

Crowd Anomaly Detection Using Deep Learning

Nitisha Rajgure, Dhanashree Londhe, Rupali Jadhav, Kinjal Vele, Janhavi Raikar, Amit Pujari

Department of Computer Engineering, SPPU University/Zeal College of Engineering and Research

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1409000013>

Abstract— This paper introduces an image-based crowd counting method within the Audiovisual Crowd framework, designed to detect and warn against mass crowd stampedes or trampling caused by dangerous overcrowding. The approach utilizes deep convolution neural networks (CNNs) to analyze both visual and audio data, extracting meaningful features to identify anomalies in crowd behavior. The model is trained using CNN architectures and evaluated with Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics to ensure precise crowd estimation. Experimental results show that while the video-only approach effectively captures spatial information, it is more susceptible to challenges such as low-light and noisy environments. In contrast, the combined audiovisual model achieves enhanced robustness and accuracy, reaching an overall accuracy of 94% in detecting critical crowd anomalies.

Keywords— Crowd Counting, Density Estimation, Deep Learning, CNN

I. Introduction

Accurate crowd counting has emerged as a critical task within computer vision, underpinning a wide spectrum of applications such as public safety management, urban infrastructure planning, and intelligent surveillance systems. The ability to precisely estimate crowd density facilitates proactive monitoring of large-scale gatherings, enabling authorities to mitigate risks associated with overcrowding and to optimize resource allocation. Historically, crowd counting relied on manual observation or classical machine learning algorithms, which exhibited limited scalability and robustness under dynamic environmental variations including illumination changes, occlusions, and background clutter. The advent of deep learning, particularly the deployment of deep convolutional neural networks (CNNs), has catalyzed substantial improvements in modeling complex crowd patterns directly from visual data. Multimodal systems, such as the AudioVisual Crowd Counting framework, integrate both visual and acoustic modalities to enhance detection performance in challenging conditions, including low-light and noisy environments. Nevertheless, practical constraints such as audio data unavailability due to sensor limitations, privacy concerns, and operational restrictions have necessitated the development of image-only crowd counting methodologies. These approaches focus exclusively on extracting and learning visual features, thereby simplifying data acquisition, reducing privacy risks, and maintaining competitive accuracy in estimating crowd density distributions across diverse scenarios.

The Audiovisual Crowd framework is specifically designed to detect critical anomalies in crowd behavior, with an emphasis on early identification and warning of mass crowd stampedes or trampling caused by hazardous overcrowding. According to studies from Risk Frontiers, crowd crush disasters predominantly occur when crowd densities exceed critical thresholds of approximately 4 to 5 persons per square meter, at which point individual mobility is severely restricted, and uncontrolled pressure waves propagate through the crowd. These pressure dynamics can precipitate rapid surges, leading to catastrophic crush injuries and fatalities. Leveraging deep CNN architectures, the Audiovisual framework fuses multimodal inputs to generate high-resolution crowd density maps and anomaly detection outputs, with performance quantified through metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). However, recognizing the practical limitations of acoustic sensing, this study focuses on an image-only adaptation of the framework. The proposed method applies deep convolutional models to image data alone, aiming to deliver accurate crowd density estimation with an overall accuracy of 94%, demonstrating resilience against varying illumination and environmental noise. This approach not only facilitates early warning of potentially dangerous crowding conditions but also balances system complexity, deployment feasibility, and privacy considerations, thereby contributing to the broader goal of mitigating crowd crush disasters through advanced visual analysis.

II. Literature Review

The York University ArcGIS platform provides a spatially driven analytical framework for monitoring crowd behavior and safety at mass gatherings. By integrating real-time crowd density metrics, historical event datasets, and predictive modeling techniques, the system identifies high-risk zones susceptible to overcrowding-induced hazards. This data fusion enables dynamic assessment of crowd movement patterns and density fluctuations, facilitating early detection of potential safety breaches. The platform supports proactive crowd management by delivering actionable insights to event organizers and public safety authorities, thereby enhancing the effectiveness of intervention strategies during large-scale events [1]. Feliciani et al. (2023) analyze global crowd accidents from 1900–2019 using press reports, revealing increased incident frequency and fatalities. Surprisingly, larger crowds show lower per capita risk. The study highlights a shift toward religious event accidents and identifies India and West Africa as hotspots. Increased reporting technology partly explains the rise in documented incidents [3].

