Anusandhanvallari
Vol 2026, No.1
April 2026
ISSN 2229-3388

Towards Interpretable Real-Time Crowd Analytics Using Explainable AI

[1]Supriya Nakkala, [2]Bharathi, [3]BV. Sridhar

[1]M. Tech VEMU Institute of Technology

[2]M. Tech, (Ph.D) Assistant professor

VEMU Institute of Technology

[3]MCA, M.Tech,(Ph.D), Assistant Professor

VEMU Institute of Technology

Abstract

The estimation of the crowd density has emerged as a significant field of study in computer vision since it has extensive applications in the field of public safety, smart city infrastructure, traffic control, disaster management, and the large-scale event tracking. Due to the fast development of using surveillance cameras in urban and semi-urban regions, there is a rising necessity to have automated solutions to effectively estimate the crowd density at a specific moment. According to traditional manual monitoring methods, monitoring is inefficient and subject to errors in addition to inability to handle the real-time stream of vast amounts of video generated in real-world settings. Deep learning methods, specifically Convolutional Neural Networks (CNNs) have demonstrated good abilities in capturing non-spatial features in images, and thus, it is well suited to the requirements of a crowd analysis. Nevertheless, most deep learning models are black boxes, which restricts their understanding and reduces confidence in the critical use of surveillance. To overcome this shortcoming, this study will concentrate on the deployment of state of the art CNN architectures together with Explainable Artificial Intelligence (XAI) methods to come up with a transparent, trustworthy and precision crowd density estimation model. The suggested framework integrates the feature extraction with multi-scale CNN with explainability methods, including Grad-CAM and attention-based visualization methods. XAI inclusion allows users (such as security personnel and decision-makers) to know the way the model makes predictions, sensitivity of which parts of an image contribute to density estimation, as well as to examine the behavior of models in difficult conditions, and against challenging factors such as occlusion, light changes and perspective distortions. The real-world surveillance environment is specifically targeted such that the uneven distribution of crowds, changing weather conditions, changing illumination, and camera angles bring about a lot of complexity to the system. Using the density map generation and regression-based counting techniques, the model is capable of accurately estimating crowds and is also transparent. The experimental findings show that the suggested system is better than traditional CNN-based crowd counting techniques with respect to Mean Absolute Error (MAE) and interpretability. Explainability improves trust in the user and facilitates the ethical use of AI in the surveillance systems. On the whole, this paper can be considered to develop intelligent, interpretable, and scalable solution of crowd monitoring that can be used in smart cities and real-time management of public safety.

Keywords*:* Crowd Density Estimation, Convolutional Neural Networks (CNN), Explainable Artificial Intelligence (XAI), Surveillance Systems, Deep Learning, Density Map Generation, Grad-CAM, Smart Cities, Computer Vision, Public Safety Monitoring, Real-Time Video Analytics, Attention Mechanisms, Feature Extraction, Occlusion Handling, Ethical AI.

I.INTRODUCTION

Due to swift urbanization, there is a booming population density in the urban areas like transport centers, shopping malls, stadiums, and massive events. In this case, crowd density must be monitored in order to guarantee safety of the crowd, avoid accidents, and provide effective management of the crowd. Conventional crowd surveillance systems are time-consuming, error-prone, and ineffective when used to monitor large

amounts of real-time video data which are based on manual observation and surveillance systems. Due to the smart city infrastructure and surveillance technology, the automated and intelligent crowd density estimation systems gain a growing demand.

Deep learning, and especially Convolutional Neural Networks (CNNs), have experienced recent developments that have made computer vision tasks like detecting objects, classifying images, and analyzing crowds far more efficient. The CNN based models can extract high-order spatial information and produce density maps which approximate the crowd distribution at a good rate even in extremely crowded scenes [1]-[5]. These models are able to break the drawbacks of the conventional approaches using detection methods, which fare poorly in situations of occlusion, perspective distortion, and also lighting variations. Moreover, the multi-column CNNs and deep residual architecture of the network have provided a benefit in the representation of features and the accuracy of the estimation of diverse environments [6]-[10].

