CHAPTER 2
LITERATURE REVIEW

In recent years, vision-based crowd analysis has gained colossal attention to maintain crowd security and safety. Conventional machine learning and deep learning approaches have been used to accomplish different tasks of crowd analysis. This thesis focuses on three crucial research domains of crowd analysis: crowd count and density estimation (CCDE), crowd congestion-level analysis (CCA), and crowd behavior analysis (CBA). The primary objective of this chapter is to provide a detailed survey of the existing methods in each of the selected research areas, understand the current research trends, and find out possible research gaps.

2.1 Literature Review on Crowd Counting and Density Estimation Approaches

Over the past few years, several research communities have shown a massive interest in crowd count and density estimation (CCDE). The CCDE is also an interdisciplinary research area that spans several research domains such as cell counting [16], crop counting [17], and vehicle counting [18]. The CCDE constitutes two different words, i.e., "crowd count" and "density estimation." The former defines the number of people present in the crowd scene, whereas the latter resembles a technique to perform crowd count. However, these two words are combined in the literature that counts the people present in the crowd scene. The CCDE is a challenging task, specifically in the presence of occlusion, cluttered background, illumination changes, varying crowd densities, and crowd shape change due to perspective distortion. Various techniques have been developed to handle these challenges, such as detection-based, regression-based, and density map-based regression. A comprehensive literature review is required to understand the current research direction and identify possible research gaps. Therefore,

this literature review mainly discusses the taxonomy of CCDE approaches, followed by a brief review of image-based and video-based CCDE techniques.

2.1.1 Taxonomy of CCDE approaches

The CCDE approaches can be categorized in many ways. Followings are the taxonomy of CCDE approaches based on four criteria such as,

- Mode of implementation.
- Dealing with labelled data.
- Learning mechanism.
- Dataset modality.

2.1.1.1 Taxonomy of CCDE-based on Mode of Implementation

As mentioned in Figure 2.1, the vision-based CCDE approaches can also be classified into three types based on the mode of implementation, such as

- Detection-based Approaches.
- Regression-based Approaches.
- Hybrid approaches.

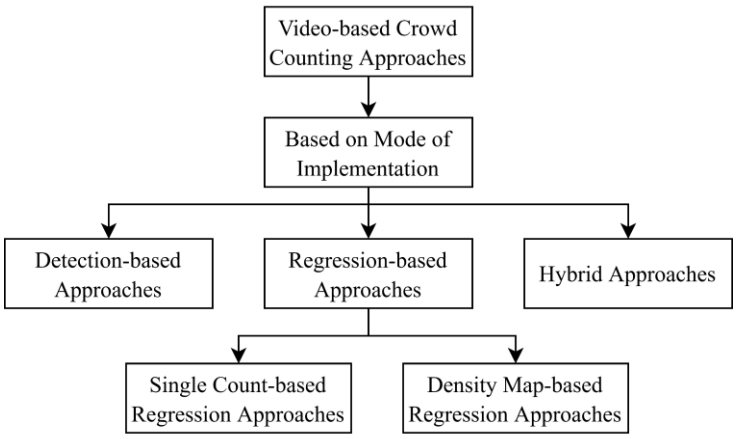


Figure 2.1: Categorisation of vision-based crowd counting approaches based on mode of implementation

The detection-based approaches [19]–[21] utilize body part detectors to localize people present in the crowd scene, which is used to predict the total count of the crowd. Detection-based approaches perform better in low crowd densities, but their performance gets degraded in very dense crowd scenes. The regression-based CCDE approaches [22]–[28] overcome such drawbacks. These approaches are divided into single count-based regression approaches [22]–[25] and crowd density map-based regression approaches [26]–[28]. The single count-based regression approaches extract meaningful features from the crowd scene and map onto the ground-truth crowd count values. Several researchers have applied the count-based regression approach at the patch-level and global or scene-level. The advantage of such an approach is that it considers global features for crowd counting but does not consider the crowd distribution during regression.

On the other hand, the crowd density map-based regression approaches consider the crowd distribution during regression by mapping the scene-level features onto the ground-truth crowd density maps. However, this approach requires ground-truth head annotations for generating crowd density maps which is a laborious task. Researchers are now adopting different combinations of counting approaches to enhance the crowd counting process, also known as hybrid approaches. He *et al.* [29] recently utilized a combined detection-based technique and DMR approach for crowd counting. The authors adopted YOLO V3 as a detection-based approach.

2.1.1.2 Taxonomy of CCDE-based on dealing with labelled data

The vision-based CCDE approaches can also be classified into four categories based on dealing with the labeled dataset. Figure 2. 2 shows the vision-based CCDE can be classified into,

- Supervised [30]–[34]

- Semi-supervised [35]
- Weakly-supervised [36]
- Self-Supervised [37]

The supervised models use all the labeled datasets during training. However, labeling the datasets is a tedious and time-consuming task. So, various research communities have proposed weakly supervised, self-supervised, and semi-supervised models to deal with the lack of labeled datasets. The semi-supervised models are trained using a smaller number of labeled training samples. Recently, Olmschenk *et al.* [38] proposed a semi-supervised model for CCDE. Authors designed dual objective semi-supervised GAN, which has two outputs: one for regression and another for classification. The classification output determines whether the input sample is real or fake. The model performs regression in a supervised manner and classification in an unsupervised manner. Combining these two methods, the discriminator can learn more robust features and work well in a few labeled data.

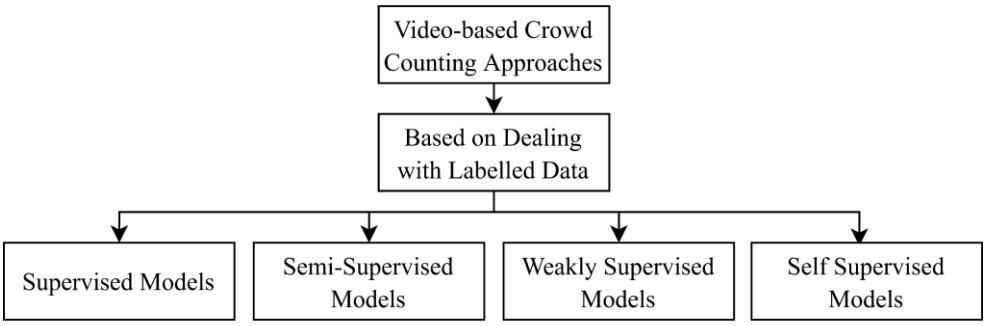


Figure 2.2: Categorisation of vision-based crowd counting approaches based on dealing with labelled data

The weakly supervision models have labeled data where some samples are falsely labeled. In CCDE, the DMR approaches require ground-truth density maps, which are obtained from head annotations from the crowd scene. However, the head annotations are done manually and are prone to error. So, to address such type of problem, weak

supervision models for CCDE have been proposed in the literature. Based on this issue, Lei *et al.* [39] proposed a weak supervision model for CCDE. The authors performed several auxiliary tasks to train the CCDE model and achieved better density map prediction capability in the presence of a weakly annotated labeled dataset.

On the other hand, the self-supervised CCDE models learn the hidden structure of the crowd scenes and can thus work on the larger dataset without relying on the labeled datasets. Liu *et al.* [37] observed that the total number of people in any sub-part of the image is less or equal to the total number of people in the image. Based on this finding, the authors [37] proposed a self-supervised CCDE model that works well on large-scale unlabelled datasets.

2.1.1.3 Taxonomy of CCDE-based on Learning Mechanism

According to Figure 2. 3, the vision-based CCDE approaches can also be divided into two types based on the learning methodology: first conventional machine learning approaches, and second, deep learning approaches.

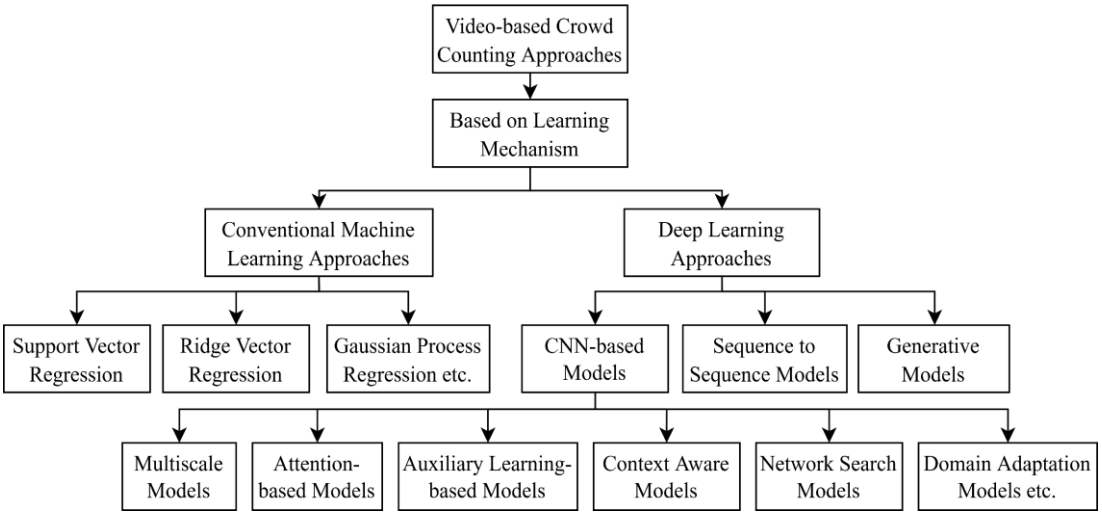


Figure 2.3: Categorisation of vision-based crowd counting approaches based on learning mechanism.

The conventional machine learning approach utilizes different regressions to perform crowd counting. Such approaches mainly rely on handcrafted features like

LBP[40], GLCM [41][42] Fourier Features[41][42], Scale Invariant Feature Transform (SIFT) [42], HOG[43], and motion features [44] for learning mechanisms. On the contrary, the deep learning-based CCDE approaches utilize different deep models for crowd counting. Deep learning techniques using CNN [12], RNN (Bidirectional-LSTM, Conv-LSTM [12], Transformers[45]), Encoder-Decoder [46], and GANs[38] have been proposed in the literature. Different types of CNN-based models are vastly used for CCDE to address different issues and challenges. For example, multiscale models [30], [47] have been designed to handle scale variation issue in the crowd scene, attention-based models [48] are designed to give more attention to different parts of the crowd image, auxiliary learning-based models [49][50] have been designed to infuse the learning of related tasks such as crowd density classification or single count regression in CCDE to learn efficient crowd density maps. Similarly, context-aware models [51][52], network search model [53] and domain adaptation models [54][55] using CNN have also been proposed for CCDE.

2.1.1.4 Taxonomy of CCDE-based on Dataset Modality

In Figure 2.4, the CCDE approaches can be categorized into two types based on the available dataset modality, such as

- Image-based approaches
- Video-based approaches.

Most image-based CCDE approaches utilize different types of spatial features for counting, while the video-based CCDE approaches extract spatial-temporal features from the video sequences and then perform crowd counting. Both conventional machine learning and deep learning approaches have been developed for both the modality of crowd datasets.

