Video analytics using deep learning for crowd analysis: a review

Md Roman Bhuiyan · Junaidi Abdullah · Noramiza Hashim · Fahmid Al Farid

Received: 18 January 2021 / Revised: 2 February 2022 / Accepted: 9 March 2022
Published online: 29 March 2022
© The Author(s) 2022

Abstract
Gathering a large number of people in a shared physical area is very common in urban culture. Although there are limitless examples of mega crowds, the Islamic religious ritual, the Hajj, is considered as one of the greatest crowd scenarios in the world. The Hajj is carried out once in a year with a congregation of millions of people when the Muslims visit the holy city of Makkah at a given time and date. Such a big crowd is always prone to public safety issues, and therefore requires proper measures to ensure safe and comfortable arrangement. Through the advances in computer vision based scene understanding, automatic analysis of crowd scenes is gaining popularity. However, existing crowd analysis algorithms might not be able to correctly interpret the video content in the context of the Hajj. This is because the Hajj is a unique congregation of millions of people crowded in a small area, which can overwhelm the use of existing video and computer vision based sophisticated algorithms. Through our studies on crowd analysis, crowd counting, density estimation, and the Hajj crowd behavior, we faced the need of a review work to get a research direction for abnormal behavior analysis of Hajj pilgrims. Therefore, this review aims to summarize the research works relevant to the broader field of video analytics using deep learning with a special focus on the visual surveillance in the Hajj. The review identifies the challenges and leading-edge techniques of visual surveillance in general, which may gracefully be adaptable to the applications of Hajj and Umrah. The paper presents detailed reviews on existing techniques and approaches employed for crowd analysis from crowd videos, specifically the techniques that use deep
learning in detecting abnormal behavior. These observations give us the impetus to undertake a painstaking yet exhilarating journey on crowd analysis, classification and detection of any abnormal movement of the Hajj pilgrims. Furthermore, because the Hajj pilgrimage is the most crowded domain for video-related extensive research activities, this study motivates us to critically analyze the crowd on a large scale.

Keywords Deep learning · Crowd analysis · Abnormal behavior · Video surveillances

1 Introduction

1.1 Background

For the last 14-centuries, the Hajj has been a sacred and religious ritual for Muslims worldwide. That is precisely the context in which all Muslims in the world believe in visiting Mecca and Kaaba at a given date and time. Due to the ever-increasing number of pilgrims, the management of large crowds has become a major issue. Several research works have revealed the massive devastating effect on pilgrims due to lack of crowd management by the authorities. In the last three years, the number of casualties has increased by at least 1426 [1, 2]. The use of efficient crowd analysis can potentially help the stakeholders to reduce a large number of casualties at the scene. Traditional approaches to crowd analysis that are merely based on CNN are inefficient. This is mainly because they cannot address the complex requirement associated with the highly dense crowd.

Based on the above consideration, the use of a modified CNN for crowd analysis and monitoring techniques in video surveillance has become important. Apart from the modified CNN, other approaches in crowd analysis have enabled the use of enhanced crowd analysis systems [43]. This is to ensure a significant reduction in the number of unexpected incidents during pilgrimages.

1.2 Motivation

Over the years, crowd analysis has shown steady improvement due to the emergence of novel approaches. Deep learning techniques have been increasingly used for many applications due to the discriminatory power and the efficient functional extraction revealed. Many approaches used in traditional crowd analysis were unsuitable for modern surveillance due to certain limitations. Ordinarily, modern surveillance systems are characterized by intense uncertainties and dynamicity in crowd motion trends, and the operating circumstances of surveillance tools. This diverse characteristic can complicate the use of current techniques in the monitoring and analysis of the dense crowd. Crowd analysis researchers should develop novel techniques to respond to the concern in the new environment where computer vision is increasingly needed to monitor and analyses many people from video feeds of the surveillance cameras in real time. This includes estimating the diversity of the crowd as well as the density distribution across the entire collection region. Identifying areas above protection can help in issuing previous alerts and could prevent crowd crushes. The estimate of the number of crowds also helps to quantify the importance of the event and the logistics and infrastructure of the event.
1.3 Challenges and gaps

