Semantic Video Analytics Framework using SpinalXNet for Crowd Management in Large Public Events

Shreedevi. P[1], *H. S.* Mohana[2*], *Khalid* Nazim Abdul Sattar[3], *Pradeep* Kumar. S[4], and *Bala* Dhandayuthapani. V[5]

[1]Department of Electronics and Communication Engineering, Navkis College of Engineering, Hassan, India
[2]Department of Computer Science and Engineering, Rajeev Institute of Technology, Hassan, Affiliated to Visvesvaraya Technological University, Belagavi, India
[3]Department of CSI, College of Science, Majmaah University, Al Majmaah, Saudi Arabia
[4]Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Nitte (Deemed to be University), Bengaluru, India
[5]Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Sciences, Shinas, Oman

Abstract. In recent years, crowd management (CM) in large public events, such as concerts, sports gatherings, and festivals, has become increasingly critical for ensuring safety and preventing accidents. This task becomes more challenging owing to the high crowd density, severe occlusions, varying camera perspectives, and complex motion patterns that limit the reliability of traditional surveillance systems. However, conventional monitoring methods struggle to provide a real-time analysis and semantic understanding of crowd behavior. Hence, this study proposes a Semantic Video Analytics Framework (SVAF) for crowd management in large-scale public events. The framework integrates a SpinalNet-based Convolutional Neural Network (CNN) for feature extraction and density map generation to estimate crowd distribution, enabling an accurate assessment of crowd density and spatial patterns. Semantic modeling is applied to map low-level video features onto high-level contextual knowledge. The proposed SVAF-CM was evaluated using the UCF_CROWD_50 and UCF_QNRF datasets, where preprocessing included normalization, perspective correction, and noise reduction. Experimental results show that the SVAF-CM method achieves significant improvements in Mean Absolute Error (MAE) of 0.01 and Mean Squared Error (MSE) of 0.012 compared to the existing DCNN model.

1 Introduction

Among a range of applications, the management of crowds and their associated behavior is one of the most complex and crucial areas, as crowds are often unpredictable and prone to unexpected disasters and crime-related events [1]. The Hajj pilgrimage, one of the world's largest recurring religious events, attracts millions of Muslims annually, creating unique challenges for public safety, emergency response, and logistical coordination [2]. Religious gatherings or pilgrimages that experience a high influx of people are vulnerable to incidents such as stampedes, with events such as Hajj and Kumbh Mela witnessing tragic casualties despite detailed crowd management plans [3]. An effective crowd monitoring and management system should be able to detect fire incidences, early signs of stampedes, congestion, or other preventable disasters [4]. In addition, public transport systems must be evaluated for their responses to different demands and operational policies to ensure reduced waiting times and controlled crowding conditions [5]. The paper presents an application of machine learning and computer vision technologies designed to promote crowd safety in public settings [6]. Advanced technologies are essential for detecting and analyzing crowd density, enabling researchers and event managers to identify potential hazards early and prevent tragedies [7]. The proposed methodology, based on 1-deniability, transmits bloom filters securely across the network, enabling server-side intersections that facilitate detailed analysis of people flow within defined time windows, ensuring a privacy-conscious and robust crowd monitoring solution. [8]. The use case chosen in this paper involves managing the crowds of the Umrah event, which is challenging because of the large number of pilgrims from different cultures, languages, age groups, and nationalities. Controlling pandemic-related risks has recently become a critical aspect to address [9]. Therefore, the proposed system integrates real-time monitoring, health tracking, and crowd behavior analysis to ensure safety during large gatherings. Intelligent, technology-driven approaches are essential for mitigating risks such as stampedes, disease spread, and emergency situations [10].

