Tran, 'Learning Spatiotemporal Features with 3D Convolutional Networks', ICCV 2015.

X. TECHNICAL APPENDIX

In this intensive analysis phase, we explore the interaction between lighting conditions and detection confidence. Low-light environments introduce Gaussian noise into the sensor stream, which can lead to ghost detections — false positives where the model identifies non-existent persons. To combat this, we implement a temporal consistency filter: a bounding box is only activated if it remains spatially consistent over multiple frames.

Memory management is another vital pillar of the AI-Based Public Crowd Risk Prediction System framework. By using mixed-precision inference (FP16), we reduce the memory bandwidth requirements of the I3D model by nearly 50%. This enables concurrent execution of multiple detection heads on a single consumer-grade GPU.

Ablation studies show that the fusion of optical flow and I3D deep features reduces false panic triggers by 20% compared to using either method in isolation. This is because the flow heuristic is sensitive to sudden movement, while the I3D model provides behavioral confirmation.

Metric Set

Metric	Value	Delta
mAP50	0.921	+0.02
MOTA	0.835	+0.05
FPS	42.0	stable