In the last several years, the use of a Fully Convolutional Neural Network (FCNN) has steadily gained prominence for crowd analysis and monitoring. Nowadays, it is very important to perform video analysis and monitor the crowd density of the pilgrims and detect any abnormal movement. To achieve this, there is a need for state-of-the-art technologies such as deep learning. There is a big challenge in analyzing images or videos that involve the movement of large numbers of pilgrims with density ranging between 7 to 8 people per square meter. Besides, it is very difficult to determine density and spot suspicious activity due to the need to recognize effective monitoring features in the extremely concentrated nature of hajj. The issue of using the non-stationary tracking camera as a feed for the crowd video also needs to be addressed.

Most current works analyses the density of the crowd based on face detection to count individual people in the crowd. However, the analysis of crowds based on face detection in highly dense circumstances (> 2000 people) presents several difficult situations. The videos suffer from intense occupation, making the conventional face/individual detectors ineffective. Furthermore, the variety of angles that introduce the problem of perspective may make the capture of crowd videos difficult. These issues require the estimated model to be scale-invariant for large-scope adjustments when the group is not uniformly scaled. Besides, annotating a dense crowd is an exhausting task and it is also not easy to get or to build a good data set that represents huge crowd numbers where terrible incidents might occur from it.

1.4 Research questions

This research will attempt to answer the following key questions:

- What are the main difficulties faced in crowd monitoring?
- What are some challenges faced during Hajj?
- What are the impacts of crowd monitoring on the pilgrims?
- What are the major algorithms involved in the crowd analysis domain?
- What are the most important datasets in this field of research?

1.5 Contributions

This article focuses on reviewing the latest crowd video analysis technology from the current video surveillance system. The latest approach for crowd analysis relies on the deeply discovered features from the use of the Fully Convolutional Neural Network (FCNN) architecture. We have categorized the related works into two main branches: network-based and image-based. We reviewed the Convolutional Neural Network (CNN) strategies to illustrate the shortcomings and core characteristics in each branch. In addition, we provide detailed analyses on the different approaches in each of the two branches (in terms of n Mean Absolute Error (nMAE) and a full output basis in separate data sets such as UCF, World Expo (WE), Shanghai Tech PartA (STA) and Shanghai Tech PartB (STB).
1.6 Comparison of the proposed work with existing works

Existing works Bendali-Braham, M. et al. [9] analyzed numerous crowd analysis publications. Crowd analysis has two main branches: statistics and behavior. Crowd behavior analysis often discusses anomaly detection. Anomalies can occur in any of the crowd behavior analysis subtopics. The aim of this study is to find unexplored or understudied crowd analysis sub-areas that might benefit from Deep Learning.

Kumar and Arunnehru [42] reviewed the literature on crowds, including methodologies for crowd surveillance and behavior analysis. The author also described the datasets and methodologies used. Various methodologies and current deep learning ideas have been evaluated. This work explains the many modern methodologies for crowd monitoring and analysis.

Albattah, W. et al. [3] proposed an image classification, crowd management, and warning system for the Hajj. Images are classified using CNN, which is a deep learning technology that has recently acquired an interest in numerous applications of image classification and speech recognition for the scientific and industrial communities. The goal is to train the CNN model on mapped picture data to classify crowds as severely packed, crowded, semi-crowded, lightly congested, and normal.

Proposed work This study examines current methods and approaches for crowd analysis from crowd videos, with an emphasis on deep learning techniques for detecting anomalous behavior. These findings motivate us to go on a time-consuming yet fascinating journey of crowd analysis, classification, and detection of any abnormal Hajj pilgrim activity. This study also pushes us to critically evaluate the crowd on a huge scale since the Hajj pilgrimage is the most crowded arena for video-related intensive research activities.