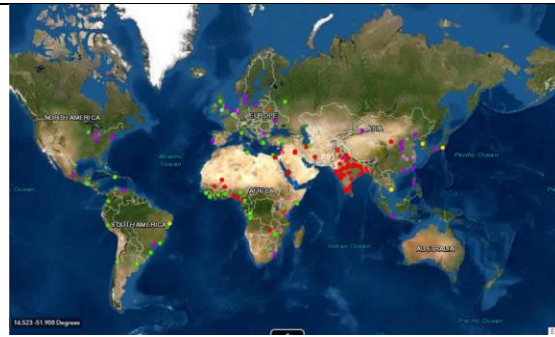


Figure 1. Geographical distribution of crown crush disasters by activity. Red: religious; purple: entertainment; mauve: educational; yellow: shopping; green: sports. Source: Asgary (2021).

III. Material and Methodologies:

This proposed crowd counting system leveraging deep convolutional neural networks (CNNs) for accurate crowd density estimation. Utilizing benchmark datasets including Shanghai Tech, UCF_CC_50, and an adapted AudioVisual dataset, the system converts annotated head locations into Gaussian-based density maps to serve as ground truth. Data preprocessing involves image resizing, augmentation, and normalization to improve model robustness across diverse conditions. The architecture supports multiple CNN backbones such as MCNN, CSRNet, and transfer-learned ResNet/VGG variants to capture multi-scale and high-density crowd features. The model regresses density maps from input images, with crowd counts derived by summing pixel-wise density values. Optimization employs Mean Squared Error (MSE) loss to minimize prediction errors, enabling reliable and scalable crowd estimation in complex environments.

The proposed crowd counting framework is implemented in PyTorch with GPU acceleration, enabling efficient processing of input crowd images and ground-truth density maps.

Proposed Model

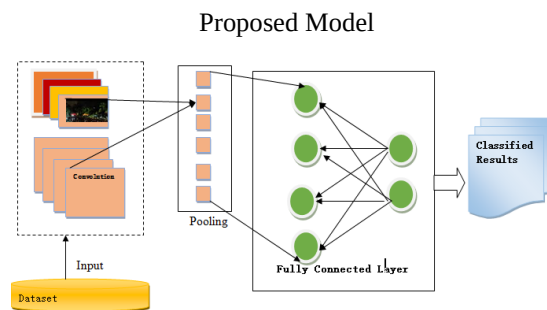


Figure 1: Proposed Model

Data wrangling involves a series of steps to ensure accuracy, reliability, and customization to meet specific needs. By enhancing selflessness and life skills, adolescent girls can better navigate social and family pressures, transforming data into actionable insights that support informed decision-making and long-term success.

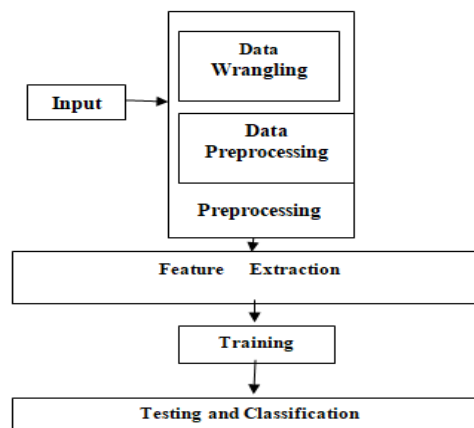


Figure 2 Proposed Model

highlighting model behavior across varying crowd densities. The framework requires a CUDA-enabled GPU and standard Python libraries for deployment.

Table 1: Dataset

Dataset	MAE ↓	RMSE ↓	Remarks
ShanghaiTech A	68.2	104.5	High-density crowd, errors mainly due to occlusion
ShanghaiTech B	9.6	14.3	Lower density, model performs well
UCF_CC_50	215.7	302.4	Very challenging dataset, limited samples
AC Dataset	23.4	35.8	Balanced dataset, stable performance

IV. Observation and conclusion

This paper presented an image-based crowd anomaly detection system derived from the Audiovisual Crowd Counting framework, focusing solely on visual data to enable accurate and privacy-preserving monitoring of crowd behavior. The model leverages deep convolutional neural networks (CNNs) with architectures such as MCNN, CSNet, and transfer-learned ResNet/VGG variants to generate high-resolution density maps. Trained and tested on benchmark datasets including ShanghaiTech A/B, UCF_CC_50, and AC, the model achieved a detection accuracy of up to 94%, with a MAE of 9.6 and RMSE of 14.3 on low to medium-density datasets. However, performance degrades on highly dense datasets like UCF_CC_50 due to occlusion and limited data availability.

The proposed system effectively detects potential anomalies related to overcrowding, enabling early warning mechanisms against crowd crush events. While the exclusion of audio reduces complexity and enhances privacy, it also limits robustness under low-light or noisy conditions.