The key contribution of the paper:

1. Proposes a Semantic Video Analytics Framework for Crowd Management (SVAF-CM) that combines a SpinalXNet-enhanced CNN and semantic modeling to extract multilevel spatial

* Corresponding author: hsm@rithassan.ac.in

features and generate accurate crowd density maps.
2. It integrates low-level video features into high-level contextual knowledge, enabling the precise interpretation of crowd behavior, spatial patterns, and potential hazards.
3. Enhances feature representation and reduces computational complexity, achieving real-time performance with lower memory usage compared with traditional RNN, LSTM, and BiLSTM models.
4. Validates the effectiveness of the proposed method through extensive evaluation on UCF_CROWD_50 and UCF_QNRF datasets, demonstrating superior accuracy and robustness across all performance metrics.

Organized as follows, the remainder of the paper includes: Section 2 reviewing related work, Section 3 explaining the proposed SVAF-CM method, Section 4 presenting results and comparisons, Section 5 discussing findings, and Section 6 concluding with contributions and future directions.

2 Literure Review

Sultan Daud Khan et al. [11] developed deep fusion architecture leverages hierarchical features across multiple convolutional layers to optimize performance with both efficiency and precision The method developed maps input image to a density map by combining multiscale features from the network's shallow to deep layers. By summing the peaks in density map, the final crowd count is obtained. Experimented results were obtained using three benchmark datasets. The developed method outperformed the other advanced methods by achieved lower MAE and MSE values. However, the developed method struggles with low-resolution inputs and an extremely small head size in dense scenes, limiting its accuracy.

Silky Goel et al. [12] developed the MaskFormer EfficientNetB7 instance segmentation framework for efficient crowd counting. MaskFormer was used as a mask classification model to predict binary masks linked to global classes, whereas EfficientNetB7 employed a compound-scaling algorithm to balance feature extraction. Experiments were conducted using the UCF-QNRF, ShanghaiTech Part A, ShanghaiTech Part B, and Mall datasets. The approach achieved a lower MAE and Root Mean Square Error than the existing methods. The method demonstrated strong adaptability across environments; however, a disadvantage was its lack of integration between semantic and instance segmentation, which limited its ability to fully capture crowd density and diversity in highly congested scenarios.

Zhiqiang Zhao et al. [13] developed CL-DCNN integrates dilated convolutions with cross-level connections to extract rich contextual information while preserving local details. The methodology introduced a dilated contextual module (DCM) with cross-level connections between feature maps to integrate contextual information while preserving local details. MAE values of 52.6, 8.1, 1.55, 181.8, and 96.4 were achieved on the ShanghaiTech Part A, ShanghaiTech Part B, Mall, UCF_CC_50, and UCF-QNRF datasets, during model evaluation. Outperforming advanced methods, the model still faced limitations in practical scenarios like UAV platforms because it lacked lightweight CNN integration.

Yongqi Chen, Huailin Zhao et al. [14] developed a weakly supervised hybrid lightweight crowd-counting network, integrating GhostNet backbone for local feature extraction and a modified swin-transformer block for global context. A Pyramid Pooling Aggregation Module and cross-attention module enhanced scale variation handling and feature flow, whereas a simplified regressor avoided density map generation. The approach improved the MAE and MSE accuracy compared with ARNet and PDDNet with reduced parameters. However, the limitation was the need for larger and more diverse datasets, which restricted the scalability and generalization of the model.

Lidia Fotia et al. [15] developed a novel training methodology for accurate crowd counting in moderately and highly crowded scenarios using point detection, where each point represented the head of an individual. The methodology applied Curriculum Learning to mimic human cognition, starting from simple to complex tasks, supported by a new indicator that considered the number and distribution of people. Experiments on multiple public datasets demonstrated significant improvements over the baseline methods. Limited adaptability to diverse camera perspectives, such as ego-view, is a key disadvantage.

2.1 Problem Statement

Accurate crowd density estimation and behavior analysis in large public events is challenging due to the high crowd density, severe occlusions, varying camera viewpoints, and complex motion patterns. Traditional surveillance and existing deep learning models often fail to provide real-time semantic understanding, struggle with extremely dense crowds, and incur high computational overhead, thereby limiting their effectiveness in ensuring safety and proactive crowd management.