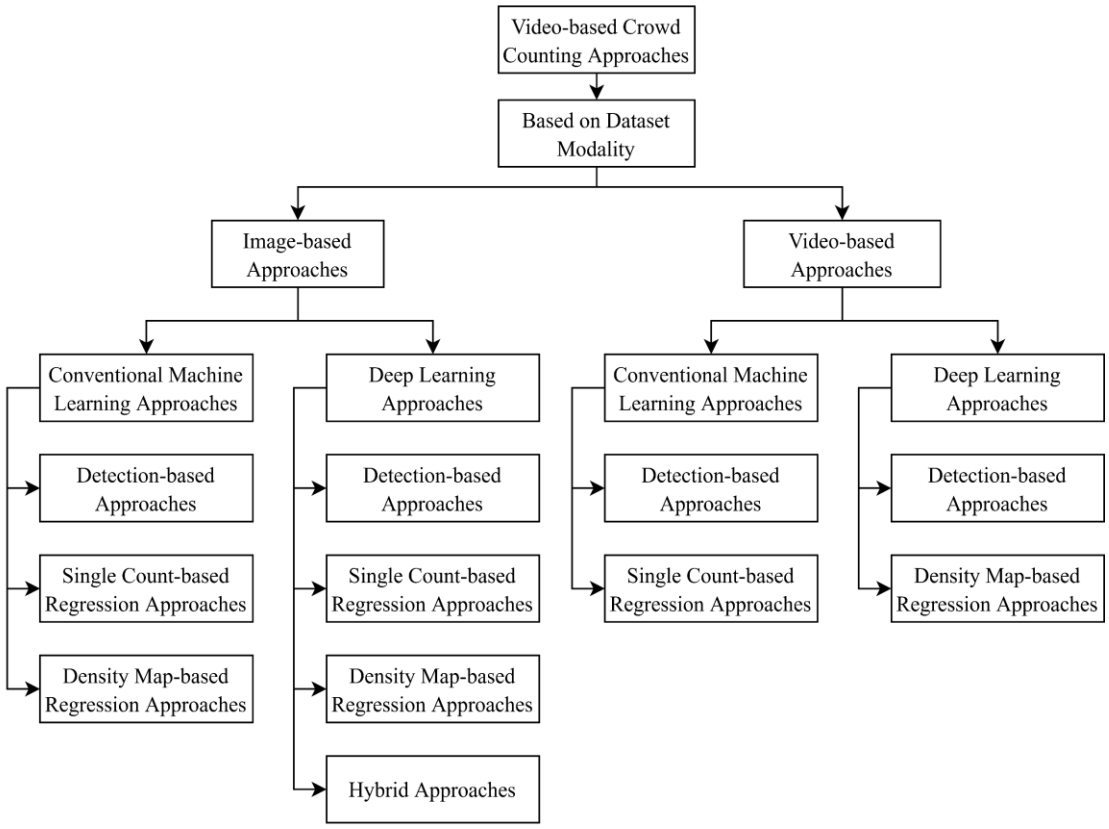


Figure 2.4: Categorisation of vision-based crowd counting approaches based on dataset modality

Methods like detection-based, single-count, and density map-based regression approaches have been vastly explored regarding image-based CCDE using conventional machine learning and deep learning. However, some hybrid approaches have been proposed for image-based CCDE using deep learning. Compared to image-based CCDE, fewer works have been found for video-based CCDE. Detection-based and single count regression-based approaches have been proposed for video-based CCDE using conventional machine learning. On the other hand, the deep learning methodology for video-based CCDE has been exploited in detection-based and density map-based regression approaches.

This review is mainly focused on discussing the current research trends and possible research gaps as far as image and video-based CCDE. The following subsections discuss the review of these approaches in detail.

2.1.2 Review on Image-based CCDE

The image-based CCDE approaches extract meaningful spatial features from the crowd image and adopt a learning mechanism to regress on the ground-truth crowd counts. Approaches like detection-based [56], regression-based [22]–[24], [57]–[60], and hybrid approaches have been explored in the literature to show their effectiveness in crowd counting. The following subsections discuss various methods and models of image-based CCDE based on conventional and deep learning techniques.

2.1.2.1 Conventional Machine Learning Approaches for Image-based CCDE

Most conventional approaches are detection-based, and very few are based on single-count regression. The following subsection illustrates a brief review of these state-of-the-art.

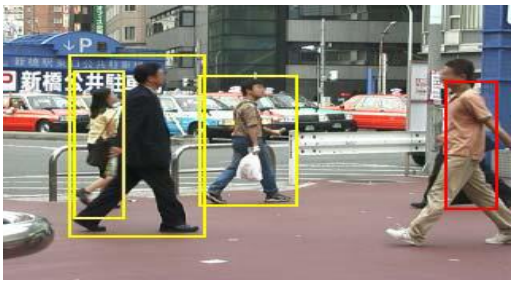
2.1.2.1.1 Detection-based Approaches

The detection-based approaches perform crowd counting by locating each pedestrian fully or partially in the crowd image. Such approaches extract hand-crafted features from the crowd scenes. Dalal *et al.* [19] obtained feature descriptors using a Histogram of Oriented Gradients (HOG) for human detection. Leibe *et al.* [61] proposed a pedestrian detection model which can perform well for partially occluded pedestrians. The authors extracted scale-invariant difference of gaussian (DoG) interest points from all the training images. Wu *et al.* [20] proposed “Edgelet" features to detect a human from the still image. The model can detect a human in the presence of partial occlusion. Sabzmeydani et al. [62] extracted “Shapelet” features using oriented gradients to detect pedestrians from still images. Yan *et al.* [63] proposed a model to detect multiple pedestrians in the crowd images. Authors utilized HOG features to represent the images. Felzenszwalb et al. [43] extracted HOG features from the image whose dimensionality is reduced using PCA to detect pedestrians. Dollar *et al.* [64] extracted a multiscale gradient

histogram to detect pedestrians. Li *et al*. [65] proposed a detection technique for detecting head and shoulder parts. The authors extracted HOG features from the image. Figure 2.5 illustrates some examples of detection-based approaches based on full-body and part-of-body detection methods.

The detection-based approaches follow the sliding window technique to classify or detect pedestrians from the still images. Dalal *et al*. [19] used SVM to detect pedestrians. Yan *et al*. [63] and Felzenszwalb *et al*. [43] adopted a Latent SVM (LSVM) to detect pedestrians. On the other hand, some works [20], [62] have also used AdaBoost to detect pedestrians. Dollar et al. [64] proposed a pyramidal classifier to detect pedestrians. Pedestrian detection using the clustering technique [61] has also been proposed in the literature.

a) Full body detection results obtained by Wu et al. [20].



b) Full body detection results obtained by Leibe et al.[61].

c) Head shoulder detection results obtained by Wu et al. [20].

d) Head shoulder detection results obtained by Leibe et al.[61].

Figure 2.5: Some results of detection-based approaches obtained from literature.

Although these models are good at detecting humans in very low density and partially occluded crowd scenes but do not perform well in dense crowd scenes with high occlusions.

2.1.2.1.2 Single Count-based Regression Approaches

Few works have been reported under this category in the literature. Bansal et al. [66] proposed a hybrid feature fusion model for crowd counting in dense image scenarios. Authors fused features like SIFT, Fourier analysis, GLCM, wavelet decomposition, and low confident head detections and adopted SVR for crowd counting. However, scale variation and clutter background affect the performance of this model. Chouhan *et al.* [41] proposed a similar approach for crowd counting in still images. Features like SIFT, Fourier, and head shapes are extracted and inputted to SVR for counting people in the crowd scenes. Lamba *et al.* [67] proposed a still image crowd counting approach by extracting texture-based manifold feature extraction followed by SVM to estimate patch-based crowd count. The authors applied Gaussian Markov random field to obtain uniformity in crowd count. The model performs better in low-density crowd images than in highly dense crowd scenes.

Recently, Saleem *et al*. [42] extracted and fused multiple local handcrafted features from the still images. Features like statistical spectral features using Fourier transform, texture features (like GLCM, HOG, and LBP), wavelet features using Haar transform, segment features, and structure features are extracted from the image patches. The fused features are given to the ensemble regression (EREG) to perform crowd counting. However, the experimental analysis is minimal, and the difference between MAE and MSE is very high, which is not desirable for a robust prediction system.

2.1.2.1.3 DMR-based Approaches

The DMR-based image-based CCDE uses conventional machine learning approaches to extract meaning full handcrafted features from the images and map them on ground truth crowd density maps by adopting a regression function. Lempitsky *et al*. [68] learn the density function by mapping the image's pixel on ground-truth head points.

The Maximum Excess over Sub Arrays (MESA) distance was used to find the loss between the predicted and ground-truth values. The authors adopted convex quadratic programming to optimize the model.

Fiaschi *et al*. [69] followed a similar approach as proposed by Lempitsky *et al*. [68], but with few changes. Instead of working on raw data, the authors computed standard filter responses from the raw images followed by non-linear mapping. The non-linear mapping is demonstrated by using regression forest (RF). Pham *et al.* [70] proposed a DMR approach known as Count Forest (CF). The authors extracted low-level features from the image patches, followed by patch-wise density estimation by developing a CF dictionary. The final density maps are obtained by combining the results obtained from all the patches. However, such an approach lacks in handling scale variation and clutter background issues in the crowd scene.

Wang *et al*. [71] proposed a density estimation-based visual object counting (DE-VOC) model. The simple model does not address the crowd counting system's main issues. Again, Wang *et al.* [72] proposed a manifold VOC (M-VOC) approach for crowd counting. Authors adopted a locally linear embedding technique to reconstruct the object density by preserving local geometry.

The state-of-the-art DMR-based image CCDE using conventional machine learning approaches provides different solutions for crowd counting, but such approaches suffer from severe drawbacks. These approaches neglect temporal dependencies between frames while working on video datasets. The feature extraction approaches are handcrafted and are not adaptive to varying crowd densities. Significant issues like shape variation and the effect of cluttered backgrounds are not handled. Moreover, the experiments are conducted on datasets with crowd counts of less than 100.

2.1.2.2 Deep Learning Approaches for Image-based CCDE

The deep learning approaches have immensely been used for image-based crowd counting. Most of the approaches are regression-based, and very few approaches are developed for a detection-based techniques using deep learning techniques. In this section, a comprehensive review is performed on different models of image-based CCDE.

2.1.2.2.1 Detection-based Approaches for Image-based CCDE

Deep learning approaches using CNN [73]–[78] and LSTM [79] are explored for pedestrian detection in crowd scenes. Gao *et al.* [73] proposed AdaBoost and a CNN-based model for people counting based on head detection. The authors used the AdaBoost algorithm to obtain head region proposals. Then the authors developed a CNN-based model to classify whether the region proposal belongs to the head or not. Based on the classification results, crowd count is then performed. The main drawback of such an approach is that it is time-consuming. Stewart *et al.* [79] proposed an end-to-end network using GoogLeNet and LSTM to detect a human in the crowd scene. The model without standard post-processing provides better precision. Zhang *et al.* [76] proposed occlusion-aware pedestrian detection in crowd scenes. The authors proposed R-CNN to achieve their objective function. Huang *et al.* [75] proposed a body structure aware deep model for crowd counting in still images. The authors infuse semantic body part information for crowd counting.

Shami *et al.* [77] proposed crowd counting in dense images using the people head detection method. The authors proposed the CNN model for head detection from the image patches. Figure 2.6 illustrates the results produced by Shami *et al.* [77]. The model provides better results in different crowd datasets. However, the detection accuracy of the proposed model in dense crowd scenes is inferior because of severe occlusion and perspective distortion.

Figure 2.6: Sample of crowd counting results based on people head detection [77]

2.1.2.2.2 Single-Count-Based Regression Approaches

Generally, the count-based regression approaches using deep learning techniques extract meaningful spatial features and maps on ground truth crowd count values for image-based CCDE. Most approaches under this category utilize CNN and LSTM to achieve the objective. Wang *et al.* [22] proposed a multilayer of CNN to perform crowd counting. The authors tried to minimize background effects by adding negative samples during training. The negative samples are those samples that do not have any crowd. However, the approach looks like ridge regression and does not handle scale variation due to perspective distortion.