1.7 Paper organization

The rest of the paper is organized as follows: Section 2 gives the information about the research works on the crowd analysis domain, Section 3 describes selected studies on crowd analysis Section 3.1 highlights the unsolved problem that still exists in the domain and possible future research direction. Section 4 presents different categories of CNN techniques. Section 5 concludes the review and details are shown in Fig. 1.

2 Background studies on crowd analysis

In the previous section, the focus was on the introductory aspects of this work. This section focuses on existing research works related to crowd analysis. We consider crowd analysis using global regression, deep learning, scene labelling data-driven approaches, detection-based methods, CNN-based methods, optical flow detection, Object Tracking, Convolutional Neural Network (2D), 3D Convolutional Neural network, crowd anomaly detection, abnormal event detection for deep model, feature learning based on the PCANet, representation of neural event patterns with deep GMM.
2.1 Crowd analysis by global regression

Many experiments are being developed for the monitoring of pedestrian crowds by sensing or clustering of trajectories [11, 77]. However, the techniques are restricted by serious occlusions among people. Specific methods for global count predictions were introduced using low-level-trained regression [13, 14]. These methods are more suitable for crowded situations and are more efficient in computation.

Lempitsky et al. [43] proposed crowd detection analysis based on the regression of the pixel-level object density map. Fiaschi et al. [27] subsequently employed a random forest to reduce the density of objects to improve training efficiency. In addition to the consideration of space information, a further advantage of regression-based methods is their ability to estimate the number of objects in a video region. Taking advantage of this, the users were able to unveil an interactive object counter system, which can visualize regions, to determine the relevant feedback efficiently [14].
To solve the problem of occlusion, regression-based crowd counting techniques have been developed. The core concept behind regression approaches is to learn how low-level imaging patches are assigned to the number \([17, 61]\). The removed functions include forecast, border components, textures, and gradient features like local binary pattern (LBP) and Histogram of Oriented Gradients (HOG). Returns include linear regression \([14]\), and part linear regression \([17]\). These methods improve past identification-based methods while neglecting knowledge of spatial distribution by crowds. To use spatial distribution information, a density map is replaced by a solution suggested by Lempitsky et al. \([43]\). A linear representation of local patch functions and maps are obtained, and, by combining the entire density maps, the total number of images can then be determined. Pham et al. \([59]\) can draw a non-linear map of local density patches using the random forests. The density map is based on the most recent approaches to regression.

### 2.2 Scene labelling data-driven approaches

According to the other well-published large-scale crowd application, data-driven methods are recommended in non-parameter format \([48, 60, 70]\). Such methods can quickly be enhanced as they do not require preparation. Data-driven methods move the marks from training image to the test image by looking for the most appropriate training photographs and thereby suit the test picture. A non-parameter image parsing technique, as suggested by Liu et al. \([48]\), searches for a dense area of deformation in between images. Powered by methods of data-driven scene labelling, we get identical scenes and audience patches from the training scenes for an unknown destination location.

### 2.3 Detection-based methods

To identify and count the pedestrians, detection-based crowd tracking techniques were proposed \([23]\). Some authors suggested extracting particular features from appearance-based crowds to count crowds \([71]\). However, these methods led to limited recognition of large crowds. Researchers used partial methods to detect parts of crowd bodies, like the head or shoulder, to count footmen to deal with this problem \([28]\).

### 2.4 Optical flow detection

Optical flow is a fully vector-based approach that estimates movement on crowd objects through picture frames with matches \([12]\). Optical approaches based solely on flow can be used to locate moving crowd objects independently while the subject is turning. It is an extremely dynamic and complex approximation algorithm. The space-time filtering approach uses multiple adjoining frames to remove and acquire one-off images dependent on time series. This approach cannot be used in real-time implementations for objects that are not moving at all.