2.2 Objective

To address these challenges, this paper proposes the SVAF-CM method, which integrates a SpinalXNet-enhanced CNN with semantic modeling. This method effectively extracts multilevel spatial features, maps low-level video data into high-level contextual knowledge, and reduces the computational complexity. The features are processed to generate accurate crowd density and spatial distribution estimations, improving the monitoring accuracy and efficiency on the UCF_CROWD_50 and UCF_QNRF datasets.

3 Proposed Method

This paper proposes the SVAF-CM method, which integrates a SpinalXNet-based Convolutional Neural Network (CNN) for crowd management in large-scale public events. The architecture extracts low-level visual features and generates density maps to estimate crowd distribution, thereby enhancing crowd density assessment and understanding of spatial patterns. To improve contextual understanding, semantic modeling is incorporated to map video features into high-level knowledge, enabling an accurate interpretation of crowd behavior. Preprocessing involves normalization, perspective correction, and noise reduction to refine the video quality. The extracted features and density maps are then processed to produce crowd analytics for management and monitoring. The overall work flow of proposed SVAF-CM is shown in Figure 1.

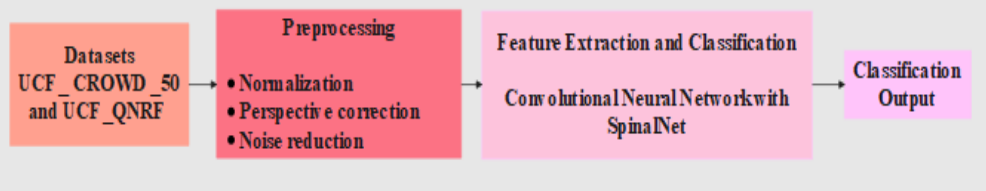

Fig. 1. Overall work flow of the SVAF-CM method.

3.1 Description of UCF_ CROWD _50 and UCF_QNRF datasets

The UCF_CC_50 dataset [16], developed at University of Central Florida, includes 63,075 annotated head centers and 50 images. It covers scenes like marathons, stadiums, protests, and concerts, where densities range from 94 to 4543.Perspective distortion varies due to different viewing angles, as shown in Figure 2. The UCF_QNRF dataset [17], comprising 1,535 high-resolution crowd images and 1,251,642 annotated head centers, is a highly challenging dataset that captures a broad as shown in Figure 3, surveillance cameras with diverse viewpoints and angles recorded images that exhibited head counts from 49 to 12,865.

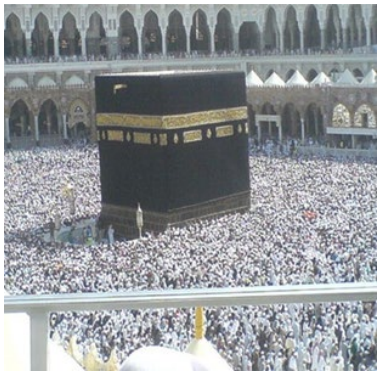

Fig. 2. Illustrative example from UCF-CROWD-50.

Fig. 3. Illustrative frame from the UCF-QNRF dataset.

3.2 Preprocessing

For UCF-CROWD-50 and UCF-QNRF datasets, first step in preprocessing involved image resizing, where all images were scaled to a fixed resolution to standardize the input dimensions for the model. Next, normalization involves scaling the pixel values to the range [0, 1] through division by 255, ensuring uniform data representation and facilitating faster convergence during training. Additionally, noise reduction is performed using Gaussian blur, which smooths the images and reduces unwanted artifacts, thereby enhancing the feature extraction and improving the overall quality of the input data. These steps collectively prepare datasets for effective model training and accurate crowd density estimation.