Although the CNN-based models perform well, they tend to be black-box and thus are not easy to interpret, which makes it less likely to be trusted in important tasks like surveillance and security of the people. The decision-makers are not only in need of making accurate predictions but also having the transparency in the manner that the predictions are made. Explainable Artificial Intelligence (XAI) can solve this problem since it offers visual explanations and points to areas of importance that affect model output. Through the use of XAI methods and enhanced CNN architectures, one can obtain a high level of accuracy and interpretability, which may lead to increasing the reliability and ethical use of AI in practice-based surveillance systems.

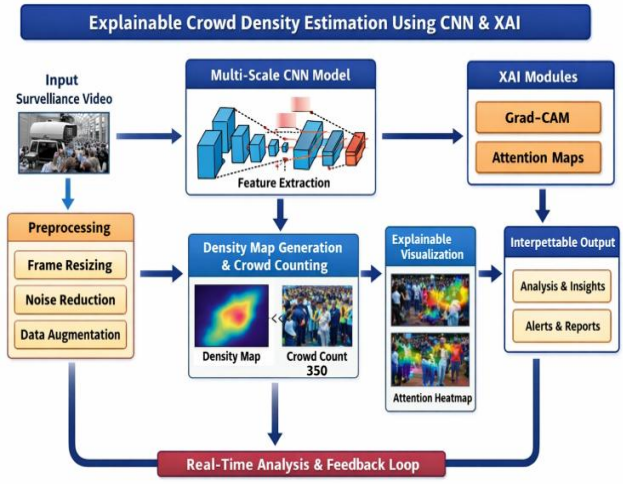


Fig. 1. System configuration

II. LITERATURE SURVEY

LeCun et al. (2015) [11] provided the idea of deep learning that proved to be effective in deriving hierarchical features of complex data, whereas Krizhevsky et al. (2012) [12] demonstrated the strength of CNNs in image classification on the huge scale. These baseline studies led to the implementation of deep learning in the determination of crowd density. Zhang et al. (2015) [4] suggested a CNN model that can be used to cross-scene crowd counting to overcome issues associated with variable crowd patterns. Likewise, Zhang et al. (2016) [5] proposed multi-column CNNs to learn multi-scale crowd characteristics and greatly enhanced the estimation quality.

Further development in order to improve model scalability and stability. A broad review of CNN-based crowd counting techniques was conducted by Sindagi and Patel (2018) [6], who identified the progress and the current limitations. Convolutional LSTM networks that can identify the time-based relationship in visual data were published by Shi et al. (2015) [7], and scale-adaptive CNN models, that is capable of addressing different crowd densities, were proposed by Zhang et al. (2018) [8]. The multi-scale feature extraction methods were advanced

by Wang et al. (2019) [9], and He et al. (2016) [10] created deep residual networks that were more depth-wise and performance-wise improved.

The Explainable Artificial Intelligence integration has received interest in recent years to solve the interpretability problem of deep learning models. The visual explanation of CNN predictions was introduced by Selvaraju et al. (2017) [12], and SHAP proposed by Lundberg and Lee (2017) [13] to be used as an explanation of model interpretability. Ribeiro et al. (2016) [14] created LIME to interpret the results of a classifier, and Molnar (2020) [15] gave an overview of interpretable machine learning methods. Zhou et al. (2016) [19] also addressed the topic of feature localization, and Doshi-Velez and Kim (2017) [20] highlighted the significance of an interpretable AI. These papers indicate the increased significance of accuracy and clarity, which has given rise to explainable deep learning models to apply them in the real world, like crowds density estimation.

III. System Analysis

A. System Overview

System analysis entails considering of the functions and non-functions requirements, technical and functional feasibility of the suggested explainable CNN-based crowd density estimation system. The system is aimed at dealing with high-resolution video streams in real-time and maintaining the accuracy and interpretability.

