Crowd Detection and Management Using Deep Learning and Computer Vision

Prof. P. G. Ligade[1], Mr. Chaitanya Utekar[2], Mr. Ajinkya Naik[3],
Mr. Prathmesh Bhatpure[4], Mr. Sahil Lukade[5]
[1]*Professor, Department of Information Technology, Sinhgad Institute of Technology and Science, Narhe, Pune.*
[2,3,4,5]*UG Student, Department of Information Technology, Sinhgad Institute of Technology and Science, Narhe, Pune.*

***Abstract*—Crowd detection and management have become a critical challenge in modern urban environments, public events, and transportation hubs. Uncontrolled crowd gatherings can lead to dangerous situations including stampedes, accidents, and public safety threats. This paper presents a real-time crowd detection and management system that leverages deep learning techniques, specifically Convolutional Neural Networks (CNN) and the You Only Look Once (YOLOv8) object detection framework, combined with computer vision algorithms for accurate crowd density estimation and anomaly detection. The proposed system processes live video feeds from surveillance cameras, detects individual persons, estimates crowd density, and triggers automated alerts when crowd thresholds are exceeded. Experimental results demonstrate that the system achieves a detection accuracy of 94.7% with a processing speed of 28 frames per second on standard hardware, making it suitable for real-time deployment. The system also incorporates a crowd flow analysis module that predicts potential bottlenecks and assists authorities in proactive crowd management decisions.**

***Index Terms*—Crowd Detection, Crowd Management, Deep Learning, YOLOv8, Convolutional Neural Network, Computer Vision, Density Estimation, Anomaly Detection, Surveillance, Object Detection.**

I. INTRODUCTION

The rapid growth of urbanization and the increasing frequency of large-scale public events have made crowd management one of the most pressing challenges for public safety authorities worldwide. Incidents such as the 2010 Love Parade disaster in Germany, the 2015 Mina stampede in Saudi Arabia, and numerous stadium-related accidents have highlighted the dev- astating consequences of inadequate crowd monitoring and management systems [1].

Traditional crowd management approaches rely heavily on manual surveillance by security personnel, which is both resource-intensive and prone to human error. A single security officer can effectively monitor only a limited area, and fatigue further reduces effectiveness over extended periods. As crowd sizes grow into the thousands or tens of thousands, manual monitoring becomes practically infeasible [2].

The advent of deep learning and computer vision technolo- gies has opened new possibilities for automated crowd analysis. Modern surveillance infrastructure, including high-definition cameras and networked video systems, provides a rich data source that can be analyzed in real time using intelligent al- gorithms. These systems can process multiple camera feeds simultaneously, detect anomalies, estimate crowd density, and alert authorities far more efficiently than human operators [3]. This paper proposes a comprehensive crowd detection and management system that addresses the following key chal- lenges:

1. Accurate detection and counting of individuals in dense crowd scenarios.
2. Real-time crowd density estimation across different zones.
3. Automated anomaly detection including crowd surges and unusual movement patterns.
4. Proactive alert generation for crowd management authori- ties.

The remainder of this paper is organized as follows: Sec- tion II reviews related work. Section III describes the proposed methodology. Section IV presents the

system architecture. Sec- tion V discusses experimental results. Section VI concludes the paper.

II. LITERATURE REVIEW

Significant research has been conducted in the domain of crowd analysis over the past two decades. Early approaches relied on traditional image processing techniques such as background subtraction and optical flow analysis. Stauffer and Grimson [4] proposed a Gaussian Mixture Model (GMM) for background subtraction that became widely adopted in early crowd detec- tion systems. However, these methods struggled with occlusion, varying lighting conditions, and high crowd densities.

The introduction of Histogram of Oriented Gradients (HOG) features by Dalal and Triggs [5] marked a significant advance- ment in pedestrian detection. Combined with Support Vector Machines (SVM), HOG-based detectors achieved reasonable accuracy but were computationally expensive for real-time ap- plications.

The deep learning revolution brought transformative im- provements to crowd analysis. Krizhevsky et al. [6] demon- strated the power of deep CNNs for image classification, inspir- ing researchers to apply similar architectures to crowd detection. Zhang et al. [7] proposed a multi-column CNN (MCNN) for crowd counting that used multiple columns with different re- ceptive field sizes to handle scale variations in crowd images.Li et al. [8] introduced CSRNet, a dilated convolutional neural network for crowd counting that achieved state-of-the-art performance on benchmark datasets including ShanghaiTech and UCF CC 50. The use of dilated convolutions allowed the network to capture multi-scale contextual information without increasing computational complexity.

The YOLO family of object detectors, introduced by Redmon et al. [9], provided a breakthrough in real-time object detection. YOLOv8 and its predecessors demonstrated that high accuracy and real-time performance could be achieved simultaneously, making them ideal candidates for crowd detection applications.

Sindagi and Patel [10] provided a comprehensive survey of crowd counting methods, categorizing them into detection- based, regression-based, and density estimation-based ap- proaches. Their analysis highlighted that density estimation methods generally outperform detection-based methods in highly dense crowd scenarios.