Similarly, Hu *et al*. [23] proposed a deep CNN structure to perform dense crowd counting from still images. Authors developed a crowd counting dataset, AHU-Crowd, whose crowd density ranges from 58 to 2201; critical challenges like scale variation and minimizing background influence are not addressed. Shang *et al*. [24] proposed an end-to-end crowd counting model which jointly learns the local and global crowd counts. The end-to-end model is developed using a pre-trained GoogLe-Net [25] followed by an LSTM to learn local crowd counts. The local counts are then inputted into a dense layer

for global crowd counting. Zhang *et al*. [80] proposed a crowd counting model called count-net, developed using CNN. The model takes advantage of auxiliary learning to perform crowd counting. Recently, Li *et al*. [81] proposed a weak supervision crowd counting model using the Swin Transformer (CCST). The model can handle scale variation due to perspective distortion, but the authors do not take any measures to minimize the effect of the background, and the performance of the model needs to be improved in the presence of low-density crowds.

Similarly, Savner *et al*. [45] proposed a weakly supervised crowd counting model using Transformer (known as CrowdFormer). The model works well as far as cross-scene crowd counting is concerned. However, the scale issue and background minimization are not addressed in CrowdFormer. Table 2.1 shows a comparison of the above approaches. Although both supervised and weakly supervised approaches have been developed for this category, these approaches do not consider the spatial distribution of crowd during crowd counting and also do not consider any measures to minimize background effects. Wang *et al*. [22] tried to minimize the effect of clutter background by learning from non-crowd scenes, but still, every scene has a background that is to be removed during crowd counting.

2.1.2.2.3 Density Map-based Regression Approaches

The DMR-based approaches extract spatial features from the still image and perform regression on ground-truth crowd density maps to perform CCDE. Most of the existing single-image CCDE approaches address the scale variation issue in the crowd scene, and very few take measures to minimize the effect of cluttered background and handle cross-scene crowd counting.

Table 2.1: Comparative analysis of Image-based CCDE approaches

References	**Learning Type**	**Implementation Mode**	**Challenges Handled**			**Attention**	**Limitations**
			Scale-Variation	**Cluttered Background**	**Cross-Scene**		
Leibe *et al.* [61]	Supervised and Conventional ML	Detection	×	✓	×	×	The detection is done through series of iterative evidence aggregation step. Trained and Tested with limited samples with only two backgrounds[83].
Dollar *et al*. [64]	Supervised and Conventional ML	Detection	✓	×	×	×	Performance degraded in presence of dense crowd with severe occlusion.
Li *et al*. [65]	Supervised and Conventional ML	Detection	×	✓	×	×	Performance degraded in presence of very dense crowd with severe occlusion.
Bansal *et al*. [66]	Supervised and Conventional ML	SCR	×	×	×	×	Poor performance due the effect of cluttered background and scale issues.
Lamba *et al.* [67]	Supervised and Conventional ML	SCR	×	×	×	×	Poor performance in very dense crowd scenes.
Saleem *et al*. [42]	Supervised and Conventional ML	SCR	×	×	×	×	Difference between MAE and MSE is very high which is not desirable for a robust detector.
Lempitsky *et al*. [68]	Supervised and Conventional ML	DMR	×	×	×	×	Directly maps the image on to the ground truth density maps. The model is more like ridge regressor.
Pham *et al*. [70]	Supervised and Conventional ML	DMR	×	×	×	×	Does not handle the issues like scale variation and cluttered background.
Wang *et al.* [72]	Supervised and Conventional ML	DMR	×	×	×	×	Does not handle the issues like scale variation and cluttered background.
Wang *et al*. [22]	Supervised and Deep Learning	SCR	×	✓	×	×	Model acts more like ridge regression. Model does not consider spatial distribution of crowd.
Hu *et al.* [23]	Supervised and Deep Learning	SCR	×	×	×	×	The model does not consider the spatial distribution of the crowd during counting.
Shang *et al*. [24]	Supervised and Deep Learning	SCR	×	×	×	×	The model does not consider the spatial distribution of the crowd during counting.
Zhang *et al.* [80]	Supervised and Deep Learning	SCR	×	×	×	×	The model does not consider the spatial distribution of the crowd during counting.

Li *et al*. [81]	Weakly Supervised and Deep Learning	SCR	✓	×	×	✓	Does not perform well in the low-density crowd scenes.
Savner *et al*. [45]	Weakly Supervised and Deep Learning	SCR	✓	×	✓	✓	The model does not consider the spatial distribution of the crowd during counting. Scale variation and minimizing background effect are still unexplored.
Luo *et al.* [26]	Supervised and Deep Learning	DMR	×	×	×	×	Scale variation and minimizing background effect are still unexplored.
Wang *et al.* [82]	Supervised and Deep Learning	DMR	×	✓	×	×	Scale variation issue is not handled.
Zhang *et al.* [27]	Supervised and Deep Learning	DMR	✓	×	×	×	The model requires lots of pre-training and the regression exhibits as ridge type.
Rubio *et al.* [28]	Supervised and Deep Learning	DMR	✓	×	×	×	Don't take any measures to minimize effect of cluttered background.
Boominathan *et al*. [84]	Supervised and Deep Learning	DMR	✓	×	×	×	Limited in scale representation.
Sam *et al.* [85]	Supervised and Deep Learning	DMR	✓	×	×	×	Requires lots of pre-training. Ignores large-scale contextual information.
Zhang *et al.* [86]	Supervised and Deep Learning	DMR	✓	×	×	×	The fusion of multiscale features in the middle of the network ignores the further layer's potentiality.
Gao *et al.* [35]	Supervised and Deep Learning	DMR	✓	×	×	✓	Does not take any measures to minimize effect of cluttered background.
Kumar *et al.* [88]	Supervised and Deep Learning	DMR	×	×	×	×	Does not handle issues like scale variation and effect of cluttered background.
Sindagi *et al.* [89]	Weakly Supervised and Deep Learning	DMR	×	×	✓	✓	Does not handle issue like scale variation due to perspective distortion.
Wang *et al.* [90]	Supervised and Deep Learning	DMR	✓	×	×	×	The generation of crowd density maps for different datasets is inconsistent, hence the model overfitting the datasets.
He *et al.* [29]	Supervised and deep learning	Hybrid	✓	×	×	×	The experimental analysis is limited with one dataset.

Luo *et al*. [26] proposed a crowd counting model using CNN for high-density crowd scenes. Authors proposed a feature fusion strategy for crowd density estimation, but such a model lacks in addressing scale variation due to perspective distortion. Wang *et al*. [82] designed a de-background detail CNN (DDCN) to minimize background influence for crowd counting. The DDCN performs the crowd counting on the crowd scene's detail layer (information). Some constructive works have been done to handle crowd shape change due to perspective distortion in the still images. Zhang *et al*. [27] developed a multicolumn CNN to address the crowd scale variation issue by arguing that multi-CNN structures with different sizes of receptive fields can obtain scale-invariant features from the scene.

According to [28], MCNN [27] suffers from several drawbacks, like it requires much pre-training of the columns, and the model behaves as a ridge regressor. Rubio *et al*. [28] proposed two counting models, Counting-CNN and Hydra-CNN, where the latter addresses the scale variation issue. The crowd counting is done by performing non-linear regression. Boominathan *et al.* [84] handled the scale variation issue by extracting shallow and deep CNN features from the image pyramids, but this approach has limited scale representation, and ground-truth density maps do not consider the head points spatial distribution.

Zeng *et al*. [91] addressed the scale variation issue by designing a single-column CNN with different multiscale blocks. The advantage of such a model is that it contains fewer network parameters. Sam *et al*. [85] proposed Switch-CNN, which constitutes four CNNs, among which three are used for regression, and the fourth one is a switch-classifier. The Switch-CNN handles the scale variation issue, but the approach requires lots of pre-training and results in a greedy solution. On the other hand, Zhang *et al*. [86] handle the scale variation issue by extracting multi-layer features of a single column CNN

(also called scale-adaptive CNN (SA-CNN)) with a fixed kernel size. However, Zou *et al*. [92] argued that constructing multiscale features in the middle of the network ignores the further layer's potentiality. Ranjan *et al*. [93] found that most existing counting systems do not generate high-quality density maps, which motivated the authors to propose an iterative crowd counting model to generate high-resolution crowd density maps. The model utilized two-column CNNs, among which the first column generates low-resolution crowd density maps, which the second column uses to obtain high-resolution crowd density maps.

Li *et al*. [94] proposed a congested scene recognition network (CSR-Net) that mainly focused on developing high-quality crowd density maps. The CSR-Net comprises two components: a frontend network containing a CNN built on VGG-16 and a backend network containing a dilated CNN. The backend network uses the extracted frontend features to generate high-quality crowd counting density maps. Deb *et al*. [34] developed a multicolumn aggregated dilated-CNN (AMDCN) to handle scale variation due to perspective distortion. Gao *et al*. [35] identified that approaches like [30], [85], [95] depend on the local-appearance features and neglect large-scale contextual information for CCDE. Based on this, the authors developed a Spatial Channel-Wise Attention Module (SAM) for the CCDE. Kumar *et al.* [88] adopted a multi-task paradigm for CCDE. The authors utilized the VGG-16 as the frontend, shared by a dilated CNN for density estimation and a CNN for density classifier. The drawback of such a system is that authors do not handle scale variations.

Different variants of encoder-decoder models have also been developed for CCDE. A variant like the Trellis Encoder-Decoder model is proposed by Jiang *et al.* [96] for CCDE. The model extracts multiscale features by maintaining similarities in the local coherence and correlation between different density maps. Recently, Sindagi *et al.* [89]

proposed a hierarchical attention-based crowd counting network (HA-CCN) motivated to perform well in highly congested crowd scenes. The HA-CCN is built on the VGG-16, followed by two attention networks: a spatial attention module (SAM) and a global attention module (GAM). The SAM improves the low-level features by fusing multi-spatial segmented features, while GAM improves the channel-wise information in the higher layers. Wang *et al.* [90] proposed a multi-density map fusion strategy for crowd counting. The model is built on VGG-16 and handles crowd scale variation in the scene. Three different parallel branches follow the VGG-16. Each branch extract features on a different scale. Instead of fusing multiscale features for final density map generation, the authors fused multiscale generated density maps for final crowd counting. The problem with such a model is that they used different density map generation techniques for different datasets. Shi *et al.* [97] proposed a real-time crowd counting model by developing a compact-CNN (CCNN). The model can extract multiscale features to handle scale variation issues and process the frames in real-time. The comparative analysis of the above approaches is illustrated in Table 2.1.

2.1.2.2.4 Hybrid Approaches

A very few hybrid approaches have been developed for image-based CCDE. Liu *et al.* [98] proposed a hybrid model to handle crowd scale variation and varying crowd density. The hybrid network constitutes two sub-modules: A Detection Network (DNet) and an Encoder-decoder Network (ENet). The DNet detects several detectable humans and segments the crowd scene into detectable and non-detectable ones. On the other hand, the ENet estimates crowd density from the non-detectable part by employing a modified Xception model and a dilated-transposed network. He *et al.* [29] proposed a hybrid approach using detection and DMR methods. The detection-based module contains the

YOLO V3 to detect the people near the camera, and the DMR approach utilizes an inception-dilated CNN to estimate people far from the camera.

The above-reviewed approaches are made for single image crowd counting. These models do not consider temporal features in the video dataset and treat the frame as isolated. The video-based CCDE approaches have been developed to overcome the drawbacks of single-image crowd counting. In the following sub-section, we briefly review video-based CCDE approaches.