3.3 Semantic video Analytics

Semantic video analysis is employed to identify and extract core content conveyed in the video. Depending on the application, different semantic aspects are integrated into the analysis. Semantic video analysis employs human abstraction, where a priori information translates semantics into rules applied through an algorithm. Examples comprise template matching, face detection, moving-object segmentation, and extraction of captions and text. When the object shape is known a priori, template matching is applied. Pixels whose spectral characteristics fall within a specific region of chromaticity diagram are identified for face detection. Motion information can be used as semantics for moving objects. For segmenting moving objects, change detection is commonly used. Using a static camera, change detection separates moving objects (foreground) from static background. Semantic analysis is applicable to moving cameras after global motion compensation and also addresses the static camera problem. To determine whether the frame difference results from noise or an object, the locally adaptive threshold $\tau(x, y)$, models noise statistics and conducts a significance test. Let the null hypothesis H_0 refer to the assumption where $g(x, y)$ represents the sum of absolute frame differences in an q observation window around (x, y). Given additive Gaussian camera noise with variance σ,

Equations (1-2) indicate that the conditional probability density function of the frame difference follows a χ_q^2 distribution.

$$f\ (g(x,y)|H_0) = \frac{1}{2^{q/2}\sigma^q\Gamma(q/2)} g(x,y)^{(q-2)/2} e^{-g^{(x,y)^2}/2\sigma^2} \quad (1)$$

Where $\Gamma(\cdot)$ is the gamma function. To ensure robustness and accuracy, $q = 25\ (5 \times 5\ window)$. The significance test is performed as follows:

$$P\{g(x,y) \geq \tau\ (x,y)|H_0\} = \frac{\Gamma(q/2, g(x,y)^2/2\sigma^2)}{\Gamma q/2} \quad (2)$$

If this probability is smaller than the significance level α ($10^{-2}\ to\ 10^{-6}$), the pixel is labeled as a moving object. This change-detection process segments moving objects from the background and enables the extraction of object metadata combined with video tracking.

3.4 Feature Extracting using CNN with SpinalNet

3.4.1 CNN

This section presents details of the proposed method for crowd management in large public events.

The crowd-counting task in the proposed architecture involves fusing feature maps from different levels, where each level encodes features at distinct scales. Following the standard pipeline of popular crowd-density estimation techniques, the proposed architecture consists of two interconnected networks. In the proposed method, the encoder extracts hierarchical features from the input image, while the decoder constructs the density map, together forming the two-network architecture. The number of people per pixel of input image is what the final density map represents. In DenseNet design, every convolutional layer except the first generates its output feature map using the output from the preceding layer, while the initial layer takes the input image directly. The Densely Connected Convolutional Network reflects the DenseNet design, in which each layer maintains connections with all other layers. Between the L levels, number of direct connections equals $L(L+1)/2$. Feature maps from preceding layers act as inputs to the current layer, while those from the current layer are passed to subsequent layers. In DenseNet, a convolutional network is applied to a single X-ray image, X_0. Each of the L layers in the network performs a nonlinear transformation $H_1(.)$, where l indexes the layers. Several iterations of the DenseNet architecture exist, including but not limited to DenseNet-121, DenseNet-160, and DenseNet-201. In this paper, DenseNet-121 is analyzed and improved by adding extra layers to achieve higher levels of precision.

3.4.2 SpinalNet

The evolution of CNNs has enabled researchers to reach several milestones over time. Through the use of CNNs, machines and humans have become closer to each other. While CNNs offer multiple advantages, they tend to lose spatial information post-pooling and require deeper layers to enhance prediction precision. The SpinalNet structure provides the same precision, but delivers better results. In SpinalNet, the input sublayer is divided into two segments that are repeatedly assigned to hidden sublayers across its three sublayer structures. The structure of SpinalNet is illustrated in Figure. 4.

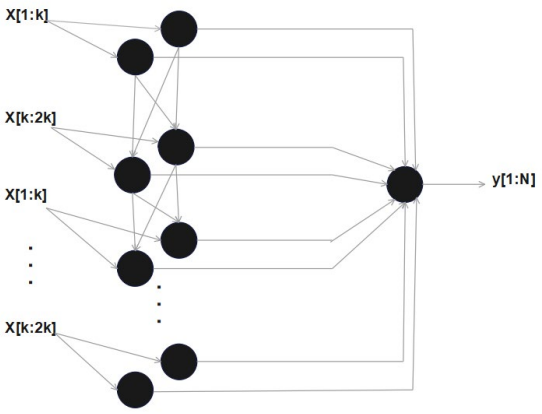

Fig. 4. Structure of SpinalNet.