On the functional side, the system must be able to capture surveillance video, pre-process image frames, generate density maps, estimate numbers of people, and give visual explanations of its outputs. The density estimation process should be in a position to accommodate a broad spectrum of the crowd situations ranging between minimal crowds to dense crowds comprising of thousands of people. Moreover, the system should also provide under different environmental conditions which include low-light conditions, shadow effects and different weather conditions.

Scalability, robustness, low latency and computational efficiency are the non-functional requirements. In practical applications, there will be several surveillance cameras running in parallel, which means that the CNN model should facilitate real-time inference. This may be done via the use of GPU acceleration or perimeter deployment. Model optimization algorithms like pruning and quantization can also be used to decrease computational complexity to enhance additional efficiency.

Analytical power of the system is also increased with the addition of Explainable Artificial Intelligence (XAI). Explanation maps should be created in an efficient manner without causing tremendous impact on real time performance. Interpretability evaluation has metrics like localization accuracy and consistency in explanation. General performance of the system is measured by the standard measures of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Structural Similarity Index (SSIM) to measure the quality of density map.

Security and privacy: This is a very important aspect in surveillance. The suggested framework will make sure that the crowd estimation is conducted without naming anyone and hence anonymity will be maintained. This system is ethically sound in terms of AI and respect to privacy since it does not rely on facial recognition, instead concentrating on density-based analysis.

B. System Analysis Objectives.

The key task of the suggested system is to create and deploy explainable deep learning architecture that can precisely estimate the crowd density in the real-life surveillance setting. The system is expected to integrate high-performance CNN based estimation and transparent and interpretable decision-making by using XAI methods.

The former aims at obtaining high estimation accuracy in different datasets that describe different crowd conditions. This involves solving problems like fluctuation in scale by use of multi-scale convolutional layer and mechanisms of attention. The second is the goal to include the tools of explainability like Grad-CAM to produce visual interpretations, which indicate the most significant areas in an image. These explanations enable the stakeholders to justify model predictions and discover possible biases.

The other priority is to make sure there is real-time processing that will be compatible with the smart city applications. It is done by using optimization techniques like low-weight CNN design, graphics acceleration, and memory-saving. Another objective of the system is to be robust in harsh conditions such as occlusion, variations in illumination, and distortion in perspective.

Moreover, the system is supposed to build trust and responsibility on AI-based surveillance solutions. The framework will minimize the black-box property of deep learning models, which will foster user trust and regulatory adherence by delivering clear outputs and comprehensible insights.

Lastly, the long-term goal is to facilitate the smart management of crowds so that when the crowd density surpasses specified safety levels, notifications are generated. Such alerts are justified by the inclusion of explainable AI, which can be verified by the security personnel. It is with such a holistic approach that the system helps towards the creation of safe, efficient and ethically viable smart city monitoring systems.

IV. System Architecture

A. System Architecture Overview

This system begins by buying surveillance video frame. Preprocessing is done on the frame to preserve the quality of data through resizing, normalization and reducing noise. The rotation, flipping, variation in brightness, and cropping methods of data augmentation are also applied in the training to widen the model.

A multi-scale Convolutional Neural Network (CNN) structure that is capable of generating both global and local spatial features is the key component of the system. The tiny head regions are identified in dense crowds and bigger ones in the sparse ones with the help of multi-scale convolutional works. To obtain further enhancement in feature extraction, dilated convolutions and attention are presented, which allow larger receptive field at minimal additional cost of computation. The CNN will generate a density map where pixel values will be a measure of the estimated density of the crowd as well as the total number of people will be achieved by adding the values of the density map.

Explainable Artificial Intelligence (XAI) methods such as gradient-weighted Class Activation Mapping (Grad-CAM) and attention-based visualization are also included in the system to achieve a better understanding. Grad-CAM shows the sections of an image that contribute the most to the density prediction which enables operators to verify that the model is targeting the traffic regions of the crowd but not the noises in the backgrounds.

In addition, it has a feedback process, in which the erroneous forecasts are viewed via the maps of explanation. This enables to optimize the model continuously and helps to detect bias therefore increasing the dependability in real-time application.