2.1.3 Review on Video-based CCDE approaches

The video-based CCDE approaches overcome the drawback of static image-based CCDE by considering the temporal consistency between sequences of frames. In this section, a brief review of both conventional machine learning and deep learning approaches has been conducted, and the comparative analysis is illustrated in Table 2.2.

2.1.3.1 Conventional Machine Learning

Most conventional machine learning approaches extract the foreground of the scene followed by feature modeling and regression or classification (for detection) for crowd counting. This section reviews various methods and models using conventional machine learning, which are based on detection and single count-based regression approaches.

2.1.3.1.1 Detection-based Approaches

Some state-of-the-art detection-based approaches have been proposed in the literature. Most of these approaches depend on the motion and appearance features to detect a human from the crowd scene. Viola *et al.* [44] extracted features like image intensity and motion features and fused them to detect humans using AdaBoost classifiers. However,

such an approach detects humans at the rate of 4 frames per second, and the performance degrades in dense crowd scenes with occlusion.

Wu *et al*. [99] proposed multiple human detection techniques by modeling partial or complete body parts using edge-type features. Authors adopted AdaBoost to classify humans from the scene. The model can work well in the presence of partially occluded humans but not in very dense crowd scenes. Figure 2.7: a) shows an example of the detection results of Wu *et al*. [99].

a) Part detection responses (yellow for full-body; red for head shoulder; purple for torso; blue for legs Wu et al. [99]

b) Pedestrians detection in the crowd [100] scene

Figure 2.7: Some samples of the results of detection-based approaches using conventional machine learning techniques

Walk *et al*. [100] detected humans in the crowd scene by utilizing the motion and spatial features from the video sequences. The authors extracted features like HOG, HOF, and Colour Self Similarity (CSS) from the frame sequences and inputted these features into the MLPBoost algorithm to detect pedestrians. The model can also work in the low density partially occluded crowd scenes. Gal *et al*. [101] detected humans from the video scene using Hough forest. Xu *et al*. [102] proposed multiple human detections in real-time video surveillance applications based on detecting human heads. Authors extracted handcrafted features like a color histogram and oriented gradients and inputted them into

a cascaded AdaBoost classifier to detect human heads. The model can work well in moderate crowd scenes. Subburaman *et al*. [103] proposed people counting in the crowd video dataset by detecting human heads from the crowd scenes. Authors proposed cascading of boosted features to detect human heads. However, the model can work in less occlusion moderate crowd scenes. Yoon *et al*. [104] proposed a conditional marked point process to detect and count humans from the crowd scenes. The authors used a mixture of Bernoulli shapes to determine the shape size from a given location in the frame to count humans. The model can work well only in low and moderate crowd scenes. Ge *et al*. [105] proposed a sampling-based approach to detect pedestrians for crowd counting. The authors proposed a reversible jump Markov Chain Monte Carlo (RJMCMC) sampling method to achieve the objective function. However, the model cannot localize people in dense crowd scenes.

Although different detection-based approaches have been proposed using conventional machine learning techniques. The main focus is to detect an object by minimizing the effect of the background, but the performance of such approaches gets degraded in the presence of dense crowd scenes because of severe occlusion.

2.1.3.1.2 Single-Count Regression Approaches

The single count-based regression using conventional machine learning techniques extracts meaningful handcrafted features from the patch or whole image and then maps these features onto ground truth crowd count values. So, these approaches follow two steps first feature extraction and second regression. The handcrafted features such as HOG, HOF, LBP, SURF, shape features, spatial-temporal interest points, etc., have been extracted to perform regression on the ground truth values.

Table 2.2: Comparative analysis of Video-based CCDE approaches

References	Learning Type	Implementation Mode	Challenges Handled			Attention	Limitations
			Scale-Variation	Cluttered Background	Cross-Scene		
Wu et al. [99]	Supervised and Conventional ML	Detection	×	✓	×	×	Poor performance in dense crowd scenes.
Walk et al. [100]	Supervised and Conventional ML	Detection	×	✓	×	×	Poor performance in dense crowd scenes.
Subburaman et al. [103]	Supervised and Conventional ML	Detection	×	✓	×	×	Poor performance in dense and occluded crowd scenes.
Yoon et al. [104]	Supervised and Conventional ML	Detection	×	✓	×	×	Poor performance in dense and occluded crowd scenes.
Conte et al. [106]	Supervised and Conventional ML	SCR	×	×	×	×	Handcrafted features are not adaptive to dense crowd scenes and thus the degrades in performance.
Chan et al. [59]	Supervised and Conventional ML	SCR	×	✓	×	×	Does not address scale variation due to perspective distortion. The holistic features are scene specific and are adaptive for crowd counting.
Tan et al. [107]	Semi-Supervised and Conventional ML	SCR	×	✓	×	×	Does not address scale variation due to perspective distortion. The holistic features are scene specific and are adaptive for crowd counting.
Fradi et al. [108]	Supervised and Conventional ML	SCR	✓	✓	×	×	holistic features are scene specific and are adaptive for crowd counting.
Jiang et al. [109]	Supervised and Conventional ML	SCR	✓	✓	×	×	holistic features are scene specific and are adaptive for crowd counting.
Zeyad et al. [110]	Supervised and Conventional ML	SCR	✓	✓	×	×	The threshold for level of occlusion decision in a frame is predefined and not dynamic or adaptive.
Khan et al. [21]	Supervised and Deep Learning	Detection	✓	×	×	×	Counting performance depends on the crowd head detection output which is less effective in very dense crowd scene.

Xiong et' al [12]	Supervised and Deep Learning	DMR	×	×	×	×	Does not handle scale variation due to perspective distortion and also minimize the effect of cluttered background.
Zhang et' al [111]	Supervised and Deep Learning	DMR	×	×	×	×	Does not handle scale variation due to perspective distortion and also minimize the effect of cluttered background.
Miao et' al [32]	Supervised and Deep Learning	DMR	×	×	×	×	Does not handle scale variation due to perspective distortion and also minimize the effect of cluttered background.
Gao et al. [112]	Supervised and Deep Learning	DMR	×	×	✓	✓	Does not handle scale variation due to perspective distortion and also minimize the effect of cluttered background.

Conte et al. [106] proposed a regression-based crowd counting method based on the SURF algorithm. The authors extracted salient points using the SURF algorithm and then clustered the moving points into distinct clusters. The features from each cluster are extracted and mapped onto the ground truth count using the SVR algorithm to count people in the video sequence. Chan *et al*. [59] proposed a privacy-preserving crowd counting approach in crowd video. At first, the authors proposed to use the mixture of dynamic textures motion model for crowd segment. Then holistic features such as segment and edge features are extracted and inputted into the gaussian process regression (GPR) for crowd counting.

Chan et al. [113] proposed Bayesian Process Regression (BPR) for crowd counting in video sequences. The authors used low-level features to achieve the objective function. Tan et al. [107] proposed a semi-supervised Elastic Net regression approach for crowd counting in crowd video datasets. The authors extracted the foreground image by applying a running average followed by feature modeling for crowd counting. The authors extracted six sets of features concerning foreground masks and foreground images.

Fradi *et al*. [108] applied Gaussian Mixture Model (GMM) for foreground map extraction. Authors extracted frame-wise normalized features and then predicted crowd count using GPR. The authors applied perspective normalization to minimize the effect of distortion. Jiang *et al*. [109] performed crowd counting by extracting a new set of low-level features from the crowd video and fused relevance vector regression (RVR) and GPR for crowd prediction. Zeyad *et al*. [110] proposed a crowd counting approach by adopting dynamic feature selection and occlusion handling. On the other hand, Xu *et al.* [8] extracted HOG features and GPR for crowd counting in videos. Although various conventional-based count-based video CCDE approaches have been proposed in the literature and have shown their effectiveness on low crowd density video datasets, the effectiveness of such models is degraded because of the use of handcrafted features which are not adaptive to crowd scenes.

2.1.3.2 Deep-Learning-based Techniques

The deep learning approaches for video-based CCDE are the current research trends. Very few approaches have been proposed for detection and density map-based regression approaches. In this section, we will summarise these techniques.

Khan *et al*. [21] proposed a detection-based approach for crowd counting in sports videos. Authors generate scale-aware head proposals, which are given to the CNN to obtain the presence of probability of people across the scale, but the counting performance depends on the detection output, which is less effective in very dense crowd scenes. Such a model does not take advantage of temporal features during crowd counting. Nevertheless, the SCR and DMR approaches take advantage of spatial and temporal features during crowd counting. Xiong *et al.* [12] proposed a bidirectional Conv-LSTM to perform spatial-temporal modeling of video frames followed by non-linear regression on ground-truth crowd density maps. Zhang *et al.* [111] proposed a hybrid model based

on a fully convolution neural network (FCN) and residual Lang short-term memory (FCN-rLSTM) for car counting as well as crowd counting. The authors designed an FCN to extract mid and very deep spatial features from video frames. An rLSTM is then designed to exploit temporal features between frames, followed by density map regression. Miao *et al.* [32] proposed a spatial-temporal CNN (ST-CNN) for video crowd counting. The 2D-CNN and 3D-CNN are used to extract spatial-temporal features from the video sequences, and the fused feature is then merged to perform density map regression for crowd counting. Recently, Gao *et al.* [112] proposed a domain adaptation-based crowd counting approach for video surveillance applications. The authors proposed multilevel feature-aware adaptation (MFA) and structured density map alignment (SDA) modules to enhance the crowd counting performance.

Although the video-based CCDE models [12], [32], [111] depend on the spatial-temporal features but do not consider measures to minimize background details' influence and neglect the attention of each feature type's response for crowd counting.

2.1.4 Summary of Vision-based CCDE

A brief overview of the state-of-the-art approaches for vision-based CCDE and their comparative analysis is illustrated in Table 2.1 and Table 2.2. The initial approaches vastly explored conventional machine learning techniques for image- and video-based CCDE. However, the current research trend shows a massive implementation of several deep learning approaches for vision-based CCDE. The DMR-based approaches provide better solutions for varying crowd densities among several modes of implementation approaches. The following conclusions can be made from the literature review,

- The detection-based approaches for image and video datasets perform poorly in dense crowd scenes.

- The single-count regression and DMR approaches overcome such challenging situations.
- The single-count regression approaches do not consider the spatial distribution of crowd during regression, but the DMR approach does.
- The hybrid approaches increase the model complexity.
- The image-based CCDE does not consider temporal dependencies between frames of the crowd video dataset.
- The video-based CCDE approach extracts spatial and temporal features from the crowd scene. The video-based CCDE using deep learning techniques performs better than the handcrafted feature learning techniques.
- However, the video-based CCDE using deep learning approaches is lacking in handling crowd shape change due to perspective distortion in the video and minimizing the effect of cluttered background from the video scenes.

2.2 Literature review on crowd congestion-level analysis

Crowd congestion-level analysis (CCA) is one of the prominent research fields in crowd analysis. The terms crowd congestion-level analysis and crowd density classification are interchangeably used in the literature, where each density level defines a particular degree of congestion. The existing works can be categorized into two types based on the learning method.

- Conventional Machine Learning-based approaches and
- Deep Learning-based approaches.

Table 2.3 illustrates a brief literature review of the state-of-the-art approaches for CCA using both conventional and deep learning approaches.