3.5 Density map

Density map generation methods are applied for network training, with h_i indicating the location of the pedestrian in the image. For all pedestrian positions in the image, the delta function $(h\ - h_i)$ f is formulated in Equation (3):

$$H(h) = \sum_{j=1}^{n} \delta(h\ - h_i) \quad (3)$$

The variance of kernel G is represented by σ. The variance of kernel. This density function is applicable to scenes captured from an orthogonal perspective, where object size remains constant. In pedestrian crowd scenes, the camera is usually tilted, causing perspective distortions. To address this, a Gaussian kernel is used to compensate for perspective distortions and generate a density map. H is convolved with G to obtain a continuous density function, , as shown in Equation (4).

$$H(h) = \sum_{j=1}^{n} \delta(h\ - h_i) * G\sigma(h) \quad (4)$$

Equation (5) formulates the Euclidean loss L_E, which enables the network to learn parameters by quantifying the difference between actual and predicted density maps.

$$L_E(\theta) = \frac{1}{N}\sum_{j=1}^{N} ||K_d(S_j;\theta) - D_j||^2 \quad (5)$$

Here, θ stands for the network parameters, N represents the number of images, S_j identifies the current image, and D_j refers to its density ground-truth map j.

4 Result and Discussion

The implementation of SVAF-CM employs PyTorch and is trained on NVIDIA TITAN Xp GPU, with setup involving PyTorch 1.4, CUDA 10.2, Anaconda 3, and Ubuntu 16.04 (64-bit).

4.1 Evaluation Metric

Performance of the proposed SVAF-CM method is evaluating using mean absolute MAE, MSE, and RMSE, as listed in Table 1.

Table 1. Evaluation metric of proposed SVAF-CM method.

Metric	Formula
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{n=1}^{N}\lvert\mu_n - G_n\rvert$
Mean Square Error	$MAE = \frac{1}{N}\sum_{n=1}^{N}(\mu_n - G_n)^2$
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

4.2 Performance Evaluation

Performance of the proposed SVAF-CM method on two benchmark crowd-counting datasets, UCF_CROWD_50 and UCF_QNRF, is evaluated and compared with traditional models such as RNN, LSTM, and BiLSTM. In terms of MAE, RMSE, and MSE, the proposed SVAF-CM method achieved values of 0.010, 0.012, and 0.013 for UCF_CROWD_50, and 0.010, 0.015, and 0.018 for UCF_QNRF, respectively, significantly outperforming existing methods. Similarly, the proposed method demonstrates lower error rates across all evaluation metrics than the baseline methods. Overall, the results show that the SVAF-CM method is highly effective for estimating crowd density. Table 2 lists a performance comparison of the SVAF-CM method with the baseline methods.

Table 2. Performance analysis of SVAF-CM method.

Model	Metrics		
	MAE	MSE	RMSE
UCF_CROWD_50			
RNN	0.045	0.052	0.060
LSTM	0.032	0.038	0.045
BiLSTM	0.020	0.024	0.030
Proposed (SVAF-CM)	0.01	0.012	0.013
UCF_QNRF			
RNN	0.055	0.070	0.085
LSTM	0.040	0.052	0.065
BiLSTM	0.025	0.032	0.040
Proposed (SVAF-CM)	0.010	0.015	0.018

4.3 Computational Complexity

The proposed SVAF-CM method is assessed in terms of training time, testing time, and memory consumption, and its performance is compared with that of the baseline models. The SVAF-CM model achieved a training time 96 s and a testing time 10 s, which is lower than those of the RNN (141 s, 28 s), LSTM (133 s, 21 s), and GRU (100 s, 15 s). In addition, the memory requirement of SVAF-CM is only 100 MB, while the baseline models consume considerably more memory resources. These results confirm that the SVAF-CM model offers reduced computational overhead, making it efficient for deployment in practical crowd-counting applications. Figure 5 illustrates a comparison of the computational complexity of the different models.

