CHAPTER 6
MULTISCALE FLOW ATTENTIVE DEPTH SEPARABLE CNN FOR MULTITASKING CROWD ANALYSIS

6.1 Introduction

Crowd disaster management requires accurate analysis of crowd scenes. The crowd analysis is done by a collective understanding of crowd count and density estimation, crowd flow, crowd behavior, and crowd congestion. Nevertheless, crowd counting and behavior understanding are essential to minimize crowd disaster, which is the main focus of the proposed work. Variation of crowd shape and influence of scene background information degrades the performance of crowd analysis. The current research trends focus on exploiting deep learning techniques for several tasks of crowd analysis. Deep models like convolution neural network (CNN), long-short term memory (LSTM), encoder-decoders, and generative adversarial networks have been vastly exploited. Different models for single-image-based crowd counting have extracted scale-invariant features to handle crowd shape variation, whereas a few efforts have been proposed to minimize the background influence. However, there is a lack of models to address such issues in video-based crowd counting and crowd behavior prediction. Most of the solutions are single-task-based. The solution to crowd analysis using different single-task models would increase the computation overheads and have synchronization issues. Hence, a multitasking crowd analysis model is highly required, lacking in the literature. This may be because of the unavailability of a multitasking crowd analysis dataset. So, the following things could be drawn which need to be addressed.

- There is a lack of availability of a multitasking crowd analysis model.
- There is a lack of availability of a multitasking crowd analysis dataset.

- Two challenging issues: scale variation and minimization of background, have to be handled as far as video-based crowd analysis (crowd counting and crowd behavior prediction).

To fulfill the above research gaps, a multiscale flow attentive multilayer depth separable CNN has been proposed for multitasking crowd analysis from crowd video datasets. The followings are the main contribution,

- A large-scale multitasking crowd analysis (i.e., for crowd counting and crowd behavior prediction) datasets are generated using the publicly available crowd behavior datasets (i.e., MED and GTA). More than 1,20,000 frames (from 45 video sequences) have been used for annotation.
- A multitasking crowd analysis model is proposed, which effectively handles crowd shape variation and minimizes the effect of backgrounds.
- The backbone of the model is designed using the depth-wise separable CNN. A flow attentive module is designed to minimize the effect of background details from different scales of spatial-temporal features.
- The scale-invariant features are extracted to handle scale variation issues in the crowd videos.
- Comprehensive results analysis and extensive ablation study have been conducted to show the effectiveness of the proposed model.

6.2 The Proposed Method and Model

6.2.1 Overview

Multitasking crowd analysis focusing on crowd counting and crowd behavior prediction is essential to draw effective crowd management strategies, public space design and provide a better visual surveillance system to minimize crowd disasters. The efficiency of any crowd behavior classification model is mainly affected by two major issues in video sequences: crowd shape changes in the video sequence and the effect of cluttered background. The proposed model is designed to address these issues. The overall architecture of the proposed model is illustrated in Figure 6.1. The working of the proposed model can be explained by using the following major points..

- Pre-processing.
- Network Overview.
- Spatial-Temporal Feature Modelling using Depth Separable CNN.
- Working of Flow Attention Block.
- Multiscale Feature Modelling.
- Multitasking Crowd Analysis and Optimization.

6.2.2 Pre-processing

During pre-processing the video sequences, the frames are obtained and converted into their grayscale level. For each timestamp, volume of frames is obtained. Each volume constitutes of three consecutive frames at time stamp, t, t-1 and t-2. Let the volume of frames for each time stamp of video sequence is represented by a set $S = \{s_1, s_2, \dots \dots .., s_N\}$, where N is the total number of frames of a video sequence.

6.2.3 Network Overview

As shown in Figure 6.1, the proposed model is built using a backbone network called Depth Separable CNN (DSCNN). The input to the DSCNN is the batches of the volume of frames. The DSCNN consists of four depth separable convolution layers, each of these layers is followed by a convolution layer with several kernels with size (1×1). Each convolution layer is followed by a down sampling layer which is the max-pooling layer used in the proposed model. The size of the max-pooling layer is set to (2×2) with a stride of 2. All the layers are padded with zeros. The arrows with different colors represent the shapes of the tensors passing from one layer to another during training the model in a mini-batch manner. Four flow attention blocks (FAB) are proposed; each takes the output of each convolution layer as input, respectively. The FABs are applied at different scales of feature maps of the DSCNN network.