2.2.1 Conventional Machine Learning-based Approaches for CCA

The traditional methods for crowd congestion or density level classifications depend mainly on handcrafted features and conventional machine learning approaches. Such handcrafted features vary significantly between different congestion levels, thus attracting researchers to exploit significant handcrafted features for the CCA. Handcrafted features based on shape, texture, edge, moments, spectral (Fourier), and wavelets were utilized for CCA. Marana *et al*. [115] explained that the texture information of crowd scenes changes dramatically from sparse to very dense and extracted texture features using Gray Level Dependency Matrix (GLDM) from the crowd scene and classified different classes using Self Organisation Map (SOM). However, the performance of the model [115] is poor. Again Marana *et al*. [116] characterize the crowd densities using Minkowski Fractal Dimension (MFD) and classify different density levels using SOM. The authors achieved 75% correct classification.

Rahmalan *et al*. [117] exploited shape, texture, and moment features using MFD, GLDM, and Translational Invariant Orthonormal Chebyshev Moments (TIOCM) for different crowd density classes and classified them using SOM. However, the cluttered background, shadow, and noise degrade the model's performance. Marana *et al*. [118] proposed a real-time model by exploiting texture histogram and low-pass filter for classification correction in a distributed environment but obtained 73.89% accuracy only. Su *et al*. [119] extracted crowd regions by adopting Maximally Stable Externally Region (MSER) [120] and then applied projection on it, followed by shadow removal. The authors extracted histogram-based statistical features and trained the model using SVM, but the model resulted in moderate performance.

Ma *et al*. [121] modeled crowd frames using Gradient Orientation Co-occurrence Matrix (GOCM). The authors created a bag of visual words to normalize the GOCM

descriptors followed by k-mean clustering. The model can handle background noise and scene changes. Wang *et al*. [122] extracted texture features from the grayscale and gradient image using the Local Binary Pattern Co-Occurrence Matrix (LBPCM). The SVM is used to train the model and achieves 94.25% accuracy. However, the frame processing time is very slow and cannot be applicable in real-time applications. Kim *et al*. [123] extracted feature descriptors for both moving and static crowds. The authors employed Combined Local-Global (CLG) optical flow and accumulated magnitude map to find the moving area and then applied a threshold value to find a normalized moving area. Authors capture GLCM to capture contrast information. The idea is noble but suffers from a high misclassification rate in varying scenes and lighting changes. Yang *et al*. [124] extracted spatial-temporal features from the moving crowd scene. Such features are extracted using sparse-spatial-temporal local binary pattern (SST-LBP) descriptor followed by spectral analysis. Authors adopted SVM for multiclass classification.

Fradi *et al*. [40] proposed a patch-level crowd density classification technique by extracting LBP features from the crowd patches. The extracted features are followed by a discriminant subspace analysis using LDA and PCA. The final feature descriptors were inputted to SVM with RBF kernel for multiclass classification. Lamba *et al*. [125] obtained rotational invariant spatial-temporal LBP features for the moving crowd. At first, key interest points were detected using Hessian Detector [126] for a volume of frames. Then RIST-LBP features were extracted from the volume of spatial-temporal key points, and then SVM was employed to classify four crowd density levels. The authors did not consider descriptors for the static crowd in the scene.

Table 2.3: Comparative analysis of Vision-based CCA approaches

Ref.	Congestion Classification at		No. of Density Levels	Feature Type		Features	Learning Algorithm	Advantages	Limitations
	Local-Level	Global-Level		Spatial	Temporal				
[116]	×	√	Five	√	×	GLDM	SOM	Accurately recognizes very low-density frames.	Couldn't capture variations between different density levels therefore results are poor performance.
[117]	×	√	Five	√	×	GLDM, MFD, Spectral (Fourier)	SOM, Bayesian Classifier, Curve Fitting	First to implement MFD for this problem	Can't extract discriminant features, so results in poor performance.
[118]	√	×	Five	√	×	GLDM, Minkowski, Chebyshev Moments	SOM	Achieves better result using Chebyshev moments	Small samples for testing. Performance is affected by noise, shadow and cluttered background.
[119]	×	√	Five	√	×	Texture Histogram	SOM	Real time	Poor performance.
[120]	×	√	Five	√	×	Histogram statistical features	SVM	Better crowd descriptor.	Assumed that all people are of same size and they are located in same horizontal plane which is not applicable to real world crowd scenarios.
[121]	×	√	Five	√	×	Gradient Orientation Co-Occurrence Matrix	Bag of words using k-means clustering.	Handles scene change and background noise.	Poor performance.
[122]	√	×	Four	√	×	LBP-Co-occurrence matrix	SVM	Good crowd descriptor and better performance	Not suitable for real time application.
[123]	×	√	Five	√	√	Normalized Moving area and normalized contrast.	MLP with BP	Crowd density estimation is done for moving and stationary crowds.	The feature descriptor dimension is only two which is not suitable to distinguish large scale dataset.

[40]	√	×	Five	√	×	Subspace LPB features using LDA and PCA.	SVM with RBF kernel.	Reduce dimensional	Poor performance.
[124]	×	√		√	√	SST-LBP	SVM	Good descriptor for moving crowd.	Poor performance. Didn't consider static crowd information.
[114]	√	×	Four	√	×	CLBP	SVM	Good descriptor	Not applicable to real time application.
[115]	√	×	Five	√	×	Deep features using ConvNets	Deep Learning	Faster implementation	Poor performance.
[125]	√	×	Four	√	×	Deep features using GoogleNet and ResNet	Deep Learning	Created 160K density annotated scenes for 31 crowd sequences.	Poor performance.

In the other scenario, Alanazi et al. [114] proposed a local crowd density classification technique in which CLBP features were utilized to describe four different density or congestion levels. Multi-class SVM was used for classification. Although the method showed some promising results, it cannot be applicable in real-time as a single-frame processing time is 7 seconds.

2.2.2 Deep Learning-based Approaches for CCA

A very few models have been proposed using the deep-learning algorithm for CCA. Fu et al. [115] proposed three deep CNN models: modified multi-stage CNN (MS-CNN), optimized CNN, and cascaded CNN. The main focus was to extract deep spatial features for CCA and increase the processing speed by deleting weights of the neurons which have similar receptive fields.

However, the proposed model has some limitations like i) One has to manually identify complex samples from the dataset for cascade ConvNets which is quite impossible to find in a real-world application, and ii) Extracting only spatial features cannot increase the performance as the crowd scene contains both spatial (for the static crowd) and temporal (moving crowd) information. Pu et al. [125] utilized a transfer learning mechanism to solve the objective. Authors adopted GoogLeNet and VGGNet-16 to classify crowd densities for three and five classes; it still suffers to achieve better performance.

2.2.3 Summary of Vision-based CCA Approaches

A [114]A brief review of vision-based CCA has been presented. Methods and models using conventional and deep learning approaches have been reviewed and compared. The following conclusion can be drawn from the review,

- The handcrafted feature learning techniques perform poorly on varying crowd densities.
- The deep learning approaches provide fine-grained features for the CCA and are more effective than the conventional approaches.
- However, very little work has been developed for the CCA, which is lacking in utilizing scale-invariant features for CCA to minimize the effect of varying crowd shape change due to perspective distortion.

- The availability of datasets for the CCA is also limited in quantity.

2.3 Literature Review on Crowd Behavior Analysis

Crowd behavior analysis (CBA) or Crowd behavior prediction (CBP) is an essential task of CA, which provides crowd safety and security, thereby minimizing crowd disasters. Thus, the CBA has gained much attention from academicians and industrial AI. The state-of-the-art approaches can be broadly categorized into two types based on the availability of ground-truth

- One-Class Classification (OCC) [13], [126], [135]–[144], [127], [145], [128]–[134] models
- Multi-class classification (MCC) models [2], [146], [147].

The OCC-based crowd behavior prediction (CBP) approaches to learn the normal crowd behavior patterns, and during testing, the frames that deviate from the normal behavior are treated as crowd anomalies or abnormal behavior. Learning models like One-Class Extreme Learning Machine (OC-ELM) [148], Gaussian classifier [137], [149], OC-SVM [131], [138], [140], [150], reconstruction error [13], [139], [143], [145], thresholding-based approaches, reconstruction error-based regularity-score [145], similarity score between the target and balanced distribution [134] are adopted as classifier for anomaly or panic detection. The issue with OCC-based approaches is that they do not consider the dissimilarities between different types of abnormal crowd behaviors and treat them as one class. In contrast, the MCC-based CBP approaches consider the dissimilarities between different crowd behaviors by solving them as a multi-class problem. The MCC-based CBP is beneficial in identifying different types of crowd behaviors and helps control crowd disasters. However, limited models have been proposed for MCC-based CBP [2], [146], [147]. The following subsections briefly review

the OCC-based crowd behavior prediction models, where the main focus will be on crowd panic detection, and the MCC-based CBP approaches.

2.3.1 OCC-based Crowd Behavior Prediction

The OCC-based CBP is also known for crowd anomaly or outlier detection. A brief literature review is conducted on the research papers based on abnormal crowd event detection, including crowd panic behaviors. Most of these approaches follow a non-object-centric approach in which the model tries to learn the features to model normal behavior patterns, and the outliers are treated as abnormal or panic behavior. Traditional [126]–[129] and deep-learning approaches [13], [130], [139]–[145], [131]–[138] have been proposed to achieve the objective function.

2.3.1.1 Traditional Approaches for OCC-based CBP

Among traditional approaches, Lu *et al.* [126] proposed a fast crowd anomaly detection approach that utilizes the inherent redundancy between the video frames and achieves a frame rate of around 150 fps. Cheng *et al.* [127] proposed a hierarchical feature representation framework for local and global anomaly detection. Saligrama *et al*. [128] proposed a statistical approach to exploit spatial-temporal features for crowd anomaly detection. Direkoglu *et al.* [151] proposed a novel feature descriptor using motion and orientation attributes of crowd scenes. Authors adopted OC-SVM to predict normal and panic crowd behaviors. Halbe *et al.* [152] proposed a motion attribute-based energy estimation features descriptor and a thresholding approach for crowd abnormal or panic detection. Hoa *et al.* [153] exploited spatial-temporal features from the spatial-temporal volume and detected crowd panic behavior using a thresholding approach. However, the thresholding process is not adaptive and is biased, which is not suitable for real-time applications. Kumar *et al*. [154] proposed a threshold-based clustering approach to detect crowd panic or abnormal activities. Authors utilize motion attributes for anomaly

detection. Mousavi *et al.* [155] extracted histogram of oriented tracklet (HOT) features from the crowd video and fed them to Latent Dirichlet allocation and Support Vector Machines for anomaly detection. Wang *et al*. [156] modeled the crowd video by extracting a histogram of optical flow orientation (HOFO) and inputted these features to OC-SVM for crowd anomaly detection. Wu *et al.* [157] proposed a Bayesian model for crowd escape like panic behavior detection using location and motion (magnitude and direction) features. Zhang *et al*. [158] proposed an enthalpy-based crowd motion modeling approach for crowd panic detection. The authors observed that the crowd panic scenes have higher enthalpy than the normal crowd scenes.

Similarly, Zhang *et al*. [159] proposed a thresholding approach-based energy-based feature descriptors for crowd panic detection. Such thresholding approaches are not adaptive, and, most of the time, they are biased toward achieving better results. Recently Lamba *et al.* [129] proposed a trajectory-based approach for crowd anomaly detection. The traditional approaches solely depend on handcrafted features to represent crowd scenes. Recently, Aldissi *et al.* [160] proposed a frequency domain-based crowd panic detection model which can process the frames in real-time. A clustered-based detection approach was proposed to detect crowd anomalies. Shehab *et al*. [161] proposed a statistical feature learning approach to differentiate between crowd panic and non-panic events. However, the recent state-of-the-art approaches fail to address scale variation due to perspective distortion in the video. Moreover, the accuracy of the model needs to be improved.