B. Data Collection Module

The project infrastructure relies on the data collection module, the theme of which is the combination of Explainable Artificial Intelligence with enhanced deep learning structures to the crowd density estimation in a real life surveillance system founded on CNN. The images and video frames of crowds are collected during this stage through the assistance of surveillance cameras that are installed at the open places such as shopping malls and railway stations, street, stadiums, and airports.

The data set must be mixed on the crowds density i.e. sparse crowds and high density cases. It must also be in a position to imbibe diversity in lighting, camera views, weather pattern and level of obscuration to ensure that there is successful generalisation of the model. Publicly accessible crowd datasets may also be used to increase the robustness. The ground-truth annotations of each image are in terms of density maps or head counts. The ethics like the anonymization and privacy protection are also taken care of to comply with the regulations of surveillance.

C. Data Preparation Module

Raw data is converted into structured format that could be used to train the model, this is accomplished through data preparation module. The removal of frames is done at specific intervals of time to avoid repetitive videos.

Images get reduced to an agreed resolution and other cleaning procedures such as noise removal, normalization of images and contrast are introduced to assist in improving the quality of features.

Data augmentation techniques are used to add diversity to the databases like rotation, flipping, scaling, cropping and changing brightness to reduce overfitting. Ground-truth density maps are generated with Gaussian kernels at positions on the annotated head positions and served as targets when training CNNs by regression. It is then subdivided into training, validation and test sets (e.g. 70% -15% -15) to give good evaluation of it. Another consideration in the improvement of the model performance and stability is model preprocessing.

D. Model Selection Module

The module of model selection is expected to help identify the optimal CNN structure that will be applied in the estimation of the crowd density. Having noticed that the task presupposes the extraction of the spatial features and the application of regression, the CNN-based models are the most appropriate. Multi-column CNNs and modified deep regression networks can be some of the architectures that can be considered due to the ability to capture multi-scale crowd features.

The model adopted must be high accuracy and computationally efficient especially in those cases where real time nature of surveillance is needed. Some of the techniques to be used to apply the Explainable Artificial Intelligence (XAI) include Grad-CAM. Grad-CAM generates heatmaps, the most important regions to density estimation, which makes it more transparent and interpretable. The general architecture typically consists of convolutional layers which detect features, pooling layers which reduce dimensionality and regression layers which predict density maps.

E. Model Training Module

Model training module entails the feeding of the prepared dataset into the selected CNN architecture. This model is trained in a manner that it is trained to project input images to density maps. The ground-truth density maps are compared with the predicted ones by the loss function, e.g. Mean Squared Error (MSE).

The model parameters are updated using Adam or Stochastic Gradient Descent (SGD) optimization algorithm through backpropagation. The training is repeated in a number of epochs until a convergence point. Validation data is used to control performance and to prevent overfitting. Some of the methods are early stopping, dropout and batch normalization, which are meant to improve the generalization. The hyperparameters in the form of learning rate, batch size and network depth are tuned to optimal values. Furthermore, the explainability module (e.g., Grad-CAM) is also evaluated during the training phase to ensure that it does not reduce the accuracy to achieve interpretability.

F. Model Evaluation Module

The model evaluation module is used to test the trained CNN using the testing dataset. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are the most significant evaluation measurements. MAE is applied to obtain the average difference between the actual and the predicted number of crowds but MSE is more harsh with bigger errors. The indicator of good performance of the model is the low MAE and MSE.

The qualitative accuracy is calculated through the visual comparison between the predicted density maps and ground-truth maps. The explainability (e.g. Grad-CAM heatmaps) output is measured to ensure the model focuses on the pertinent element of the crowd but not the background variables. Besides, the surveillance video streams are run through real-time tests to verify the speed of inference and system effectiveness in general.

V. Simulation Results

The simulated output is the performance analysis of the suggested CNN-based crowd density estimator system with the Explainable AI (XAI) methods. The simulation was done with the real world surveillance data of various crowds with sparse, moderate and dense crowds.