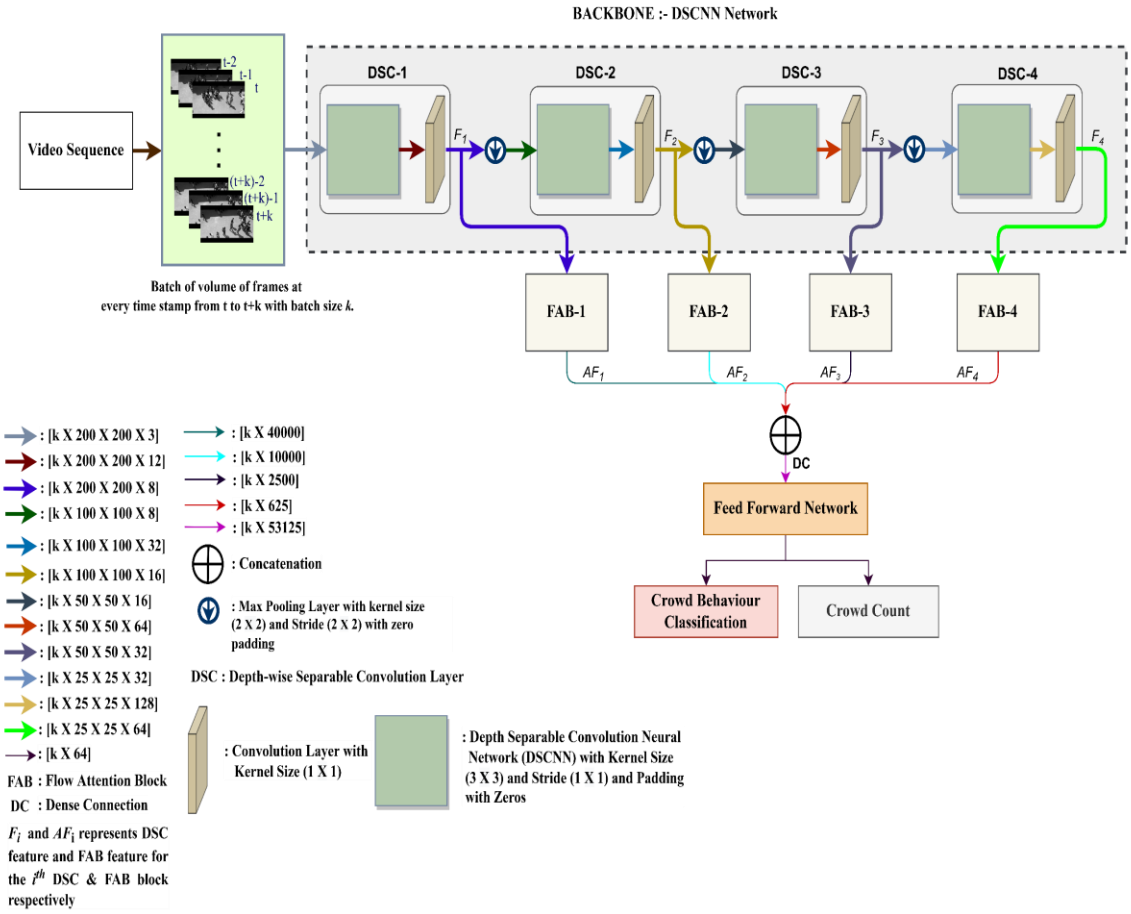

BACKBONE :- DSCNN Network
Video Sequence
t-2
t-1
t
(t+k)-2
(t+k)-1
t+k
Batch of volume of frames at every time stamp from t to t+k with batch size *k*.
DSC-1
DSC-2
DSC-3
DSC-4
F_1
F_2
F_3
F_4
FAB-1
FAB-2
FAB-3
FAB-4
AF_1
AF_2
AF_3
AF_4
DC
Feed Forward Network
Crowd Behaviour Classification
Crowd Count
: [k X 200 X 200 X 3]
: [k X 200 X 200 X 12]
: [k X 200 X 200 X 8]
: [k X 100 X 100 X 8]
: [k X 100 X 100 X 32]
: [k X 100 X 100 X 16]
: [k X 50 X 50 X 16]
: [k X 50 X 50 X 64]
: [k X 50 X 50 X 32]
: [k X 25 X 25 X 32]
: [k X 25 X 25 X 128]
: [k X 25 X 25 X 64]
: [k X 64]
: [k X 40000]
: [k X 10000]
: [k X 2500]
: [k X 625]
: [k X 53125]
: Concatenation
: Max Pooling Layer with kernel size (2 X 2) and Stride (2 X 2) with zero padding
DSC : Depth-wise Separable Convolution Layer
: Convolution Layer with Kernel Size (1 X 1)
: Depth Separable Convolution Neural Network (DSCNN) with Kernel Size (3 X 3) and Stride (1 X 1) and Padding with Zeros
FAB : Flow Attention Block
DC : Dense Connection
F_i and AF_i represents DSC feature and FAB feature for the i^{th} DSC & FAB block respectively

Figure 6.1: Overall architecture of the proposed model

Figure 6.2 shows details of FABs. The FABs utilize flow maps of the frames, which are obtained using the Lucas-Kanade optical flow algorithm [179]. The detailed working of these models is explained in the subsequent sections. The outputs of FABs are concatenated and given to the feed foreword network (FFN) to perform both crowd counting and crowd behavior prediction. The FFN contains two hidden layers with 512 and 64. All the layers of FFN are densely connected (DC) one after another. The details of FFN are illustrated in Figure 6.3.