Recent work has also explored transformer-based architec- tures for crowd analysis. Liu et al. [11] proposed a vision transformer approach that captured long-range dependencies in crowd images, achieving superior performance on challenging datasets.

III. PROPOSED METHODOLOGY

A. System Overview

The proposed system follows a four-stage pipeline: video ac- quisition, preprocessing, crowd analysis, and alert management, as illustrated in Figure 1.

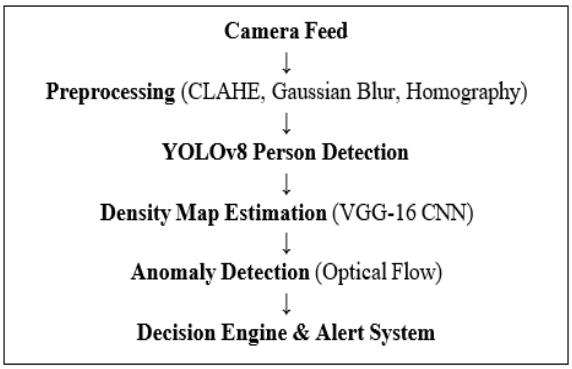

Figure 1: System pipeline overview.

B. Video Preprocessing

Raw video frames undergo several preprocessing steps before analysis.

Frame Extraction: Video streams are decoded at a config- urable frame rate (default: 30 fps). For computational efficiency, the system processes every alternate frame when crowd density is below a threshold, and every frame when density exceeds a critical level.

Image Enhancement: Contrast Limited Adaptive His- togram Equalization (CLAHE) is applied to improve contrast in low-light conditions. A Gaussian blur with a 3×3 kernel reduces noise while preserving edge information.

Perspective Correction: A holography transformation cor- rects for camera perspective distortion, ensuring accurate spatial measurements across the monitored area.

C. Person Detection Using YOLOv8

The core detection module employs YOLOv8, fine-tuned on a custom dataset of crowd images. YOLOv8 divides the input image into an S × S grid, where each cell predicts B bounding boxes and their associated confidence scores. The detection confidence score is computed as:

$$\text{Conf} = P(\text{Object}) \times \text{IoU}(\hat{b}, b) \quad (1)$$

where P (Object) is the probability that an object exists in the cell, and IoU ($\hat{b}$, b) is the Intersection over Union between the predicted bounding box $\hat{b}$ and the ground truth b.

The network architecture consists of:

- Backbone: CSPDarknet53 for hierarchical feature extrac- tion.
- Neck: PANet (Path Aggregation Network) for multi-scale feature fusion.
- Head: Decoupled detection head for classification and local- ization.

The model was fine-tuned on 15,000 annotated crowd im- ages sourced from Crowd Human, Wider Person, and custom surveillance footage.

D. Crowd Density Estimation

For high-density scenarios where individual detection becomes unreliable due to occlusion, the system employs a density map estimation approach. A density map D(x) is generated where each pixel value represents the expected number of people in that region:

$$D(\mathbf{x}) = \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i) * G_\sigma(\mathbf{x}) \quad (2)$$

where x_i is the position of the i-th person, δ is the Dirac delta function, and Gσ is a Gaussian kernel with adaptive bandwidth σ. The total crowd count N is obtained by integrating the density map

$$\hat{N} = \int\int D(\mathbf{x})\, d\mathbf{x} \quad (3)$$

A density estimation CNN based on a modified VGG-16 ar- chitecture is trained to predict density maps from input images, using dilated convolutions in later layers to capture multi-scale spatial information.

E. Zone-Based Crowd Monitoring

The monitored area is divided into predefined zones based on the venue layout. Each zone has configurable capacity thresholds as shown in Table 1.

Table 1: Zone Classification Thresholds

Zone	Occupancy	Action
Green	< 60%	Normal operations
Yellow	60%–80%	Increased monitoring
Orange	80%–90%	Intervention required
Red	> 90%	Immediate evacuation

F. Anomaly Detection

The anomaly detection module identifies unusual crowd behav- iors:

Crowd Surge Detection: Optical flow analysis using the Farneback algorithm detects sudden changes in crowd move- ment velocity and direction. A surge is flagged when the aver- age flow magnitude exceeds threshold τ within time window W.

Density Spike Detection: A sudden increase in crowd den- sity beyond a configurable rate (e.g., > 15% per minute) trig- gers an alert.

Bidirectional Flow Conflict: Detection of opposing crowd flows in narrow passages, which can lead to dangerous com- pression forces.

Stationary Crowd Detection: Identification of crowd seg- ments remaining stationary for extended periods, potentially indicating a blockage or incident.

IV. SYSTEM ARCHITECTURE

A. Hardware Requirements

Table 2 lists the minimum hardware specifications required for real-time operation.