### 2.5 Deep learning

To date, many previous experiments have applied deep learning to various monitoring systems, including the re-identification of individuals \([44]\), and pediatrics recognition \([58]\). Their popularity emerges from the influence of profound patterns. Sermanet et al. \([63]\) found
that deep models have features extracted that are more effective for many applications than a handcrafted feature. The CNN based algorithm used by [37] is largely based on the assumption that a vast number of individuals who rely on the use of a single function are very difficult to obtain. They use a blend of engineered products to overcome this: HOG mostly based on head detections, Fourier measurement, and counting points of interest. The method used the Markov Random field in several dimensions. However, the method suffers from accuracy decline and changes in weather, distortion of mood, extreme occlusion, etc. Zhang et al. use deep networks to evaluate the count of people [83]. Their model is guided by image maps. The development of such maps is a complicated procedure. Wang et al. teach a deep model for crowd estimation [74]. The network measures the crowd and the distribution of crowd density. A sample block of CNN model is shown Fig. 2 [10].

2.5.1 2D convolutional neural network

Neural networks are also known as neural networks with common weight. ConvNets are the multi-layered deeper neural networks that work with data from the real world. This is achieved by using sensitive zone fields (more commonly kernels) with the same parameters, defined as weight sharing, for all of the potential entry points. The theory is that any node from a previous layer is filled with a tiny kernel window. The distribution of weights across computing devices in the CNN system decreases the diversity of free variables, improving the overall group efficiency. Weights are repeated over the input file, allowing the translations within the data to be intrinsically insensitive.

Figure 3 displays a standard frame of convolution. A few planes (referred to Feature Map (FM)) are usually used on all layers to distinguish more than one element. The layers are known as convolutional layers. The group has a traditional gradient-descent propagation method. To define and derive spatial functions, 2D CNNs are applied to a video dataset.
2.5.2 3D convolutional neural network

The analysis uses the 3D Convolutional Neural Network platform for anomaly detection among crowds. However, a few fundamentals need to be understood before solving the anomaly detection. Convolution has two functions, m and n, which produces a third function \[ \overset{\star}{m} \times n \]. This is generally known as an altered version of one of the authentic elements, which introduces the difference between the two functions equivalent to the quantity of one of their original functions. The m and n are written as \( m \ast n \) and are lined with a megastar or an Asterisk. The measures are defined by one reversed and transferred to the integral of the product of the two features. As such, it measures a specific kind of integral transform: (m x n) \( (t) = \int_{-\infty}^{\infty} m(t) n(t-T) dT \) while the symbol t used above does not represent the time area. Figure 4 illustrates the 3D convolution.

In the time dimension, the size of the convolution kernel is 3. The shared weights are colored to match the connection settings. 3D convolution processes the overlapping 3D cubes in the input video to get motion information, as is done in 3D convolution.

Note that only one feature can be extracted from the frame cube using a 3D convolution kernel, therefore the kernel weights may be applied again to the whole cube. CNN has a general design philosophy in which they try to produce as many feature maps as possible by drawing from several sorts of features extracted from the same kind of low-level feature maps. This is done by applying numerous 3D convolution operations to the preceding layer at the same place with distinct cores as depicted in Fig. 5.

Consecutive frames may be convolved with numerous 3D convolutions to extract numerous features. Figure 1 color codes the connection set to show shared weights, making them seem to be the same color. All 6 connected sets have unique weights, and thus you get two feature maps on the right, even though they all link to the same subset of data.

2.6 Crowd anomaly detection

Event classification is usually impossible for testing, and an event variant is evaluated in standard videos by current algorithms [78]. Irregular events are detected by breaking up the proposed algorithms to identify normally occurring events into two forms, namely: (1) trajectory dependent techniques. However, abnormal trajectories are very rare as opposed to daily trajectories. 2) Local algorithms based on lines. Here, anomalies are viewed as chains that integrate dramatic case patterns. Exploring possible laws across ordinary paths defines odd behaviors as disobedient to such policies. Experience several skills, including the trajectory
interpretation, speed, and acceleration, dependent on trajectories [5]. The function of its methodology is conducted with a cluster collection, and the ultimate cluster outcomes are achieved by considering clusters with all capabilities. Anomalies are viewed as clusters with few members and tests far from these cluster centers, via adaptive particle sampling and Kalman filtering, problems with occlusion and segmentation can be dealt with [18]. In other works, traceability was also considered at the particle and function factor level. An example using an approach to particle dynamics was proposed by [78], which derives messy invariant properties from symbolic pathways. Cui et al. [20] demonstrate the direction of interest and reflect the dynamism of the crowd through the use of potential measurements.
2.6.1 Abnormal event detection for deep model

At this stage, the method proposed by [26] is defined as an element. The 3D gradient is initially calculated for each video frame. Secondly, excessive-stage skills are routinely extracted from PCANet for video events. Deep GMM is used in regular pattern modelling [26].