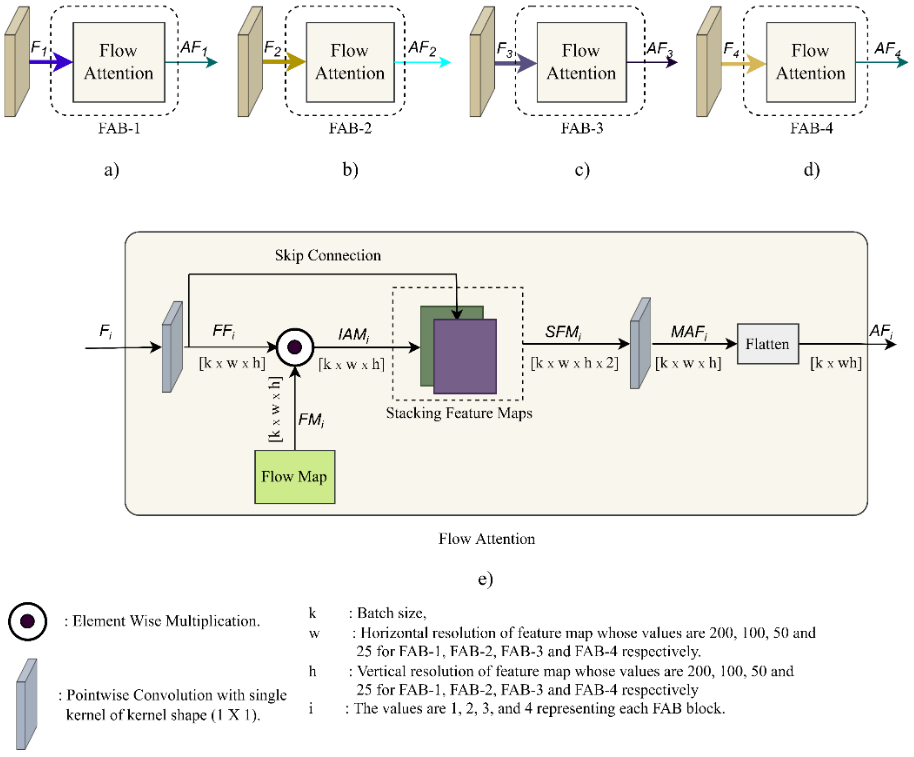

F1
Flow Attention
AF1
FAB-1
a)
F2
Flow Attention
AF2
FAB-2
b)
F3
Flow Attention
AF3
FAB-3
c)
F4
Flow Attention
AF4
FAB-4
d)
Skip Connection
Fi
FFi
[k x w x h]
IAMi
[k x w x h]
[k x w x h]
FMi
Flow Map
Stacking Feature Maps
SFMi
[k x w x h x 2]
MAFi
[k x w x h]
Flatten
AFi
[k x wh]
Flow Attention
e)
: Element Wise Multiplication.
: Pointwise Convolution with single kernel of kernel shape (1 X 1).
k : Batch size,
w : Horizontal resolution of feature map whose values are 200, 100, 50 and 25 for FAB-1, FAB-2, FAB-3 and FAB-4 respectively.
h : Vertical resolution of feature map whose values are 200, 100, 50 and 25 for FAB-1, FAB-2, FAB-3 and FAB-4 respectively
i : The values are 1, 2, 3, and 4 representing each FAB block.

Figure 6.2: Details of FAB.

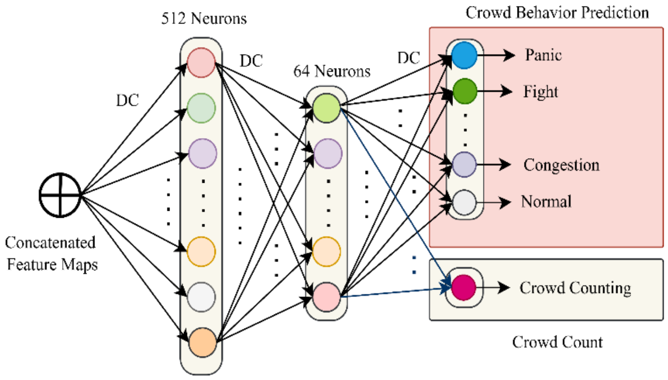


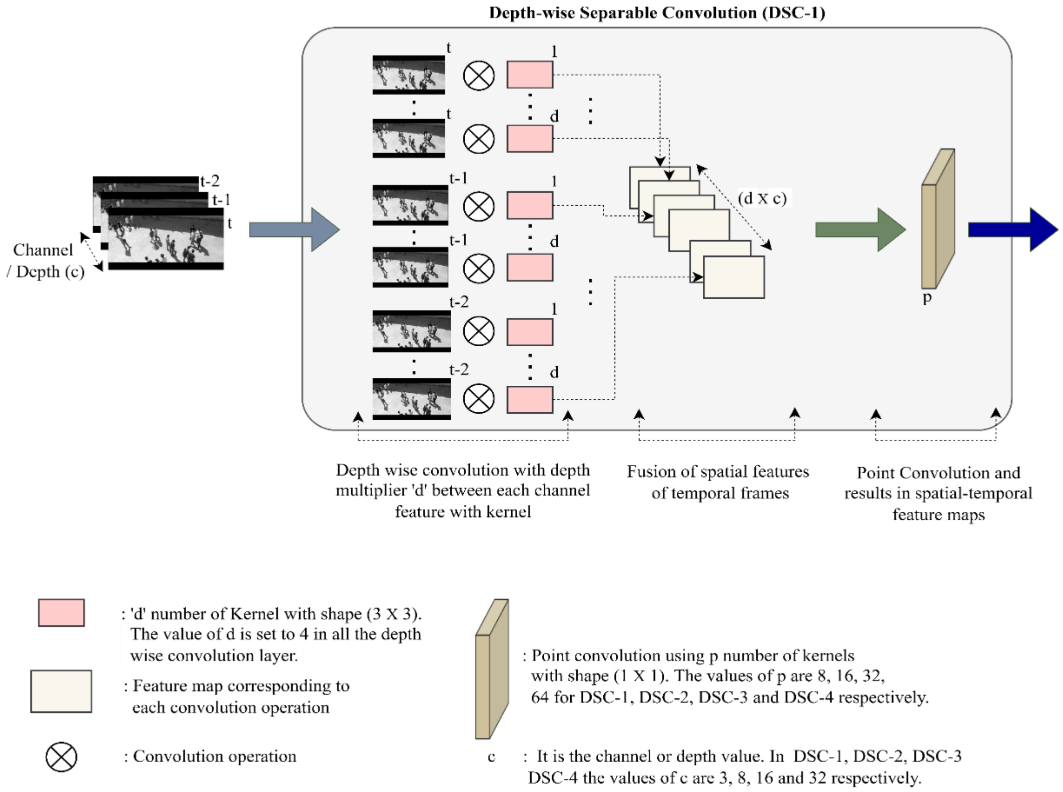
Figure 6.3: Details of Feed Foreword Network

6.2.4 Spatial-Temporal Feature Modelling using Depth Separable CNN.

The depth-wise separable convolution (DSC) layer is solely designed to minimize the total number of matrix multiplication operations during convolution. This is done by implementing the following two processes.

- Depth-wise convolution operation followed by
- Pointwise convolution operation.

The DSC layer treats the channel dimension as the depth of the image or feature maps and is generally used to obtain spatial features. However, the same DSC layer can be used to obtain spatial-temporal features from the video sequence. The only thing we have to do is input the DSC layer with the volume of frames. Further, we can exploit fine-grained depth-wise features by utilizing the depth multiplier during depth-wise convolution operation. Figure 6.4 illustrates the spatial-temporal feature modeling using DSC-1 of the proposed model with depth multiplier *d*.