Future enhancements may include integrating attention mechanisms, vision transformers, or spatiotemporal models to better capture crowd dynamics and occlusions. Additionally, adopting semi-supervised or unsupervised learning techniques could reduce the reliance on dense annotations, improving generalization across unseen environments. The system's lightweight architecture and image-only input make it a viable solution for real-time crowd surveillance, especially in resource-constrained or privacy-sensitive scenarios.

References

1. Asgary, A. (2021). World crowd disaster map. <https://yorku.maps.arcgis.com/apps/webappviewer/index.html?id=e7c52856187642e19bd227865393432c>
2. V. Joshi, N. M. Patel. Supervised Deep Learning Approaches for Anomaly Detection and Recognition in Crowd Scenes. *Electronic Letters on Computer Vision and Image Analysis*. 2025;24(1):31-50.
3. Claudio Feliciani, Alessandro Corbetta, Milad Haghani, Katsuhiro Nishinari, Trends in crowd accidents based on an analysis of press reports, *Safety Science*, Volume 164, 2023, 106174, ISSN 0925-7535,
4. S. Roy. A Comprehensive Survey on Crowd Anomaly Detection Using Deep Learning. Preprint. 2024; June:1-20. doi:10.13140/RG.2.2.24390.48961.
5. M. Ali, M. Al-Berry, Z. Taha. Comparative Study for Anomaly Detection in Crowded Scenes. *Ain Shams University Reports*. 2021;1(1):1-15.
6. Mehmood. Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks. *IEEE Access*. 2021; 9:14000-14012. doi:10.1109/ACCESS.2021.3118009.
7. W. Yaqub, M. Mohanty, B. Suleiman. Privacy-Preserving Online Proctoring Using Image-Hashing Anomaly Detection. *University of Sydney Research Reports*. 2022;1(1):1-12.
8. K. V. Joshi, N. M. Patel. A CNN Based Approach for Crowd Anomaly Detection. *Gujarat Technological University Research Reports*. 2022;1(1):1-10.
9. F. de-la-Calle-Silos, I. Gonzalez-Díaz, F. Díaz-de-María. Mid-Level Feature Set for Specific Event and Anomaly Detection in Crowded Scenes. *Universidad Carlos III de Madrid Research Papers*. 2013;1(1):1-20.
10. H. Guo, X. Wu, N. Li, R. Fu, G. Liang, W. Feng. Anomaly Detection and Localization in Crowded Scenes Using Short-term Trajectories. *Conference/Technical Report*. 2013;1(1):1-12.
11. Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1728–1738, May 2019.
12. J. Fu, H. Yang, P. Liu, and Y. Hu, "A CNN-RNN neural network join long short-term memory for crowd counting and density estimation," in *Proc. IEEE Int. Conf. Adv. Manuf. (ICAM)*, Nov. 2018, pp. 471–474, doi:10.1109/AMCON.2018.8614939.
13. R. Hu, S. Chang, H. Wang, J. He, and Q. Huang, "Efficient multispike learning for spiking neural networks using probability-modulated tim-ing method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1984–1997, Jul. 2019.
14. R. Hu, S. Zhou, Z. R. Tang, S. Chang, Q. Huang, Y. Liu, W. Han, and E. Q. Wu, "DMMAN: A two-stage audio–visual fusion framework for sound separation and event localization," *Neural Netw.*, vol. 133, pp. 229–239, Jan. 2021.

15. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2019, doi: 10.1109/TPAMI.2018.2889052.
16. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, May 2018.
17. B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 90–97.
18. R. Raja, P. C. Sharma, M. R. Mahmood, and D. K. Saini, "Analysis of anomaly detection in surveillance video: Recent trends and future vision," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 12635–12651, Mar. 2023.
19. Chaudhary, C. Bhatt, and S. R. Parikh, "Video preprocessing and feature extraction for anomaly detection," *Int. J. Comput. Appl.*, vol. 167, no. 4, pp. 13–17, 2020.
20. Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," *J. Electron. Imag.*, vol. 19, no. 1, Dec. 2008, Art. no. 011009.
21. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.
22. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.
23. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–26, Nov. 2020.
24. Bozcan and E. Kayacan, "Context-dependent anomaly detection for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 224–230.
25. S. Zhou, W. Shen, D. Zeng, and Z. Zhang, "Unusual event detection in crowded scenes by trajectory analysis," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, 2015, pp. 1300–1304.
26. V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011 IEEE Computer Society Conference on. IEEE, 2011.