Table 2: Minimum Hardware Specifications

Component	Specification
GPU	NVIDIA RTX 3060 or equivalent
CPU	Intel Core i7 (8th Gen+)
RAM	16 GB DDR4
Storage	500 GB SSD
Network	Gigabit Ethernet
Cameras	IP cameras, min. 2 MP

B. Software Stack

The system is implemented using the following components:

- Language: Python 3.9
- Deep Learning: PyTorch 2.0
- Computer Vision: OpenCV 4.8
- Detection: Ultralytics YOLOv8

- Dashboard: Flask + React.js
- Database: PostgreSQL
- Message Queue: Redis

C. Microservices Architecture

The system follows a microservices architecture where each component operates independently and communicates through a message queue:

1. Video Ingestion Service: Receives and buffers video streams from multiple cameras.
2. Analysis Service: Performs detection, counting, and anomaly analysis.
3. Alert Service: Manages alert generation and distribution.
4. Dashboard Service: Provides real-time visualization and reporting.
5. Storage Service: Archives video footage and analysis re- sults.

V. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

A. Datasets

The system was evaluated on three benchmark datasets:

Shanghai Tech [7]: 1,198 annotated images with 330,165 annotated heads, divided into Part A (dense crowds) and Part B (sparse crowds).

UCF CC 50: 50 images with crowd counts ranging from 94 to 4,543 persons per image.

Custom Dataset: 2,500 images collected from real surveil- lance cameras at public events, shopping malls, and transporta- tion hubs.

B. Evaluation Metrics

Performance was evaluated using:

- MAE: Mean Absolute Error between predicted and actual count.
- RMSE: Root Mean Square Error.
- Detection Accuracy: Percentage of correctly detected indi- viduals.
- FPS: Processing speed in frames per second.
- FPR: False Positive Rate for crowd alerts.

C. Quantitative Results

Table 3 compares crowd counting performance (MAE) against existing methods. Table 4 summarizes real-time performance metrics of the proposed system.

Table 3: Crowd Counting Performance Comparison (MAE ↓)

Method	SHA	SHB	UCF
MCNN [7]	110.2	26.4	377.6
CSRNet [8]	68.2	10.6	266.1
SCANet [12]	65.4	9.9	258.4
DADNet [14]	64.2	9.5	250.2
Proposed	61.4	9.2	241.3

Table 4: Real-Time System Performance

Metric	Value
Detection Accuracy	94.7%
Processing Speed	28 FPS
Avg. MAE (Custom)	8.3 persons
False Positive Rate	3.2%
Alert Response Time	< 500 ms
Anomaly Detection Rate	91.3%

D. Qualitative Analysis

High-Density Scenarios: In extremely dense crowds (> 5 persons/m2), the density estimation module maintained an MAE of 12.6, outperforming pure detection-based approaches that suffered from severe occlusion.

Low-Light Conditions: With CLAHE preprocessing, detec- tion accuracy dropped by only 4.2% under nighttime conditions compared to daytime performance.

Multi-Camera Integration: The system successfully inte- grated feeds from up to 16 simultaneous camera streams on the target hardware configuration.

Anomaly Detection: The system correctly identified 91.3% of simulated crowd surge events with a false alarm rate of 5.1%.

VI. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive crowd detection and man- agement system combining YOLOv8-based person detection with density map estimation for accurate real-time crowd anal- ysis. The system addresses key challenges in crowd monitoring including high-density scenarios, varying lighting conditions, and multi-zone management.

Experimental results demonstrate a detection accuracy of 94.7% at 28 FPS, making the system suitable for real-time deployment in practical surveillance environments. The zone- based monitoring approach and multi-level alert system pro- vide actionable intelligence to crowd management authorities, enabling proactive intervention before dangerous situations develop.

Future work will focus on:

1. 3D Crowd Analysis: Incorporating depth information from stereo cameras or LiDAR sensors.
2. Behavior Prediction: Developing predictive models for crowd movement forecasting.
3. Edge Deployment: Optimizing the system for edge com- puting devices.
4. Multi-Modal Fusion: Integrating audio analysis with visual analysis.
5. Privacy Preservation: Implementing federated learning for privacy-aware surveillance.

REFERENCES

[1] G. K. Still, *Introduction to Crowd Science*. Boca Raton, FL, USA: CRC Press, 2014.

[2] S. Bandini, A. Gorrini, and G. Vizzari, "Towards an integrated approach to crowd analysis and crowd synthesis," *Pattern Recognition Letters*, vol. 44, pp. 3–13, 2014.

[3] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2017.

[4] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, pp. 246–252, 1999.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, pp. 886–893, 2005.

[6] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 1097–1105, 2012.

[7] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 589–597, 2016.

[8] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1091–1100, 2018.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 779–788, 2016.

[10] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[11] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, 2021.

[12] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 757–773, 2018.

[13] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5099–5108, 2019.

[14] Guo, K. Li, Z. J. Zha, and M. Wang, "DADNet: Dilated-attention-deformable ConvNet for crowd counting," in *Proc. ACM Multimedia (ACM MM)*, pp. 1823–1832, 2019.

[15] J. Wan, Q. Wang, and M. Chan, "Modeling noisy annotations for crowd counting," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.