Depth-wise Separable Convolution (DSC-1)
t-2
t-1
t
Channel / Depth (c)
t
t
t-1
t-1
t-2
t-2
1
d
1
d
1
d
(d X c)
p
Depth wise convolution with depth multiplier 'd' between each channel feature with kernel
Fusion of spatial features of temporal frames
Point Convolution and results in spatial-temporal feature maps
: 'd' number of Kernel with shape (3 X 3). The value of d is set to 4 in all the depth wise convolution layer.
: Feature map corresponding to each convolution operation
: Convolution operation
: Point convolution using p number of kernels with shape (1 X 1). The values of p are 8, 16, 32, 64 for DSC-1, DSC-2, DSC-3 and DSC-4 respectively.
c : It is the channel or depth value. In DSC-1, DSC-2, DSC-3 DSC-4 the values of c are 3, 8, 16 and 32 respectively.

Figure 6.4: Details of Spatial-Temporal Feature Modelling using a DSC-1.

According to Figure 6.4, in DSC-1, the depth-wise convolution operation with depth multiplier ‘d’ has been performed between each channel feature or image and the d number of the convolution kernel. One thing we have to remember is that here the channels are the frames at the different time stamps. All the convolution features of different channel-wise features/images are stacked whose channel dimension is $[d \times c]$. The stacked features are given to a point-wise convolution layer which fuses all the stacked feature maps across channel dimensions and results in spatial-temporal features. The point convolution will have p number of instances of spatial-temporal features. The depth-wise separable and point convolution process is the same for DSC-1, DSC-2, DSC-3, and DSC-4.

6.2.5 Working of Flow Attention Block.

The flow attention blocks (FAB) are introduced to provide attention to the spatial-temporal features of each DSC block with the optical flow maps. Figures 6. 2 a), b), c), and d) represent the overall block diagram of the FAB-1, FAB-2, FAB-3, and FAB-4, respectively. Figure 6. 2 e) explains the detailed structure of any FAB. The FAB takes the DSC features, i.e., $F_i \in \mathbb{R}^{k\times w\times h\times ch}$, and produces Attentive Feature-maps, i.e., $F_i \in \mathbb{R}^{k\times w\times h}$ where $i\ (= 1{,}2{,}3{,}4)$ representing each DSC block, k is the batch size, w is the width of the feature map, h is the height of the feature map, and ch is the channel dimension. The values of (w, h) are $(200, 200), (100, 100), (50, 50) and\ (25, 25)$ for FAB-1, FAB-2, FAB-3 and FAB-4 respectively. The flow maps of four resolutions, i.e., $(200, 200), (100, 100), (50, 50) and\ (25, 25)$ are obtained using the famous Lucas-Kanade optical flow. The Lucas-Kanade method provides sparse optical flow between two frames with less noisy output. The optical flow represents the motion objects only; hence such techniques eliminate the background pixels. So, by imposing a flow map as attention to the fused DSC feature maps, the background pixels will be removed, thus minimizing the effects of cluttered background. The attention process is given in Equations 6.1 to 6.4.

$$IAM_i = FF_i \odot FM_i \tag{6.1}$$

$$SFM_i = Concatenate([IAM_i, FF_i]) \tag{6.2}$$

$$MF_i = Conv_{(1\times 1)}(SFM_i) \tag{6.3}$$

$$AF_i = Flatten(MAF_i) \tag{6.4}$$

, here the abbreviations like IAM_i, FF_i, FM_i, SFM_i, and AF_i are mentioned in Figure 6.2 which resembles to Intermediate Attention Map, Fused Features, Flow Map, Stacked Feature Maps, Merged Features and Attentive Features for the i^{th} block respectively.

6.2.6 Multiscale De-background Feature Modelling.

After obtaining all the flow attentive feature maps for four FABs, all these feature maps are concatenated. These fused de-background features of different scales of FABs are termed multiscale de-background features and can also be called scale-invariant features that can handle scale changes due to perspective distortion.

6.2.7 Multitasking Crowd Analysis and Optimization

The scale-invariant features are now inputted to the FFN for multitasking crowd analysis. The focus is on performing two tasks of crowd analysis, i.e., crowd behavior prediction and crowd counting. So, we will have two different outputs. The FFN contains two densely connected hidden layers of 512 64 neurons with activation as ReLU. The activation of crowd counting output of FFN is ReLU, whereas the activation of crowd behavior prediction is SoftMax. Let the sets θ_{CC} and θ_{CBP} represent all the learnable parameters connecting to the output nodes representing crowd counting and behavior, respectively. We have adopted mean squared error (MSE) and cross-entropy (CE) loss for CC and CBP, respectively. These losses are described as follows.

$$Loss_1 = Loss(\theta_{CC}) = \frac{1}{k}\sum_{l=1}^{k}\left(F_l^{pred} - gt_l\right)^2 \tag{6.5}$$

$$[L(\emptyset_{TS-MDA})]^{k} = \frac{1}{k}\sum_{i=1}^{k} L_i\left(T_{i_{CBP}}, Y_{i_{CBP}}\right) \tag{6.6}$$

$$Loss_2 = L_i(\emptyset_{TS-MDA}) = L_i\left(T_{i_{CBP}}, Y_{i_{CBP}}\right) = \left[-\sum_{p=1}^{K} T_p \log y_{p_{out}}\right]^i \tag{6.7}$$

Now, the two losses are combined by summing their weighted sum as given in Equation 6.8.

$$Loss_{Final} = \alpha Loss_1 + \beta Loss_2 \tag{6.8}$$

where $\alpha + \beta = 1$. The final loss is minimised by applying minibatch based gradient descent using Adam optimiser [170].