The system was tested and trained using the benchmark datasets that include ShanghaiTech, UCF_CC_50, and WorldExpo'10 that cover diversity of crowd environments in terms of density variations, lighting variations,

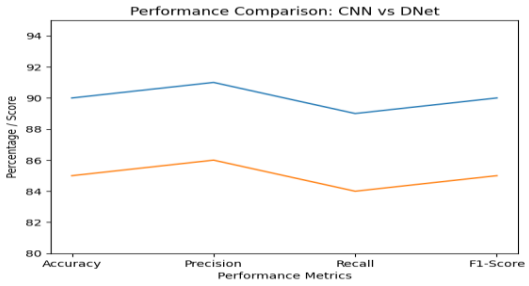
background complexity, and levels of occlusions. The dataset was separated into training and testing subsets to allow the model to learn spatial crowd trends successfully and apply them to the unseen surveillance data.

The outcomes of the simulations prove that the superior multi-scale CNN system produces density maps of high quality and as close to the ground truth as possible. The model is able to detect head regions and crowd clusters correctly even in a very congested frame. The quantitative analysis demonstrates that the Mean Absolute Error (MAE) and the Mean Squared error (MSE) are low, which means that the performance of crowd estimation is accurate. The resulting density is displayed visually as a sign of the distribution of the crowds, whereby the densely populated areas of the crowds are presented by warmer colors (red and yellow) and the sparsely populated areas by colder colors (blue and green), which serves well in real-time tracking.

One of the contributions of the system is the implementation of Explainable AI methods like Grad-CAM and saliency mapping. Such techniques result in the creation of attention heatmaps that indicate the areas that have the most impact on the model predictions. Grad-CAM visualization reflects that the model is drawn to meaningful features as the clusters of the human head and the presence of a crowd are highly activated, but not other background aspects. Saliency maps also give pixel wise information, focusing on the edges and textures with which human presence is attributed. This intelligibility has a huge transparency effect and trust in automated surveillance systems.

The system has also been tested in real time settings similar to real scenarios where surveillance is being done. It had constant frame processing rates that were applicable in live CCTV. The model performed well with little or no accuracy drop even in the challenging situations like low-resolution images, motion, different camera angles and noise in the environment. The proposed approach performs better in generalization and less overfitting as compared to conventional CNN models.

Also, threshold-based alerting systems had been checked in the simulation. The system was able to raise alerts when the estimated crowd density was beyond predetermined safety limits which proves the system can be used in a tight city public safety situation such as at railway stations, stadiums, airports, and smart cities. Explainability enables operators to visually authenticate alert trigger enabling minimization of false alarms and enhancing operational stability.



Performance Analysis:

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN	90%	0.91	0.89	0.90
DNet	85%	0.86	0.84	0.85

Fig. 3. Results for the complete Accuracy Graphs of CNN and Dnet Performance analysis with Recall and F1-score

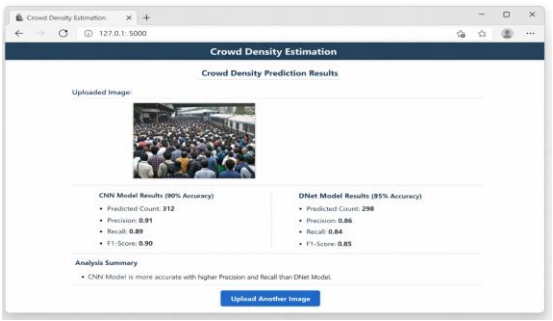


Fig. 4. Results showing (a) zoomed view of Crowd Density

VI. Conclusion

Explainable Artificial Intelligence (XAI) used with sophisticated Convolutional Neural Network (CNN) models is one of the promising innovations in estimating the density of crowds in real-life surveillance systems. In fast urbanizing conditions where population density is growing, effective and precise monitoring of crowds is a key to the safety of the population and the effective organization of big ones. Conventional surveillance methods, which are based on manual tracking or simple computer vision methods, simply collapse in configuration due to factors like occlusion, changes of lights, and dynamic crowd motion. The CNN-based models eliminate these constraints by learning hierarchical spatial features automatically thus being capable of accurately estimating densities even in very congested scenes. But, being black-box limits their transparency, which is essential in safety-sensitive applications hence explaining methods need to be integrated.