In short, conventional approaches cannot extract fine-grained features, resulting in lower performance. So, to overcome such issues, deep learning approaches have been vastly explored.

2.3.1.2 Deep-Learning Approaches for OCC-based CBP

Deep learning models such as CNN [130]–[135], recurrent neural networks [136], generative models (Autoencoders or encoder-decoder models [137]–[142], Generative Adversarial Networks [13], [143]), and hybrid models [144], [145] have been vastly explored for OCC-based CBP. Most of these approaches detect anomaly/panic at frame or pixel levels.

Both frame-level and pixel-level anomaly detection approaches are proposed by Zhou *et al.* [130] by designing a spatial-temporal CNN. The Spatial-Temporal CNN extracts feature from spatial and temporal dimensions from the volume of patches. Bouindour *et al.* [131] utilized the first two convolution layers of a pre-trained Alex-Net [162] to obtain spatial-temporal features from the volume of patches/frames. The features are inputted into an OC-SVM to train the normal frames, and during testing, the outliers are treated as abnormal events. Ravanbakhsh *et al.* [133] proposed to use of a plug-and-play CNN model for crowd anomaly detection. Authors [133] extracted both motion and semantic features from the crowd video to detect local anomalies. The advantage of such a model is it does not need any finetuning. Bouindour *et al*. [134] exploited spatial-temporal features from the volume of frames using a modified pre-trained residual 3D-CNN. Song *et al*. [135] proposed a 3D-CNN for video anomaly detection.

Besides CNN-based deep models for crowd anomaly detection, a recurrent neural network such as bidirectional-LSTM (BDLSTM) has been designed for crowd anomaly detection. Features like the histogram of oriented gradients (HOG) were extracted from the frames and given into the BDLSTM model for real-time detection of crowd violence like anomalous events in the football stadium.

Generative models like variants of autoencoders or encoder-decoder models and GANs have been exploited for crowd anomaly/panic detection. Sabokrou *et al*. [137]

extracted a set of local and global descriptors for crowd anomaly detection and localization in real-time. The global descriptors are obtained using a sparse autoencoder. Two Gaussian classifiers are used for anomaly detection and localization. Xu *et al*. [138] proposed stacked denoised autoencoders to exploit motion and appearance features from normal crowd scenes and adopted both early and late fusion followed by OC-SVM for anomaly detection. George *et al.* [139] exploit the HOFM features from the normal crowd videos' parallelepipeds of non-uniform spatial-temporal regions. The extracted feature vectors are then fed into an autoencoder model to detect abnormal/panic events. Tran *et al.* [140] proposed a convolutional autoencoder model to extract motion-related features from normal crowd videos. The authors utilized OCSVM to detect crowd anomalies. The encoder-decoder model using CNN [141] is also explored in the literature. Chong *et al*. [141] proposed a spatial-temporal autoencoder using CNN to train the normal crowd sequences. The abnormal events are detected based on the reconstruction error. Sabokrou *et al.* [142] proposed a deep cascade of 3D CNN-based autoencoders for crowd anomaly detection.

Ravanbakhsh *et al*. [13] proposed a GAN to detect crowd anomalies at frame and pixel levels. Authors exploit inherent motion patterns using GAN for normal crowd frames, and the abnormal events are detected based on the reconstruction error. Again, Ravanbakhsh *et al.* [143] proposed a GAN to detect cross-channel frame-level and pixel-level crowd anomaly detection. The model is trained using normal events, and the abnormal events are detected based on the reconstruction error.

Many hybrid models have also been explored for anomaly/panic detection. For example, Zhuang *et al.* [144] proposed a deep end-to-end network with CNN (Inception-V3) and stacked differential LSTM for understanding crow scene-like violent-based crowd abnormal events. Yang *et al.* [145] proposed a CNN-based autoencoder LSTM

model for crowd anomalous event detection. Ammar *et al*. [11] proposed a real-time detection framework for crowd panic and normal behavior detection. The authors proposed a hybrid model in which they extracted handcrafted features from the crowd video followed by an LSTM model to capture the temporal dependencies between frames. The model is trained with normal frames. The frames will be treated as panic based on the prediction error.

Overall, recent state-of-the-art approaches fail to address scale variation due to perspective distortion in the frame and across the frames in the crowd video dataset. Also, the accuracy of the model needs to be improved.

2.3.2 MCC-based Crowd Behavior Prediction

A few models have been proposed as far as MCC-based CBP is concerned. The main reason is the lack of availability of more multi-class ground-truth datasets. Nevertheless, recently multi-class crowd behavior datasets like MED, GTA, and Crowd-11 have been proposed. Conventional feature learning approaches [2] used HOG, HOF, MBH, dense trajectories, and HOT features to classify crowd behaviors. Dupont et al. [147] analyzed the performance of different deep models for CBP. Models [147] using 3D-CNN and 3D-CNN+VGG were developed for crowd behavior prediction. Lazaridis *et al.* [146] proposed a two-stream deep learning architecture for the CBP. The heat map and the optical flow of the crowd scene are inputted into the first and the second stream, respectively. The two streams were developed using convolution layers and Conv-LSTM blocks.

The state-of-the-art MCC-based CBP approaches only focuses on extracting appearance and motion attributes using deep learning techniques, but the following challenges are yet to be addressed,

a) Human shape changes due to perspective distortion and

b) Minimizing the effect of cluttered background.

2.3.3 Summary of CBA Approaches

In In the above sections, a brief literature survey has been conducted for OCC and MCC-based CBP. The review of OCC-based CBP emphasizes crowd anomaly or abnormal detection methods for crowd panic situations. Because the state-of-the-art crowd anomaly detection dataset is very limited where more dataset focuses on crowd panic situations. Other reasons for focusing [2]on crowd panic-based anomaly detection methods are,

- The existing crowd anomaly detectors do not consider the difference between different anomaly behaviors [147] and consider anything that deviates from normal [147] crowd activities to be anomalies; such a process is ambiguous. It is difficult to say that crowd behavior patterns for panic, congestion[146], and fighting are the same as anomalous events.
- Most of the crowd anomaly dataset contains crowd panic situations. Therefore, developing a crowd anomaly model with panic and normal crowd behavior seems more feasible than combining all the anomaly events into one category. The following conclusions can be drawn from the above review,
- Both conventional machine learning and deep learning approaches have been developed for OCC-based CBP.
- The deep learning-based feature learning approaches perform better than the traditional approaches.
- Different model exploration and better performance are always in deep need.
- Limited works have been done for MCC-based CBP.
- The deep learning approaches show better performance than the conventional approaches.

- Crowd shape variation and minimizing the effect of the cluttered background still need to be handled to achieve better performance.

2.4 Literature review on Multitasking Crowd Analysis

Multitasking crowd analysis focuses on designing and developing a unified model that can perform more than one crowd-related task like crowd counting, crowd behavior prediction, crowd tracking, crowd segmentation, and so on; very few works have been proposed due to the lack of multitasking crowd analysis dataset. This section briefly reviews the methods and models adopted for multitasking crowd analysis. Kang *et al.* [14] implemented CNN-based models to generate crowd density maps which were used to perform several crowd analysis tasks like crowd counting, detection, and tracking. Authors found that the low-resolution crowd density maps usually provide better counting performance, but this methodology has scale and background elimination-related issues. Marsden *et al*. [15] proposed a multitasking crowd analysis dataset containing only 100 images. Authors utilized ResNet to perform multitasking crowd analysis, including crowd counting, violent crowd behavior detection, and crowd density-level classification. The limitation of such a model is that it suffers from overfitting due to a lack of sufficient training and a testing dataset.

2.4.1 Summary of Multitasking Crowd Analysis

Thus, based on the review of the multitasking crowd analysis models, the following conclusion can be drawn,

- There is a need to develop a largescale multitasking crowd analysis dataset.
- An efficient multitasking single model is highly required to minimize the computational cost.

- The model should handle the issues like scale change due to perspective distortion and minimize the effect of cluttered background.
- The available multitasking work is limited to crowd counting, detection, and tracking. However, crowd counting and crowd behavior prediction are the most prominent task of crowd analysis.
- There is an excellent scope for developing a largescale multitasking crowd analysis dataset and developing a more effective multitasking model..

2.5 Research Gaps Identified for Vision-based Crowd Analysis

In the above section, a brief review of several methods and models of four important tasks of CA: CCDE, CCA, CBA, and multitasking CA was presented. The study identifies the following research gaps in the different tasks of CA.

2.5.1 Research Gaps Identified for Video-based CCDE

The image-based CCDE does not consider the temporal dependencies between frames of the crowd video dataset. So, the proposed models for the task of CCDE are video-based crowd counting. The state-of-the-art video-based CCDE approaches have the following noticeable shortfalls.

- The video-based CCDE approaches [12], [32], [111] rely on spatial and temporal features and ignore features corresponding to the foreground of the crowd scene by eliminating background details. In such a case designing a multi cues feature extraction can be a good choice for crowd counting, which is not addressed by the state-of-the-art video-based approaches.
- The models [12], [111] ignore each feature set's (i.e., spatial and temporal) response towards crowd counting.

- Although ST-CNN [32] utilizes different fusion schemes to obtain final density maps from spatial and temporal streams, it requires lots of pre-training to train individual streams.
- The DMR-based approaches are prone to error as it depends on crowd head points obtained manually but considers spatial distribution of crowd scenes. On the other hand, Weakly-Supervised CCDE methodology, overcome from point-level annotation error by performing regression on crowd counts, considers global crowd properties but neglects local crowd distribution. However, both local and global crowd properties are needed.
- The video-based CCDE lacks handling human shape variation and minimizing the effects of cluttered background in the crowd video dataset. Figure 2.8 shows samples of frames with variations of human head shape due to perspective distortion in the frame and across the frames of the video dataset.
- The most commonly used performance metrics for crowd counting are mean absolute error (MAE) and root mean squared error (RMSE). For a robust and reliable CCDE model, the value of MAE and RMSE and their difference should be minimal. The video-based CCDE approaches [12], [163] (except ST-CNN [32]) attain a high deviation between MAE and RMSE, which is not wise for a crowd counting system. However, the deviation between MAE and RMSE in the ST-CNN [32] is low, but the performance needs improvement.

2.5.2 Research Gaps Identified for Vision-based CCA

By observing the literature on CCA, we may have the following research gaps for the vision-based CCA,

- Most traditional methods either extract spatial features or spatial-temporal texture features from the crowd scene but lack achieving better accuracy in real-time.

- Accuracy cannot be increased using only spatial features; crowd motion information must be used.
- The existing deep approaches only extract deep spatial features to solve the objective in real-time but fail to provide better accuracy.
- The existing deep-CNN models fail to provide better accuracy in the presence of perspective change and cluttered background.
- There is scope to develop a deep model that will extract and fuse deep spatial and temporal features for crowd density classification.

2.5.3 Research Gaps Identified for Vision-CBP

The study mainly focuses on two different works under vision-based CBP

- Crowd Panic detection using OCC-based Approach.
- MCC-based CBP.

So, the research gaps for each of the types are identified as below.

2.5.3.1 Research Gaps Identified for OCC-based Crowd Panic Detection (CPD)

Although the state-of-the-art [11], [148], [160], [161], [164], [165] crowd panic detection models adopt different ways to utilize motion attributes in the crowd scene but possess the following shortcomings.

- State-of-the-arts [11], [148], [160], [161], [164], [165] do not address human shape variations due to perspective distortion in the crowd scene. Figure 2.8 shows an example of a scale issue due to perspective distortion in a frame and across frames of the Pets-2009 crowd panic dataset.
- The state-of-the-art [29–34] rely on temporal or motion attributes for crowd panic detection but neglects the spatial features of the crowd scenes.