6.3 Multitasking Crowd Analysis Dataset and Performance Metrics

6.3.1 Multitasking Crowd Analysis Dataset

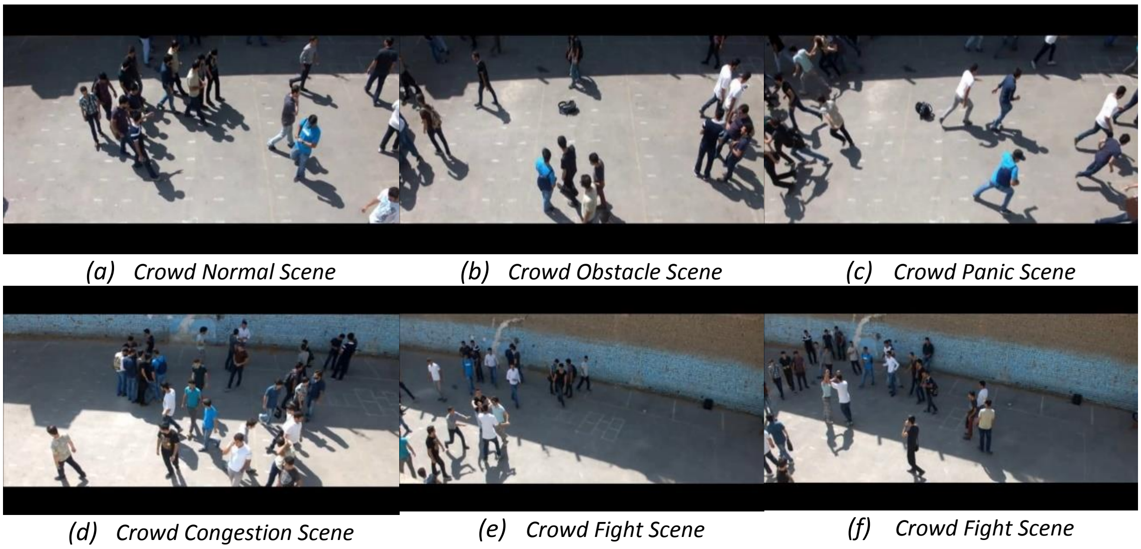
In In the literature, there is a lacking of availability of multitasking crowd analysis datasets focusing on crowd counting and crowd behavior prediction. So, to fulfill such an issue, a multitasking crowd analysis dataset is created using publicly available benchmark crowd behavior datasets like MED [2] and GTA [146]. Combinedly these two datasets contain video frames of around 1,20,000 frames. All these frames were manually annotated for obtaining ground truth crowd counts. The details of the multitasking crowd analysis dataset are illustrated in Table 6.1.

Table 6.1: Stats of multitasking crowd analysis dataset focusing on crowd behaviours and crowd counting

Dataset Name	**Description**	**Environment**	**Modality**	**No. of Sequences**	**Crowd Counting Range**	**Resolution**
The MED [2]	Real-world scenario with artificial escape like panic situation.	Walkways	Videos	31	1-36	$[480 \times 854 \times 3]$
The GTA [146]	Real-world scenario with artificial escape like panic situation.	Free View	Videos	14	0-155	$[1080 \times 1920 \times 3]$

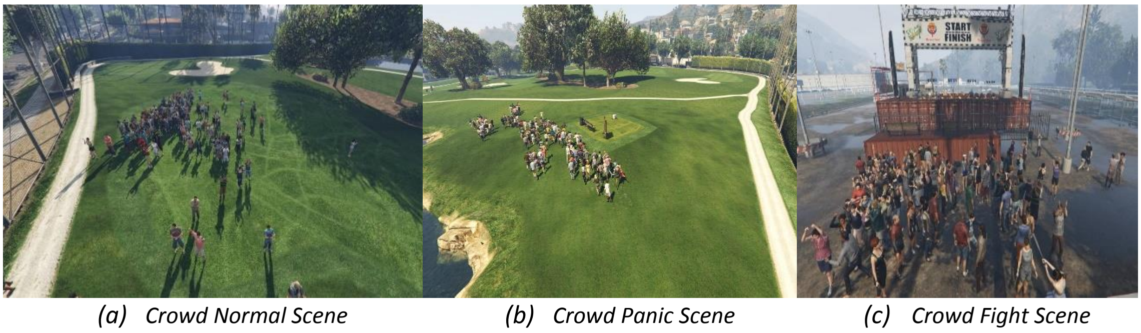
The MED dataset [2] has 31 video sequences of five crowd behaviors: Neutral, Panic, Congestion, Fight, and Obstacle or Abnormal. The resolution of frames of the MED dataset [2] is [480×854×3]. This work adopts the training and testing process as described by [2], i.e., leave-one-out validation. On the other hand, the grand theft auto v2

(GTA) [146] dataset contains 14 video sequences of three crowd behaviors: Neutral, Panic, and Fight scenes. The resolution of frames in GTA [146] is $[1080 \times 1920 \times 3]$. This work also adopted same procedures for training and testing as mentioned in [146] i.e., 10 random video sequences are used for training and rest four are used for testing. Both of these datasets provide frame-level annotation of crowd behaviors. For crowd counting all these frames are manually annotated. The MED dataset contains crowd densities that ranges from 1 to 36 whereas in the GTA dataset the crowd ranges from 0 to 155. Figure 6. 5 and Figure 6. 6 shows samples of crowd behaviors of the MED and the GTA dataset.


(a) Crowd Normal Scene (b) Crowd Obstacle Scene (c) Crowd Panic Scene

(d) Crowd Congestion Scene (e) Crowd Fight Scene (f) Crowd Fight Scene

Figure 6.5: Examples of different samples of the MED dataset.



(a) Crowd Normal Scene (b) Crowd Panic Scene (c) Crowd Fight Scene

Figure 6.6: Examples of different samples of the GTA dataset.

6.4 Experimental Setup

The program is written in TensorFlow and executed using different computing nodes of the Param Sivay Supercomputer. The batch size for all the datasets has been set to 128. The learning rate η, momentum of batch normalization, regularized parameter of L_2, decay rates for first (β_1) and second moment (β_2) of Adam optimizer [170] are initialized to 0.001, 0.95, 0.01, 0.9, and 0.999, respectively. The maximum iteration was to 500. Early stopping with patience of 30 is used to stop the training of the model and also to avoid overtraining the model.