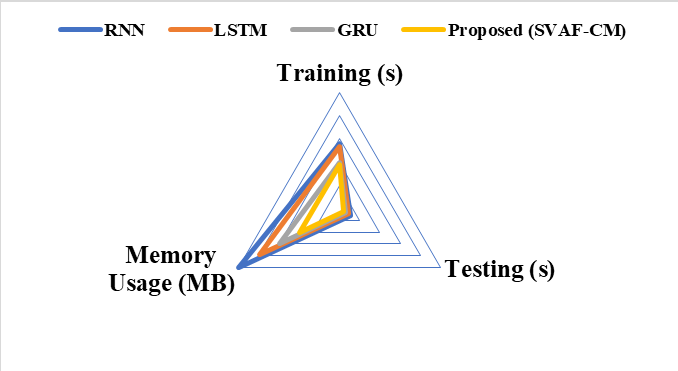

Fig. 5. Complexity of DL models.

4.4 Statistical Analysis

Compared with the traditional RNN, LSTM, GRU, BiGRU, CNN, and CNN+S+ PinalXNet models, the proposed SVAF-CM method demonstrated superior accuracy and adaptability across UCF_CROWD_50 and UCF_QNRF datasets. The improvement in error reduction was statistically significant ($p < 0.05$), as validated by a paired t-test conducted over 10 independent runs. These results confirm the reliability, effectiveness, and superiority of the proposed SVAF-CM method for scalable and precise crowd density estimation.

4.5 Ablation Study

Ablation study in this section demonstrates the performance of each modification within the proposed SVAF-CM method on UCF_CROWD_50 and UCF_QNRF datasets. The base CNN configuration recorded comparatively higher errors, while the addition of SpinalXNet enhances feature extraction and improved accuracy across the evaluation metrics. With

the integration of the complete SVAF-CM method, the method achieves the lowest MAE, MSE, and RMSE values, indicating a superior error reduction capability. Each enhancement step provides measurable gains, and the combined method significantly strengthens the effectiveness of the model for crowd density estimation, as presented in Table 3.

Table 3. Ablation study of SVAF-CM.

Model Configuration	Metrics		
	MAE	**MSE**	**RMSE**
UCF_CROWD_50			
CNN	17.8	12.3	14.9
CNN+ SpinalXNet	14.1	17.4	18.1
Proposed (SVAF-CM)	0.01	0.012	0.013
UCF_QNRF			
CNN	14.7	16.2	14.2
CNN+ SpinalXNet	20.0	20.8	15.5
Proposed (SVAF-CM)	0.010	0.015	0.018

4.6 Comparative Analysis

The proposed SVAF-CM method s evaluated on UCF_CROWD_50 and UCF_QNRF datasets and compared with existing DL methods. Using UCF_CROWD_50 dataset, the Deep-fusion model reported an MAE of 402.3 with an MSE of 434.1, while CL-DCNN achieved an MAE of 181.8. Similarly, M-SFANet produced an MAE of 343.54 and an MSE of 424.70. In contrast, the proposed SVAF-CM attained an MAE of 0.01 and an MSE of 0.012, indicating significant performance gain. On the UCF_QNRF dataset, the Deep-fusion model [11] recorded MAE of 218.2 and MSE of 357.4, while MFENet [12] obtained an MAE of 90.1 and an MSE of 150.2. Likewise, CL-DCNN [13] and ARNet [14] achieved MAE scores of 96.4 and 110.0, respectively. The proposed SVAF-CM achieved an MAE of 0.010 and MSE of 0.015, surpassing all the baseline methods. The result analysis of proposed SVAF-CM on UCF_CROWD_50 and UCF_QNRF is shown in Figures 6 and 7 respectively.