2.6.2 Feature learning based on the PCANet

Spatial and temporal characteristics such as intensity, color, gradient, and optical flow are manually selected for most current techniques. The ability to gradient 3D is calculated for video events in this paper. Research on the power and effectiveness of 3D gradient functions for unusual occasional detection [52]. In the meantime, both look and motion clues for the 3D gradient prototype. This paper uses a deep neural network to abstract high-tier capabilities, primarily based on 3-d gradients. Deep mastering in many applications of computer vision has achieved significant performance in the last few years [15]. This benefits from the non-linear multi-layer variations, which can extract meaningful and discriminatory characteristics adaptively. There are no marked unusual activities for schooling in the field of anomaly detection. The training dataset is simply supplied with the handiest regular videos. Therefore, this article will learn from video opportunities PCANet capabilities, which are simple and effective, uncontrolled approaches [15].
2.6.3 Representation of Normal event patterns with deep GMM

GMMs are used in many works to examine daily cycles of events. But the simulation of complex video occasions includes significant single Gaussian additives [8, 24, 53]. The dynamics of these approaches eventually expand dramatically. A deep GMM is now available in this paper as a standard edition of video [72]. Figure 4 indicates the activation of deep GMM. Figure 6 showed the Deep GMM pattern.

3 Selected studies on crowd analysis

This part gives a thorough overview of recently selected and examined crowd analysis investigations in specific works that have been published within the previous five years. In the current and future surveillance systems, the works have been carefully selected and adapted to satisfy the difficult requirement of much needed dense crowd analysis.

Bendali-Braham, M. et al. [9] reviewed several papers related to crowd analysis. Crowd analysis is often divided into two branches: crowd statistics and crowd behavior analysis. Anomaly detection is one of the much talks on crowd behavior analysis. Although there is no universal definition of an anomaly, each of the crowd behavior analysis subtopics may be prone to an abnormality. The goal of the study is to identify crowd analysis sub-areas that have yet to be investigated or that seem to be seldom addressed via the prism of Deep Learning.

Kumar, A. and Arunnehru, J. [42] presented literature studies for organized and unorganized crowds, as well as approaches for crowd monitoring and behavior analysis. The author also gave a description of the datasets and the techniques provided for them. [6] Different methods based on traditional techniques and on modern profound learning principles have been reviewed. This publication helps researchers to comprehend the many state-of-the-art approaches utilized for monitoring and analysis of crowds.

Albattah, W. et al. [3] presented an image classification crowd control system and an alerting system to control millions of Hajj crowds. The image classification system is heavily based on the proper dataset utilized for the formation of the CNN, which is a profound learning

![Visualizations of single Gaussian, GMM, and deep GMM distribution](88)
methodology that has lately gained attention in many applications of image classification and voice recognition for the scientific community and industry. The objective is to train and make the CNN model accessible for usage with mapped image data in the classification of crowds as heavily-crowded, crowded, and semi-crowded, light crowded and normal.

Dargan, S. et al. [21] concentrated on the basic and advanced structures, methodology, motivating factors, characteristics and limitations of deep learning concepts. In this paper, the main problems of deep learning, traditional machine learning and conventional training were also marked by considerable disparities. It is centered on studying and chronologically studying the numerous applications of deep learning, as well as methodologies and structures used in a variety of fields.

Gupta, Kumar and Garg [32] proposed to identify objects having hand-crafted functionality based on Oriented Fast and Rotated BRIEFs and Invariant Feature Transformations. Size Invariant Transforming Feature Invariant Si (SIFT) is very effective in analyzing various pictures of orientation and size. A strategy for reducing the size of the image characteristic vector is being researched. For testing the realization of the project work, K-NN, decision tree and random forest classifiers are utilized.