(a) Shape Change in the frame due to perspective distortion

(b) Shape Change across the frame due to perspective distortion

Figure 2.8: Example of crowd shape change due to perspective distortion in the Pets-2009 crowd panic dataset

(a) Human shape change in a crowd scene of MED dataset

(b) Human shape change in a crowd scene of GTA dataset

Figure 2.9: Examples of human shape change due to perspective distortion in crowd scenes

2.5.3.2 Research Gaps Identified for MCC-based CBP

The conventional machine learning [2] and deep learning [146], [147] models have been proposed for MCC-based CBP. Nevertheless, the deep learning approaches [146], [147] provide a better result than the conventional approach [2] but are limited in addressing the following challenges that exist in the crowd scenes,

- Change in pedestrian's shape due to perspective distortion. Figure 2.9 shows human shape change in the crowd scene where the person near the camera occupies more pixels than the person far from the camera.
- Minimizing the effect of cluttered background.

2.5.4 Research Gaps Identified for Multitasking CA

Minimal works have been proposed for multitasking CA, and the followings are the research gaps that we have obtained during the literature survey,

- The state-of-the-art only focuses on crowd counting and tracking and neglects the critical task for CA, i.e., CBP.
- There is a lack of largescale multitasking CA datasets covering important tasks of CA, i.e., CCDE and CBP.
- The existing models are lacking in handling crowd shape change due to perspective distortion and minimization of cluttered backgrounds.

In the above, we have discussed different research gaps in individual tasks of crowd analysis. The vision-based CA needs effective feature modeling and learning strategies to improve the model's performance. Challenges like crowd shape variation and cluttered background effects need to be handled as far as video-based CA is concerned. Also, we have found a lack of availability of a multitasking CA dataset. These research gaps are the motivations to design effective models to address them effectively.

2.6 Datasets used for Experimental Analysis

In this thesis, various datasets have been used for different tasks of crowd analysis. The following subsections describe this information in detail.

2.6.1 Datasets used for Video-based CCDE

There is a lack of variety of video-based crowd counting datasets. For experimental analysis of different models of CCDE, three publicly available crowd counting datasets are used, namely,

- The Mall dataset [57]
- The UCSD dataset [59]

- The Venice dataset [4]

Table 2.4 shows the detailed properties of these datasets. The Mall [57] and UCSD [59] datasets contain moderate crowd densities ranging from 11 to 53, whereas the Venice dataset contains high crowd densities ranging from 86 to 421 people. The Mall [57] dataset consists of 2000 frame sequences captured from a mall. The modality of these frames is RGB, and the resolution is $[480 \times 640 \times 3]$. The same setup is adopted for training and testing, as mentioned in [57]. The first 800 frames are used for training, and the rest 1200 for testing. The Mall dataset [57] includes a challenging situation like a mirror reflection.

Figure 2.10: Example of a Frame of Mall Dataset [57]

Figure 2.11: Example of a Frame of UCSD Dataset [59]

Figure 2.12: Example of a frame of Venice dataset [4]

The UCSD dataset [59] captures pedestrians walking from campus using a still camera. The benchmark dataset contains Grayscale frames with varying crowd densities, which range from 11 to 46. The resolution of the frames is [158×238]. The dataset consists of 2000 frames from which sequences from 601 to 1400 are used for training, and the rest of 1200 frames are used for testing. This work follows the same procedure described in the UCSD [59] dataset and used ROI of UCSD for experiment. The Venice dataset [4] has 167 RGB frames from four video sequences captured in Piazza San Marco, Venice. All the frames are size [1280×720×3]. The dataset has challenging situations like camera jitter and camera motions. The crowd densities range from 86 to 421. The experiments

follow the same training and testing setup as mentioned in [4]. The ROI of the Venice dataset is considered.

Table 2.4: Details of datasets used for the CCDE

Dataset	Challenges	Resolution	Total Sequences	Training Samples	Testing Samples	Minimum Crowd Count	Maximum Crowd Count
Mall [57]	Mirror reflection	$480 \times 640 \times 3$	1200	800	1200	11	53
Venice [4]	Camera jitter, Camera motion	$1280 \times 720 \times 3$	687	80	87	86	421
UCSD [59]	Low-resolution crowd frames.	158×238	2000	800	1200	11	46

2.6.1.1 Generating Ground Truth Crowd Density Maps

The famous geometric adaptive kernel (GAK) [27] is adopted to generate ground truth density maps for different models. Let a crowded frame s_i have $a\ P$ number of annotated head points. So, the frame, s_i with P number of head points can be represented as a function $\mathrm{H}(s_i)$,

$$\mathrm{H}(s_i) = \sum_{j=1}^{P} \delta\left(s_i - s_{i_j}\right) \tag{2.1}$$

As per procedures mentioned in [27], we have to calculate k nearest distances for every head point to its k-nearest head points. Let us have m different distances $\left(\left(\{d_1^j, d_2^j, \ldots\ldots d_m^j\}\right)\right)$ for a given head pixel j of a frame s_i. Next, the average of these distances is obtained by using the following Equation 2.2,

$$d^j = \frac{1}{m}\sum_{k=1}^{m} d_k^j \tag{2.2}$$

Finally, using Equation-3, the density map (DM) for a frame s_i, is generated by convolving the $\mathrm{H}(I)$ with a gaussian kernel with variance δ_j, i.e., $G_{\delta_j}(s_i)$.

$$DM(s_i) = \sum_{j=1}^{P} \delta\left(s_i - s_{i_j}\right) * G_{\delta_j}(s_i) \tag{2.3}$$

$, where\ \delta_j = \beta d^j$, β=0.3 and K=4. Figure 2.13 to Figure 2.16 shows few samples of datasets and their ground truth crowd density maps. The resolution of the output of the proposed models are $[25 \times 25]$, so the ground-truth crowd density maps are rescaled into size $[25 \times 25]$. The rescaling preserves the respective head position of the crowd scene. Let the set $GT = \{gt_1, gt_2, \ldots\ldots, gt_N\}$ represents all the rescaled ground-truth crowd density maps. Each gt_i represents the ground-truth crowd density map of the i^{th} frame, and we have N number of such frames.

The resolution of the output of the proposed models are $[25 \times 25]$, so the ground-truth crowd density maps are rescaled into size $[25 \times 25]$. The rescaling preserves the respective head position of the crowd scene. Let the set $GT = \{gt_1, gt_2, \ldots\ldots, gt_N\}$ represents all the rescaled ground-truth crowd density maps. Each gt_i represents the ground-truth crowd density map of the i^{th} frame, and we have N number of such frames.

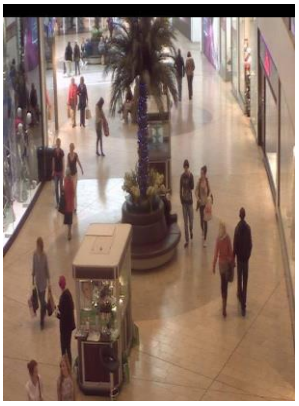


Figure 2.13: A frame of the mall dataset [57]

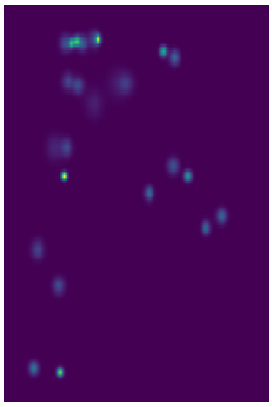


Figure 2.14: density map

Figure 2.15:A frame of the Venice dataset [4]

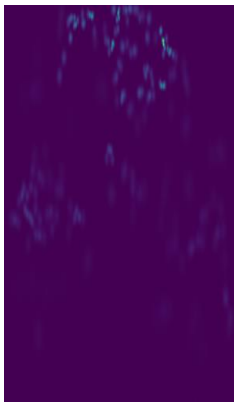


Figure 2.16: Density map

2.6.2 Datasets used for Crowd Panic Detection

Three benchmark crowd panic datasets are used: the UMN, the MED, and the Pets-2009.

Table 2.5: shows the properties of these three datasets

Dataset Name	**Context**	**Environment**	**Modality**	**No. of Samples**	**Density Level**	**Resolution**
UMN[166]	Real-world scenario with artificial escape like panic situation.	Lawn Indoor Plaza	Videos	11	Medium	[320 × 240]
MED[2]	Real-world scenario with artificial escape like panic situation.	Walkways	Videos	11	Sparse to High	[854 × 480]
Pets-2009 [167]	Real-world scenario with artificial escape like panic situation.	Campus view	Frames	855	Low	[768 × 576]

The UMN [166] dataset contains videos having real-world scenarios of normal and crowd escape-like panic behavior. The dataset is captured from three different environments: lawn, indoor, and plaza. It contains eleven crowd videos with medium crowd density. The resolution of these sequences is [320×240]. The Motion Emotion Dataset (MED) [2] is a benchmark crowd behavior dataset. The dataset contains 31 video clips that are captured from the walkways. We have adopted the steps followed by the state-of-the-art approach [11][160][161] where these approaches have used the first 11 video sequences of the MED dataset containing both "Normal" and "Panic" situations for crowd panic detection.

The crowd density varies from sparse to high. The resolution of such a dataset is [854×480]. The details of the MED dataset [2] are given in Table 2.3. The Pets-2009 [167] dataset is captured from a campus view containing real-world crowd normal and panic events. By following the process as described by the approaches [11][160][161], 855 frames (Containing Normal and Panic behavior scenes i.e., S1, S2, S3) having a resolution of [768×576] are selected for experiment. Figure 2.17 shows some examples of samples of three datasets.