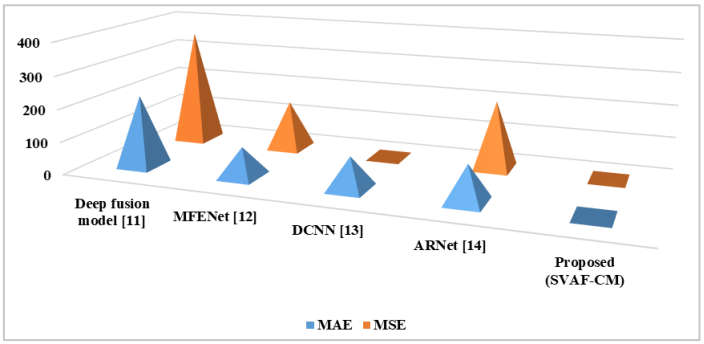

Fig. 6. Evaluation of different models on UCF_CROWD_50 dataset.

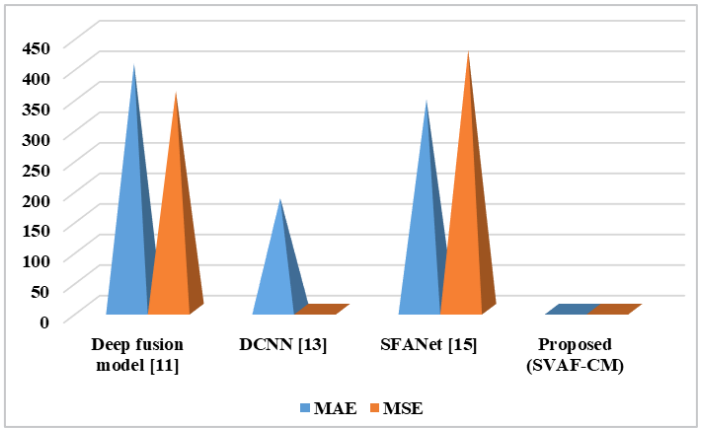

Fig. 7. Evaluation of different models on UCF_QNRF dataset.

5 Discussion

Experimental results indicate SVAF-CM achieves superior performance compared to existing models on UCF_CROWD_50 and UCF_QNRF datasets. The integration of SpinalXNet with CNN ensures the effective capture of multilevel spatial features and semantic context, strengthening the feature representation for accurate crowd density estimation. Preprocessing techniques, including normalization, perspective correction, and noise reduction, further enhance the quality of input frames. The ablation study confirms that each component contributes positively to the performance, with SpinalXNet improving feature extraction and the full SVAF-CM framework achieving optimal MAE, MSE, and RMSE values. Additionally, the method exhibits superior computational efficiency, with reduced training and testing times and lower memory usage compared to the RNN, LSTM, and BiLSTM baseline models, making it highly suitable for real-time crowd monitoring in large-scale public events.

6 Conclusion

This paper proposes the SVAF-CM method, a Semantic Video Analytics Framework for crowd management that combines CNN, SpinalXNet, and semantic modeling to accurately estimate crowd density and spatial distribution. SVAF-CM effectively improved feature representation, error reduction, and computational efficiency. Experimental results using UCF_CROWD_50 and UCF_QNRF datasets demonstrate that SVAF-CM outperforms existing models, achieving an MAE of 0.01 and 0.010, and MSE 0.012 and 0.015. The ablation study validated the contribution of each module, while computational analysis confirmed reduced training and testing times, demonstrating its efficiency. Overall, SVAF-CM provides a reliable, scalable, and high-performance solution for crowd monitoring and management. Future work will focus on extending the SVAF-CM framework to incorporate multicamera fusion and predictive modeling of crowd movement for enhanced safety and proactive risk mitigation.