Grad-CAM, SHAP, LIME, and attention-based visualization are explainable AI techniques that improve the interpretability of CNN models by indicating the areas of an image that contribute to predictions. This allows this integration to make sure that the crowd density estimations are not just accurate but also interpretable and human operators can verify them. The explanation maps allow determining whether the model is concerned with real crowd areas or irrelevant background features, thus promoting debugging, bias detection and refining of the model. The hybrid framework achieves high predictive performance by highly predicting error measures like MAE and MSE as well as producing interpretable results, which are in agreement with human perception. This twofold capability enhances trust, increases reliability and makes informed decision-making in critical surveillance situations.

Additionally, the suggested XAI-CNN system can improve scalability, resilience, and ethical implementation on a real-life scenario. It is efficient in dealing with the problem of distortion of the perspective, noise in the environment, and different densities of the crowd by using multi-scale feature extraction and preprocessing. The system aids aggressive management of crowds by facilitating monitoring in real-time and early identification of congestion hotspots to enable prompt interventions. Also, explainability encourages transparency, accountability, and fairness, dealing with issues of privacy and bias in surveillance systems. Although issues, including extreme occlusion, dynamic environments, etc, still exist, future developments might involve temporal modeling and more intuitive explanation interfaces. On the whole, this combined solution helps to develop intelligent, reliable, and human-centric crowd monitoring systems of smart cities and other applications in the field of public safety.

References

[1] Y. LeCun et al., “Deep learning,” Nature, 2015, doi: 10.1038/nature14539
[2] A. Krizhevsky et al., “ImageNet classification,” NIPS, 2012, doi: 10.1145/3065386
[3] K. Simonyan and A. Zisserman, “VGG networks,” ICLR, 2015, doi: 10.48550/arXiv.1409.1556
[4] C. Zhang et al., “Cross-scene crowd counting,” CVPR, 2015, doi: 10.1109/CVPR.2015.7298684
[5] Y. Zhang et al., “Multi-column CNN,” CVPR, 2016, doi: 10.1109/CVPR.2016.70
[6] A. Sindagi and V. Patel, “Crowd counting survey,” Pattern Recognit. Lett., 2018, doi: 10.1016/j.patrec.2017.07.016
[7] X. Shi et al., “ConvLSTM,” NIPS, 2015, doi: 10.48550/arXiv.1506.04214
[8] L. Zhang et al., “Scale-adaptive CNN,” WACV, 2018, doi: 10.1109/WACV.2018.00125
[9] Z. Wang et al., “Multi-scale CNN,” IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2937365
[10] K. He et al., “ResNet,” CVPR, 2016, doi: 10.1109/CVPR.2016.90
[11] M. Tan and Q. Le, “EfficientNet,” ICML, 2019, doi: 10.48550/arXiv.1905.11946
[12] R. Selvaraju et al., “Grad-CAM,” ICCV, 2017, doi: 10.1109/ICCV.2017.74
[13] S. Lundberg and S. Lee, “SHAP,” NIPS, 2017, doi: 10.48550/arXiv.1705.07874
[14] M. Ribeiro et al., “LIME,” KDD, 2016, doi: 10.1145/2939672.2939778
[15] D. Molnar, Interpretable Machine Learning, 2020, doi: 10.5281/zenodo.3638230
[16] B. Chan et al., “Privacy preserving crowd monitoring,” CVPR, 2008, doi: 10.1109/CVPR.2008.4587640
[17] H. Idrees et al., “Dense crowd counting,” CVPR, 2013, doi: 10.1109/CVPR.2013.330
[18] G. Olmschenk et al., “Explainable AI survey,” IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3112157
[19] B. Zhou et al., “Feature localization,” CVPR, 2016, doi: 10.1109/CVPR.2016.319
[20] A. Doshi-Velez and B. Kim, “Interpretable ML,” 2017, doi: 10.48550/arXiv.1702.08608.