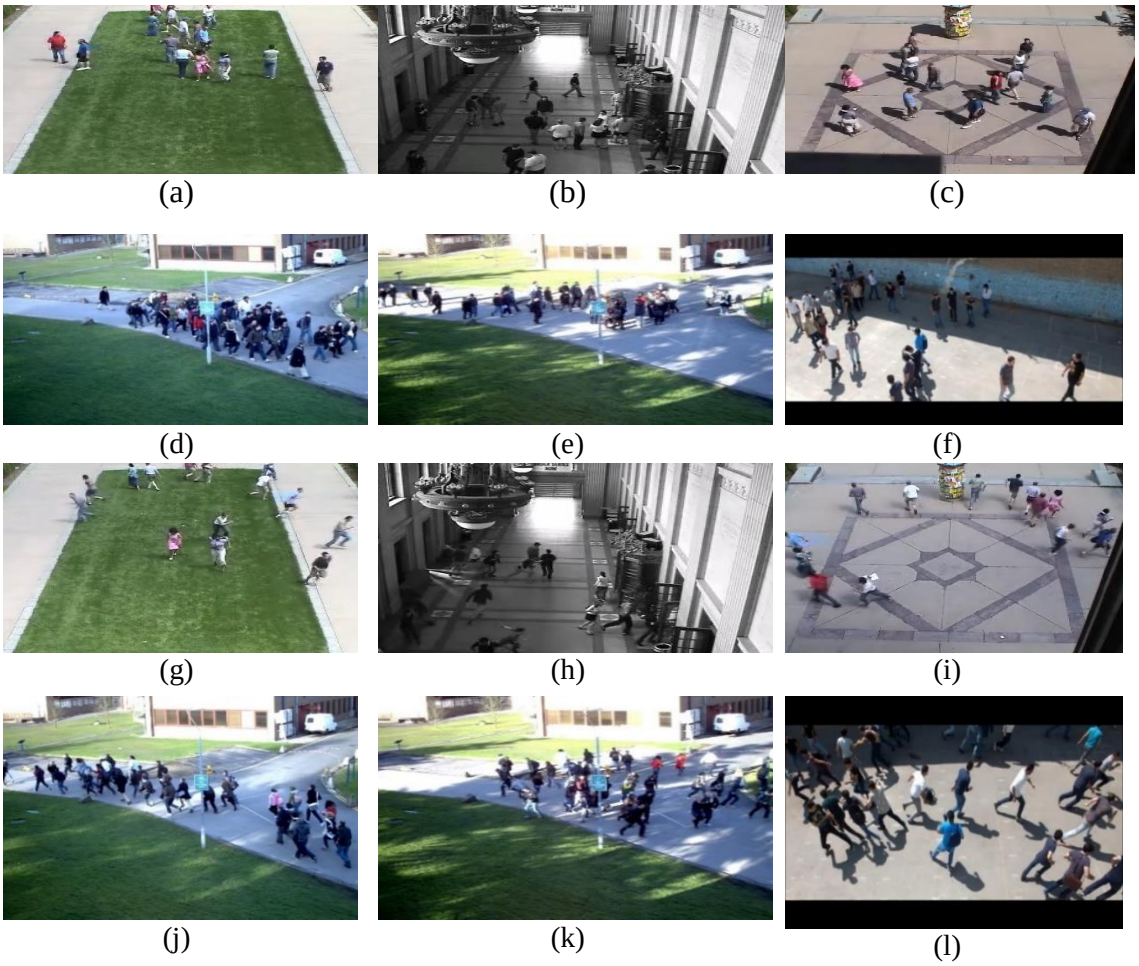


Figure 2.17: Examples of samples of the datasets. Figures (a), (b) and (c) are the examples of normal scenes of UMN S1, S2 and S3 respectively. Figures (d) and (e) are the normal scenes of Pets-2009 dataset. Figure (f) is the example of normal scene of MED dataset

2.6.3 Datasets used for Crowd Behavior Prediction

In the experiment, two largescale crowd behavior datasets are used, the MED dataset [2] and the grand theft auto v2 (GTA) [146] dataset. The MED dataset contains 31 crowd behavior sequences. There are five crowd behaviors: Neutral, Panic, Congestion, Fight, and Obstacle (or Abnormal). There are nearly 45,000 frames contained in the MED dataset. The resolution of the original frames is of size [480×854×3]. Figure 2.18 shows a few samples of the MED dataset [2]. Authors [2] adopted leave-one-out validation on the MED dataset. The GTA dataset [146] contains 14 Crowd behavior sequences. Each video contains more than 3 min video length. The GTA dataset contains only three crowd behaviors: Normal, Panic, and Fight. The frames are recorded at 60

frames per second. The resolution of the frames is of size [1080×1920×3]. Authors [146] randomly selected ten video sequences for training and four for testing on the GTA dataset. Figure 2.19 shows a few samples of the GTA dataset [146].

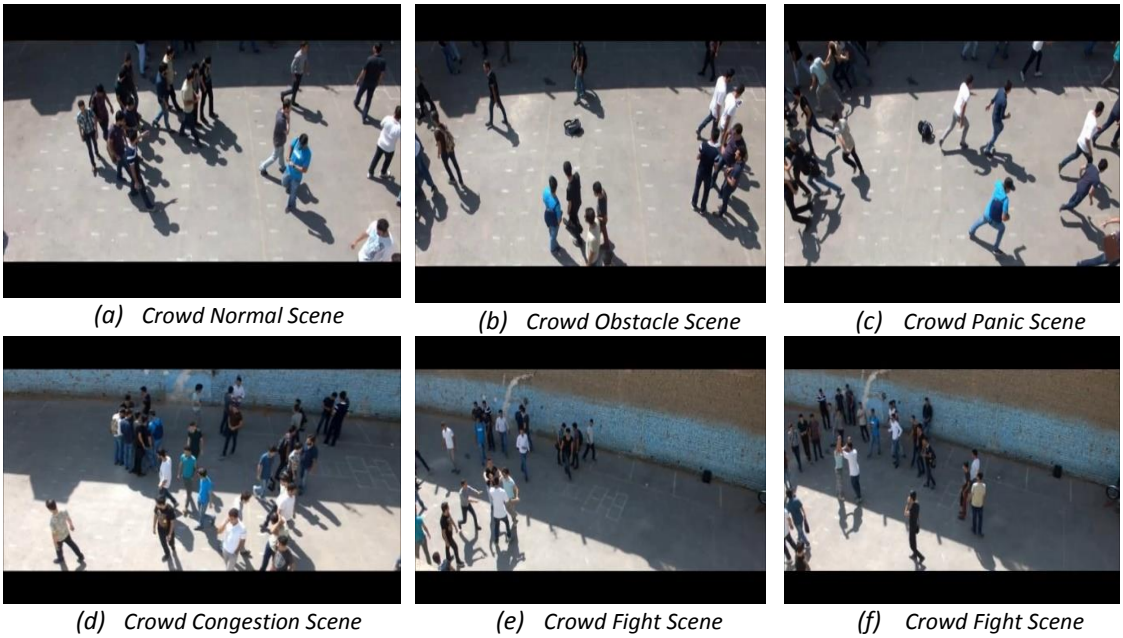


Figure 2.18: Examples of different samples of the MED dataset [2]

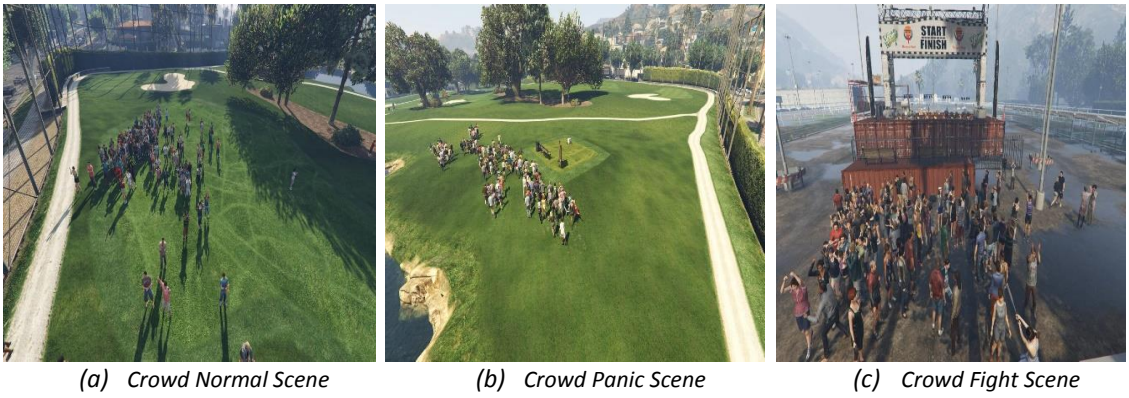


Figure 2.19: Examples of different samples of the GTA dataset [146]

2.7 Performance Metrics

The study focuses on regression and classification (binary and multi-class) problem modeling for different tasks of crowd analysis. For example, the vision-based CCDE is a regression-based problem. The OCC-based CBP adopts a binary classification

strategy as far as testing is concerned, whereas the vision-based CCA and MCC-based CBP are multi-class classification problems.

For video-based CCDE, the evaluation is conducted on standard benchmark performance metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The MAE and RMSE define the accuracy and robustness of the model. The value of MAE and RMSE should be as minimum as possible, and the difference between them should be minimum. The formula for MAE and RMSE can be defined using Equation 2.4 and Equation 2.5, respectively.

$$\mathrm{MAE} = \frac{1}{\mathrm{T}} \times \sum_{\mathrm{l}=1}^{\mathrm{T}} |\mathrm{GT}_l - \mathrm{P}_l| \tag{2.4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{T}} \times \sum_{\mathrm{l}=1}^{\mathrm{T}} |\mathrm{GT}_l - \mathrm{P}_l|^2} \tag{2.5}$$

, here T is the total number of frames, l is the frame index, *GT* is the ground-truth, and *P* predicted density maps.

The following confusion matrix (Table 2.6) and performance measures are used to evaluate OCC-based CBP approaches.

Table 2.6: Confusion matrix for binary classification

Actual Vs Predicted		**Predicted Classes**		**Total Number of Instances**
		Positive	**Negative**	
Actual Class	**Positive**	TP	FN	P=TP+FN
	Negative	FP	TN	N=FP+TN

In the case of OCC-based CBP, the positive and negative classes represent Normal and Panic behaviors. In the confusion matrix in Table 2. 6, TP, FP, FN, and TN represent True Positive, False Positive, False Negative, and True Negative, respectively. The following performance measures are used for evaluation purposes.

$$Precision = \frac{TP}{TP+FP} \tag{2.6}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.7}$$

$$F1_{Score} = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{2.8}$$

$$Error\ Rate(ER) = \frac{FP+FN}{TP+FN+FP+TN} \tag{2.9}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{2.10}$$

Now, for multi-class classification problems like vision-based CCL and MCC-based CBP, the following metrics are drawn by treating the confusion matrix (Table 2.6) as one versus all binary classification problems.

$$Recall\ (\mathrm{R_i}) = \frac{\mathrm{TP_i}}{\mathrm{TP_i+FN_i}} \tag{2.11}$$

$$\text{Precision } (\mathrm{P_i}) = \frac{\mathrm{TP_i}}{\mathrm{TP_i+FP_i}} \tag{2.12}$$

$$\text{False Positive Rate } (\mathrm{FPR_i}) = \frac{\mathrm{FP_i}}{\mathrm{FP_i+TN_i}} \tag{2.13}$$

$$\text{Specificity } (\mathrm{S_i}) = 1 - \mathrm{FPR_i} \tag{2.14}$$

$$\text{F1-Score } (\mathrm{F1_i}) = \frac{2\times \mathrm{P_i}\times \mathrm{R_i}}{\mathrm{P_i}+\mathrm{R_i}} \tag{2.15}$$

Here the index i represents to i^{th} class. Now, the global performance metric for the multi-class classification can be found by taking macro average of performance metric of individual class $i, whose\ value\ ranges\ from\ 1\ to\ K$. K is the total number of class labels. The following equations shows macro-average performance metrics.

$$\text{Recall } (\mathrm{R}) = \frac{\sum_{i=1}^{K} \mathrm{R_i}}{\mathrm{K}} \tag{2.16}$$

$$\text{Precision } (\mathrm{P}) = \frac{\sum_{i=1}^{K} \mathrm{P_i}}{\mathrm{K}} \tag{2.17}$$

$$\text{False Positive Rate } (\mathrm{FPR}) = \frac{\sum_{i}^{K} \mathrm{FPR_i}}{\mathrm{K}} \tag{2.18}$$

$$\text{Specificity } (\mathrm{S}) = 1 - \mathrm{FPR} \tag{2.19}$$

$$\mathrm{F1-Score} = \frac{\sum_{i=1}^{K} \mathrm{F1_i}}{\mathrm{K}} \tag{2.20}$$

$$Mean - Acc = \frac{1}{K}\sum_{i=1}^{K} Acc_i \tag{2.21}$$

, here K is the number of classes and Acc_i is the accuracy of the i^{th} class which is calculated as, $Acc_i = \frac{TP_i}{TP_i+FN_i}$.

2.8 Conclusion

In this chapter brief literature survey of various methods and models for different tasks of CA was conducted. The main focus was on the significant tasks of CA like CCDE, CCA, CBP, and multitasking CA. Both conventional machine learning and deep learning approaches have been reviewed. Several shortcomings of these approaches were identified and summarised as the research gaps for the study. After that review of various datasets and performance metrics used in the study for different tasks of CA was presented. The forthcoming chapters discuss the main contributions of this thesis.