Reference

1. M. Qaraqe, Y.D. Yang, E. B. Varghese, E. Basaran, A. Elzein, Crowd behavior detection: Leveraging video swin transformer for crowd size and violence level analysis. Appl. Intell. **54**, 10709-10730 (2024). https://doi.org/10.1007/s10489-024-05775-6
2. R. Alzahrani, N. Algethami, Leveraging machine learning for optimal pilgrim crowd management. Electronics **14**, 2507 (2025). https://doi.org/10.3390/electronics14132507
3. P.S. Karthika, V. Kedar, A. Verma, A walk accessibility-based approach to assess crowd management in mass religious gatherings. J. Transp. Geogr. **104**, 103443 (2022). https://doi.org/10.1016/j.jtrangeo.2022.103443
4. A. Siddiqa, W.Z. Khan, M.H. Alkinani, E.A. Aldhahri, M.K. Khan, Edge-assisted federated learning framework for smart crowd management. Internet Things **27**, 101253 (2024). https://doi.org/10.1016/j.iot.2024.101253
5. J. Soza-Parra, I. Tiznado-Aitken, J.C. Munoz, A discrete-event public transportation simulation model to evaluate travel demand management impacts on waiting times and crowding conditions. J. Public Transp. **25**, 100075 (2023). https://doi.org/10.1016/j.jpubtr.2023.100075
6. V.W. Wong, K.H. Law, Fusion of CCTV video and spatial information for automated crowd congestion monitoring in public urban spaces. Algorithms **16**, 154 (2023). https://doi.org/10.3390/a16030154
7. A. Serek, B. Amirgaliyev, R.Y.M. Li, A. Zhumadillayeva, D. Yedilkhan, Crowd density estimation using enhanced multi-column convolutional neural network and adaptive collation. IEEE Access. **13**, 146956-146972 (2025). https://doi.org/10.1109/ACCESS.2025.3597393
8. R. Rusca, D. Gasco, C. Casetti, P. Giaccone, Privacy-preserving WiFi fingerprint-based people counting for crowd management. Comput. Commun. **225**, 339-349 (2024). https://doi.org/10.1016/j.comcom.2024.07.010
9. W. Halboob, H. Altaheri, A. Derhab, J. Almuhtadi, Crowd management intelligence framework: Umrah use case. IEEE Access **12**, 6752-6767 (2024). https://doi.org/10.1109/ACCESS.2024.3350188
10. N.M. Bahbouh, S. Sendra, A.A.A. Sen, Designing a comprehensive framework for health management in crowded events. Int. J. Inf. Technol. **17**, 1641-1651 (2025). https://doi.org/10.1007/s41870-024-02051-1
11. S.D. Khan, Y. Salih, B. Zafar, A. Noorwali, A deep-fusion network for crowd counting in high-density crowded scenes. Int. J. Comput. Intell. Syst. **14**, 168 (2021). https://doi.org/10.1007/s44196-021-00016-x
12. S. Goel, D. Koundal, A MaskFormer EfficientNet instance segmentation approach for crowd counting. Sci. Rep. **15**, 13275 (2025). https://doi.org/10.1038/s41598-025-95174-9
13. Z. Zhao, P. Ma, M. Jia, X. Wang, X. Hei, A dilated convolutional neural network for cross-layers of contextual information for congested crowd counting. Sensors **24**, 1816 (2024). https://doi.org/10.3390/s24061816
14. Y. Chen, H. Zhao, M. Gao, M. Deng, A weakly supervised hybrid lightweight network for efficient crowd counting. Electronics **13**, 723 (2024). https://doi.org/10.3390/electronics13040723
15. L. Fotia, G. Percannella, A. Saggese, M. Vento, Optimizing crowd counting in dense environments through curriculum learning training strategy. SN Comput. Sci. **5**, 683 (2024). https://doi.org/10.1007/s42979-024-03029-3
16. UCF_CROWD_50 dataset Link: https://www.kaggle.com/datasets/tthien/ucfcc50 (Accessed on 03.09.2025)
17. UCF-QNRF dataset Link: https://www.kaggle.com/datasets/faihajalamtopu/ucf-qnrf (Accessed on 03.09.2025)