6.5 Results Analysis

The model is optimized by minimizing the combined loss of classification and regression using a minibatch-based gradient descent approach using Adam optimizer. Instead of giving equal weightage to two losses because of the following reasons,

- To understand the trends of performance of the model with different values of weights on the losses.
- To find out on which values of weighted parameters i.e., $\alpha,\ \beta$ the model performs better. The constraint of these two parameters is, $\alpha + \beta$ should be equal to 1.
- To understand the behavior of weighted loss functions on different datasets.

Table 6.2 shows performance of the proposed multitasking model based on different values of weighted loss parameters such as $\alpha, and\, \beta.$ α is the weight given to the classification loss and β is the weight parameter given to the regression loss. From Table 6.2, it can be observed that the behavior of the model concerning different weighted loss values fluctuates until α=0.75 and β=0.25. However, as the values of α decrease from 0.75 and its corresponding β increases from 0.25, the model is more biased towards crowd counting performance than the classification.

Table 6.2: Experimental analysis of performance of the proposed model on various values of weighted loss parameters on the MED dataset

Hyper-Parameter		Performance of the Proposed Model							
		Accuracy	MAE	RMSE	Class Wise F1-Score				
α	**β**				Panic	Fight	Congestion	Obstacle	Normal
0.95	0.05	79.19	4.55	5.78	78.44	64.30	44.75	57.17	86.89
0.90	0.10	77.6	4.89	6.08	72.3	70.19	29.08	57.43	85.22
0.85	0.15	78.77	4.84	5.9	76.33	67.01	23.75	58.88	86.26
0.80	0.20	78.91	5.37	6.58	68.8	67.64	32.37	59.23	86.42
0.75	**0.25**	**80.89**	**4.71**	**6.11**	**87.05**	**73.65**	**51.33**	**53.29**	**87.5**
0.70	0.30	78.82	5.14	6.52	69.2	76.04	14.14	58.46	86.04
0.50	0.50	76.72	3.61	4.79	49.77	71.51	0.00	53.05	85.19
0.30	0.70	75.86	3.53	4.72	37.87	60.44	33.59	53.29	84.68
0.20	0.80	72.59	3.50	4.73	24.84	52.76	14.8	60.47	81.96
0.10	0.90	71.66	3.39	4.52	0.00	47.86	0.28	56.7	81.99

So, based on these findings, the performance of the proposed model on the MED dataset for the value of α=0.75 and β=0.25 is used for performance comparison with state-of-the-art. The same procedure is adopted for the GTA dataset. Table 6.3 shows the performance of the proposed model on various values of α and β are illustrated.

Table 6.3: Experimental analysis of performance of the proposed model on various values of weighted loss parameters on the GTA dataset

Hyper-Parameter		Performance of the Proposed Model					
		Accuracy	MAE	RMSE	Class Wise F1-Score		
α	**β**				Panic	Fight	Normal
0.95	**0.05**	**85.85**	**13.08**	**16.32**	**85.85**	**13.08**	**16.32**
0.90	0.10	60.2	19.19	21.12	60.2	19.19	21.12
0.85	0.15	64.6	22.22	25.5	64.6	22.22	25.5
0.80	0.20	56.87	41.32	42.54	56.87	41.32	42.54
0.75	0.25	76.98	11.58	12.69	76.98	11.58	12.69
0.70	0.30	46.12	9.27	13.11	46.12	9.27	13.11
0.50	0.50	72.61	22.2	24.12	72.61	22.2	24.12
0.30	0.70	56.91	60.07	61.17	56.91	60.07	61.17
0.20	0.80	64.52	51.27	52.74	64.52	51.27	52.74
0.10	0.90	72.09	13.19	15.09	72.09	13.19	15.09

However, the same trend as in the MED dataset is not observed for the GTA dataset. This may be because the GTA is a computer simulation-based dataset and is not as real as the MED dataset. It can be observed from Table 6.3 that for α=0.95 and β=0.05, the performance of GTA is better in all respect and is used for comparative analysis with state-of-the-arts.

6.5.1 Results analysis for Crowd Behavior Prediction

6.5.1.1 The MED Dataset

The comparative analysis of results for crowd behavior classification with the state-of-the-art deep learning and conventional machine learning approaches are illustrated in Table 6.4. The numbers in bold letters represent highest values in Table 6.4. The proposed model obtained an overall accuracy of 80.89 % and a mean accuracy of 65.41% on the MED dataset [2]. The class-wise accuracy for Panic, Fight, Congestion, Obstacle, and Normal are 78.77%, 77.82%, 35.19%, 41.70%, and 93.59%, respectively. A limited number of models experimented on the MED dataset are available in the literature. Among the deep learning approaches, the model C3D-FC7 [147] achieves better accuracy than other listed deep models in Table 6.4. The C3D-FC7 [147] achieves an accuracy of 73.52%, having a mean accuracy of 51.22%.

Table 6.4: Comparative result analysis of proposed model for the CBP with state-of-the-art approaches for the MED dataset

Approaches		**Classification accuracy (%) per individual behavior classes**					**Mean-ACC (%)**	**Accuracy (%)**
		Panic	**Fight**	**Congestion**	**Obstacle**	**Normal**		
Deep Learning	V3G-FC7 [147]	80.72	37.41	31.18	**47.25**	71.35	53.58	62.71
	V3G-FC8 [147]	53.23	29.89	27.32	42.35	32.16	36.99	33.82
	C3D-FC7 [147]	**84.72**	32.93	16.16	29.61	92.69	51.22	73.52
	C3D-FC8 [147]	57.32	25.89	17.22	25.51	46.64	34.50	40.59
Conventional Machine Learning	HOT [2]	62.18	38.27	25.67	28.20	36.53	38.17	36.29
	DT [2]	74.82	30.47	23.43	27.94	36.88	38.71	36.10
Proposed Multitasking Model		78.77	**77.82**	**35.19**	41.70	**93.59**	**65.41**	**80.89**