In Wang et al. [76], a large congested crowd counting dataset called NWPU-Crowd has been built. This was intended to tackle the problem of small datasets, limiting the need to meet the CNN algorithms that are supervised. The built dataset includes several lighting scenes and has the widest range of densities (0 to 20, 33). Besides, they developed a benchmark website, which allows researchers to submit the findings of the trial set impartially. The data characteristics are further explained based on the proposed dataset and performance based on certain mainstream state-of-the-art (SOTA) methods.

In Zeng et al. [82], the DSPNet, which encodes multidimensional features for large crowd counts, was proposed as a modern deep learning network. It is a question addressing especially the current challenge of counting numbers in highly congested scenes because of scale variant. First, an interface and context were used in the DSPNet model. The frontline is a deep-neural network, while the deep-neural centralized backend network uses a “complete integration ratio of information at different levels. The (SCA) module cleaner is capable of effectively integrating the multiscale functions and improving image representations.

Singh, K. et al. [68] proposed to detect visual anomalies in crowded scenes using the ConvNets and the classification pool capabilities, the new principles of Aggregation of Ensembles (AOE). The plan used a collection of various finely tuned Convolutional Neural Networks based on the idea that improved feature sets from different CNN architects. For the creation of versions of the SVM type, the suggested AOE definition used the finely tuned ConvNets for a fixed-function extractor. It then combined the chances of detecting deviations in the crowd frame sequences. Experimental findings suggested that the proposed aggregation of finely tuned CNNs from various architectures is more efficient, in contrast to the other existing approaches in benchmarks.

Tian et al. [69] proposed a modern understanding called the multi-density count, to compute crowds in different densities. The Density Conscious Network consists of several sub-networks pre-trained in various densities. First of all this module collects pan-density information. Secondly, the feature enhancement layer (FEL) captures global and local qualitative characteristics and produces weight for each density function. Then, the spatial background is integrated into the Fusion Function Network (FFN), which understands these unique density functions. To help measure the efficiency of the global and local forecast, Patch MAE (PMAE) and Patch RMSE (PRMSE) metrics were also introduced. Extensive testing of four
crowd counting datasets, ShanghaiTech, UCF_CC_50, UCSD, and UCFQNRF, showed that PaDNet achieved state-of-the-art efficiencies and high energy levels.

Liu et al. [51] introduced the deeper and end-to-end trainable architecture that blends features obtained with several size receptive fields. In other words, the methodology adapted to the level of background data required to forecast crowd density accurately. This results in an algorithm that surpasses the latest methods of crowd counting, particularly in the case of strong viewpoints.

Hossain et al. [34] proposed to tackle crowd analysis problems with a revived scale-aware attention network. Their model automatically concentrated on some image-appropriate global and local scales using the attention framework commonly in recent profound learning architectures. Combining these attentions on national and local scales, the platform provides a range of state-of-the-art crowd data sets.

Gao, Wang, and Li [31] suggested a Perspective Crowd Counting Network (PCC Net) that consists of three parts: 1) Density Map Estimation (DME) that focuses on very local learning features for estimating the density maps; 2) Random High-Level Density Classification (RHDC) that extracts regional characteristics for predicting coarse density points for random image patches; To express variations in viewpoints in four directions (Down, Up, Left and Right) the DULR module is also implemented in PCC Net. Five standard data sets are evaluated with the new PCC Net to produce the most innovative performance and successful analyses with four additional data sets.

Liu et al. [49] proposed the DecidNet (Detection and Density Estimation Network) as the novel end-to-end crowd counting system. It changes the acceptable counting mode for different image positions depending on its true density conditions. DecideNet starts by measuring the multitude of density and generating different maps for identification and regression. It provides an attention element to calculate the precision of the two forms of estimates in an efficient way to detect unavoidable differences in densities. The final crowd counts are collected from all forms of density maps with the help of the attention module. Experimental findings showed that the proposed approach is at the leading edge of three rigorous data sets for crowd counting.

Li, Y., Zhang, X. and Chen, D. [45] proposed CSRNet that consists of two major components: the front end for 2D retrieval of information, the convolutional Neural Network (CNN), and the back end, a dilated CNN with dilated kernels that can be used to create wider processing areas and eliminate pooling operations. Based on its pure convolutional form, CSRNet is an easily trained model. [76] Applied CSRNet on four datasets (Shanghai Tech, UCF CC 50, World EXPO 10, and UCSD), delivering comparable results. CSRNet has a mean actual error (MAE) that is 47.3% less than the previous state-of-the-art system in the Shanghai Tech Part B dataset. They expand targeted uses, such as vehicles in the TRANCOS data collection, to other objects. Results demonstrated that, with a MAE lower than that of its previous state-of-the-art approach, CSRNet greatly increased the prediction efficiency.

Liu et al. [50] proposed the Deep Recurrence Space-Aware Network Unified Neural Networks Framework that addresses the two issues with an area-oriented process of refinement in a learning spatial transformation module. In particular, this architecture includes a recurrent spatial refinement module (RSAR) which iteratively performs two components: (i) the spatial transformers network which locates the attention-related field dynamically from the crowds’ map and renders them to the correct scale and rotation for optimal estimation of the crowd; Comprehensive experiments on four challenges demonstrate the effectiveness of approach. Specifically, we will obtain an increase of 12% in the biggest data collection, WorldExpo 10,
and a 22.8% change in the most challenging dataset UCF CC 50, relative to the current best-performing approaches.

Idrees et al. [38] suggested a novel method that tackles the counting, calculation of density maps, and position of people in a crowded crowd picture at the same time. The terminology depends on the essential presumption that the three issues are fundamentally associated, such that the breakdown feature is decomposable to clarify a profound CNN. Given the need to find images and annotations of good quality, UCF-QNRF data collection is initiated to fix vulnerabilities in previous data sets and manually involves 1,25,000 users with dot annotation labels. Finally, they employed estimation approaches, including those developed specially to count the crowds in comparison to the previous deep networks of CNN. This is the most complex data series with the most dynamic situations with the highest amount of crowd annotations, and the method.

Marsden et al. [54] suggested the deep residual architecture Resnet Crowd, monitoring of aggressive activity, and classification of crowd mass. A new 100 image data collection known as multitask crowd is developed to test and assess the emerging multi-objective system. The latest data collection is the first completely annotated computer vision data set to count the crowd, identify aggression and degree of the scale. The experiment shows that multitask approach improved individual task performance for all tasks, particularly violent behavior, which is up to 9% in the AUC (Area below the Curve) ROC curve. Many qualified versions of the Resnet Crowd have tested additional metrics underlining the superior generalization of the multi-target research models.

Sindagi and Patel [65] proposed a novel network of CNNs with cascades that collectively the classification and the computation of the density map. The division of the crowd into different categories amounts to a gross estimate of the total number of the picture. Hence, a high degree is integrated into the density measurement network. It allows the network layers to understand global discrimination, which helps to approximate increasingly defined density maps with smaller counting errors. Shared preparation fully takes place. Extensive experiments with publicly available highly challenging datasets have shown that the approach proposed achieves less count errors and better density maps compared to the latest methods.

In Bansal et al. [7] three common feature descriptor techniques are employed in this research for object identification system experimental work: Scale Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF), and Oriented Fast and Rotated BRIEF (ORB). The purpose of this article is to compare the performance of these three feature extraction approaches, especially when their combination leads in a more efficient recognition of the object. The authors conducted a comparative investigation of several feature descriptor techniques and classification models for 2D object recognition in this article.

In Elbishlawi et al. [25] this article discusses approaches for assessing congested situations that are based on deep learning. The approaches evaluated fall into two categories: (1) crowd counting and (2) crowd detection and recognition. Additionally, crowd scene datasets are analyzed. Along with the surveys mentioned above, this article presents an assessment measure for crowd scene analysis methodologies. This measure quantifies the discrepancy between the estimated and real crowd counts in crowd scene footage.