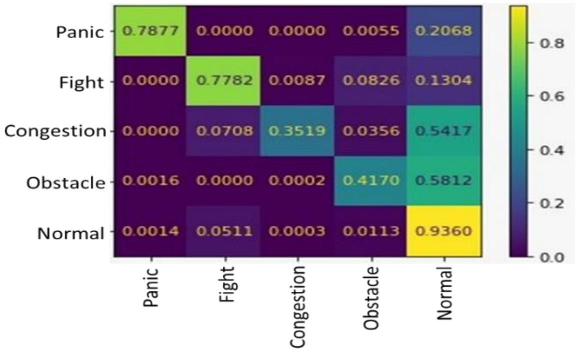


Figure 6.7 Confusion Matrix of the proposed model on the MED dataset

In contrast, the conventional machine learning techniques have very poor performance and achieved the highest mean accuracy of 38.80% using the HOG features [2]. Nevertheless, the proposed model outperforms the state-of-the-art conventional and deep learning techniques in different performance metrics. The confusion matrix of the proposed multitasking model on the MED dataset [2] is illustrated in Figure 6.7.

From the confusion matrix (Figure 6.7), it can be observed that, for congestion classes, the proposed model is more biased towards the normal crowd scenes on the MED dataset [2]. A similar kind of trend can be seen in the Obstacle class. However, the congestion and obstacle classes are not biased against each other. Due to similar appearance and motion patterns of both congestion and the obstacle with the normal scenes, most of the classes of congestion and obstacle are biased towards normal crowd scenes.

In addition to the analysis of the above results, the proposed model's performance is also compared with the recent work, i.e., the Novel-Descriptors [187] for crowd behavior prediction published in ICIP'21. Table 6.5 shows a comparison of the results of the proposed method with the Novel-Descriptors [187]. The numbers in bold letters represent highest values in Table 6.5.

Table 6.5: Comparison of results against Novel_Descriptor [187] *for the CBP on the MED dataset* [2].

Model	**Precision**	**Recall**	**F1 Score**
COF [187]	52.50	60.00	56.00
CD [187]	73.17	83.33	77.92
BD [187]	56.52	74.29	64.20
Proposed	**80.90**	**81.26**	**79.40**

The overall precision, recall and F1-Score of the proposed model are 80.90%, 81.26% and 79.40% which are far better than the Novel-Descriptor [187].

6.5.1.2 The GTA Dataset

The performance comparisons of approaches for the CBP on the GTA dataset are illustrated in Table 6.6. The values in bold letters represent highest values in Table 6.6. The proposed model achieves overall accuracy, mean accuracy of 85.85%, 72.64 respectively. The individual class accuracy for the Normal, Panic and Fight scenes are 86.15%, 31.79% and 100.00% respectively. When we compare the obtained results with the state-of-the-art method, it can be observed that the proposed model performs better. The confusion matrix of the proposed model on the GTA dataset is illustrated in Figure 6.8. An observation can be drawn from the confusion matrix that, the panic situations are almost equally biased with the normal crowd scene as well as Fight scenes. In contrast, the Spatial-Temporal Net obtains biased results on Fight scenes. Overall, the proposed model performs better than the state-of-the-art approaches.

Table 6.6: Comparison of results for CBP on the GTA dataset

Approaches	**Classification accuracy (%) per individual behavior classes**			**Mean-ACC (%)**	**Accuracy (%)**
	Normal	**Panic**	**Fight**		
Spatial-Temporal Net[23]	83.80	**61.20**	28.90	71.70	-
Proposed	**86.15**	31.79	**100.00**	**72.64**	**85.85**

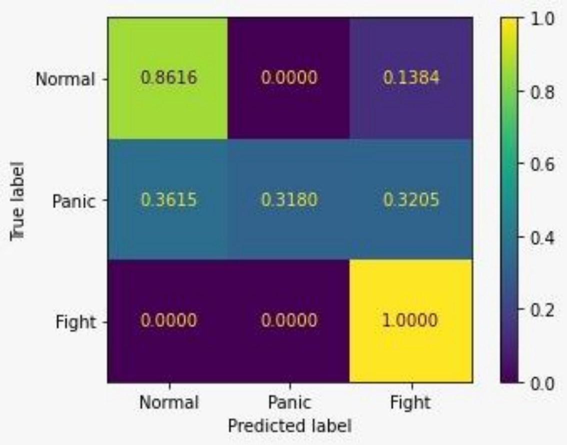


Figure 6.8: Confusion matrix of the proposed model on GTA dataset

6.5.2 Comparative Results Analysis with Crowd Counting Models

Five Five state-of-the-art models such as Real-CNN [97], CNN-Crowd [23], MCNN [27], Dense Crowd [22], and VGG-16 [188] are coded and implemented on the MED [2] and the GTA [146] for crowd counting. During the implementation of MCNN [27], Real-CNN [97], and VGG-16 [188], their output layer is replaced with a single neuron as the main focus of the study is single count regression using weak supervision and not on the density map-based regression. The input shape to these models [22], [23], [27], [97] is set to $(224 \times 224 \times 3)$.

6.5.2.1 The MED Dataset

The comparative analysis of the results with the crowd counting models is illustrated in Table 6.7. Values in bold letters represent best in Table 6.7. The proposed model achieves MAE and RMSE of 4.71 and 6.11 respectively on the MED dataset [2]. Whereas the state-of-the-art methods such as CNN-Crowd [23], Dense-Crowd [22], MCNN [27] and VGG-16 [188] obtain <MAE, RMSE> of <6.80, 12.86>, <8.65, 11.13>, <5.21, 7.84> and <7.54, 8.93> respectively. So, compared to the state-of-the-arts [23][22] [27] [188], the proposed model performs better in terms of MAE and RMSE.

Table 6.7: Comparison of results for crowd counting on the MED dataset [2]

Model	**MAE**	**RMSE**
CNN Crowd [23]	6.80	12.86
Dense Crowd [22]	8.65	11.13
VGG-16 [188]	7.54	8.93
MCNN [27]	5.21	7.84
Proposed	**4.71**	**6.11**

6.5.2.2 The GTA Dataset

The comparison of crowd counting results with the state-of-the-arts on the GTA dataset [146] is illustrated in Table 6.8. Values in bold letters represent best in Table 6.8. The proposed model achieves MAE and RMSE of 13.08 and 16.39 respectively. In contrast, the state-of-the-art approaches such as CNN-Crowd [23], VGG-16 [188], MCNN [27], Real-CNN [97] achieve <MAE, RMSE) of <33.91, 36.94>, <36.98, 39.75>, <32.06, 49.24> and <32.31, 49.89>, respectively. So, the proposed model performs better than the state-of-the-art approaches.

Table 6.8: Comparison of results for crowd counting on the GTA dataset [146]

Approaches	**Performance Metrics**	
	MAE	**RMSE**
CNN-Crowd [23]	33.91	36.94
VGG-16 [188]	36.98	39.75
MCNN [27]	32.06	49.24
Real-CNN [97]	32.31	49.89
Proposed Model	**13.08**	**16.39**

6.5.3 Ablation Study

Apart from the results analysis, an ablation study has been conducted to show the effectiveness of individual modules of the proposed model. The following modules are obtained from the proposed model for ablation study.

- CNN as Backbone: In this case, the whole DSCNN is replaced by the CNN layers as backbone.
- 2-Scale: Here, the proposed model with two scales of FABs i.e., FAB-3 and FAB-4 of scales (100×100) and (50×50) are used.

- 3-Scale: Here, the proposed model with the three scales of FABs i.e., FAB-2, FAB-3 and FAB-4 of scales (150×150), (100×100) and (50×50) are used.
- Backbone Only or No Attention: Here, no attention or FAB blocks are used.
- No Skip: Here, the proposed model without the skip connection in the FAB blocks are used.

The comparison of results of different modules during ablation study are illustrated in Table 6.9 for the MED dataset.

Table 6.9: Comparison of models during ablation study for the MED dataset [2]

Model	**Performance Metrics**										
							Class Wise F1-Score				
	Accuracy	**MAE**	**RMSE**	**F1-Score**	**Recall**	**Precision**	**Panic**	**Fight**	**Congestion**	**Obstacle**	**Normal**
CNN	78.00	5.16	6.62	76.25	77.88	78.00	64.3	72.08	41.25	54.76	85.32
2-Scale	78.50	4.50	**5.76**	77.21	78.51	77.84	75.43	69.55	32.02	**61.46**	85.90
3-Scale	78.85	4.74	6.04	76.40	78.85	79.37	79.11	60.90	38.10	54.95	86.59
No-Skip	77.74	4.76	5.97	74.87	77.78	77.02	73.93	64.83	19.22	53.36	85.98
Backbone	77.40	**4.69**	5.92	75.11	77.4	76.60	75.24	65.53	24.30	53.44	85.66
Proposed Model	**80.89**	4.71	6.11	**79.40**	**81.26**	**80.90**	**87.05**	**73.65**	**51.33**	53.29	**87.50**

During ablation study on the MED dataset, the modules like ‘CNN as backbone’, ‘2-Scale’, ‘3-Scale’, ‘No-Skip’ and ‘Backbone Only’ get <accuracy, MAE, RMSE> of <78.00%, 5.16, 6.62>, <78.50%, 4.50, 5.76>, <78.85, 4.74, 6.04>, <77.74%, 4.76, 5.97> and <77.40, 4.69, 5.92> respectively. When the backbone is replaced with the CNN, the classification accuracy as well as counting performance are also less than the proposed model. Among these modules, the 3-Scale model performs better. The backbone of the model performs poorly in terms of accuracy and slightly better in terms of MAE with respect to the proposed model. However, the proposed model takes advantages of different modules and performs better than the state-of-the-art approaches.

On the other hand, the results of the ablation study for different modules are illustrated in Table 6.10 for the GTA dataset. Values in bold letters represent best in Table

6.10. The model with CNN as the backbone performs better than other modules in terms of accuracy. i.e., 79.20%. However, in terms of MAE, the No-Skip model performs better than other modules of ablation study. However, none of these individual modules performs as better as the proposed multitasking model.

Table 6.10: Comparison of models during ablation study for the GTA dataset

Model	**Performance Metrics**								
	Accuracy	**MAE**	**RMSE**	**F1-Score**	**Recall**	**Precision**	**Class Wise F1-Score**		
							Panic	**Fight**	**Normal**
CNN	79.20	41.14	43.03	77.95	79.20	83.94	81.40	48.25	82.16
2-Scale	72.08	46.76	47.64	70.70	72.08	83.15	72.03	48.25	75.28
3-Scale	54.69	59.59	60.53	50.84	54.69	57.77	66.28	45.55	35.51
No-Skip	46.88	18.88	23.00	41.84	46.88	47.40	59.57	24.60	27.22
Backbone	55.42	36.15	45.49	53.33	55.42	54.87	79.75	23.32	32.72
Proposed Model	**85.85**	**13.08**	**16.32**	**84.24**	**85.85**	**87.55**	**88.32**	**48.25**	**89.44**

6.6 Conclusion

This chapter proposed a multitasking crowd analysis model which performed two important tasks: Crowd Counting and Crowd Behavior Prediction. A multitasking crowd analysis dataset using the available benchmarks crowd behavior datasets was also developed. The model outperforms recent state-of-the-art approaches as far as crowd behavior classification and crowd counting are concerned. For crowd behavior classification, the proposed model improved the mean accuracy by 22.01% and 1.31% concerning the SOTA for the MED and the GTA dataset, respectively. Similarly, for crowd counting, the proposed model improved the MAE by 6.17% and 59.51% concerning SOTA for the MED and the GTA dataset, respectively. However, the model bias toward normal crowd scenes when it deals with congestion and panic crowd scenes of the real-world scenario-based dataset, i.e., the MED dataset. On the other hand, the proposed model is almost equally biased towards normal and fight scenes when it deals with panic scenes of computer-simulated datasets, i.e., GTA. So, the future study will focus on addressing this limitation by proposing a more sophisticated method and model.