A variety of CNN-based approaches were adopted in crowd counting, taking advantage of the powerful capacity of CNN to understand representations. The Wang et al. [74] methodology, as a pioneer for CNN crowd counting, introduced several convolutional layers to extract features and transmitted these features into a fully connected layer for predicting the number in extremely dense crowds. In further work [45, 46, 57, 74, 81, 84, and], a network was pre-
trained for some scenes, and related training data was chosen to optimize the pre-trained network based on prospective information. The key downside is that it does not always have accessible viewpoint details. Zhang et al. [85] further suggested (MCNN) architecture to approximate the density diagram, as it noted that the densities and appearances of image patches differ greatly. Diverse columns are deliberately designed to know differences in density by multiple feature resolutions throughout their research. Given multiple image sizes, separating columns to identify density crowds is difficult, resulting in some inefficient divisions due to lack of identification. To simultaneously predict the density grading and construct a map based on high-level knowledge, Sindagi et al. [66] suggested a multitask system. They also suggested a 5-branch CNN contextual pyramid short for CP-CNN Sindagi and Patel, [65], which would incorporate popular contextual information to lower the number of errors and high-quality maps. However, CP-CNN cannot be used for real-time scenario analysis. Sam et al. [62], inspired by MCNN, uses a switch-CNN, in which the switch classifier is educated in choosing the best regression for an input patch. Switch-CNN will use a column network only in conjunction with that patch’s classified results during the prediction process, without including any of the qualified subnetworks. Not only at the overall image level, but even at the image patch level, there are high vector densities. There is also no identification efficiency on the single subnet, and the computational covariation change issue cannot be solved. Kang et al. [41] proposed the fusion of multiscale input density predictions, while Deb et al. [22] developed an aggregate multi-column dilated conversion network for the free counting of perspectives.

3.1 Challenges and future direction

| Study          | Objective                                                                                                                                   | Issues and Challenges                                                                                       | Dataset and Accuracy                                      |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------|
| Gao et al. [30] | Explored the crowd counting models, to estimate density maps based on CNN. Compared the efficiency of crowd counting based on the data sets | • The research attempts to draw rational conclusions and predictions regarding the potential growth of crowd counting and, in the meantime, will provide workable solutions in other fields to the issue of counting items.  
• The density maps and prediction outcomes for comparison and checking of certain standard algorithms in the NWPU validation kit. In the meantime, density charts and measurement instruments are also developed. | NWPU dataset for comparison and testing.                                                        |
| Wang et al. [76]| • Created NWPU-Crowd dataset evaluated the output of state-of-the-art (SOTA) methods using this data.                                    | • Current databases available are so limited that they cannot meet the criteria of the algorithms tracked by CNN.  
• NWPU dataset  
In particular, MAE dropped in the classes by 36.6%, 25.7%, 22.2% and 12.7%, respectively. |                                                            |
| Study                      | Objective                                                                 | Issues and Challenges                                                                 | Dataset and Accuracy                                                                 |
|---------------------------|---------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Ilyas, Shahzad and Kim,   | Employed ML and AI for image, classical artisan crowd counting.           | With many questions, such as occlusion, noise, and unequal distribution of objects and unequal object sizes, neural networks are promising advances in the counting of and perception of smart camera crowds. | The NWPU crowd dataset has more crowd scenes than previous datasets.                   |
| [39]                      |                                                                           |                         | This is a review paper. They have done some and compared the previous work.          |
| Li et al. [47]            | Devised the most advanced CNNs through a multi-vision framework           | With the aid of complementary data obtained from many cameras to have a clearer insight into the observation area, the issue of restricted view and occlusion in single views can be answered. | This paper used PETS2009. The accuracy of this paper went from 83.2% to 89.8%.         |
| Singh et al. [68]         | Suggested a novel AOE to improve ConvNets’ capacities and classification pools. | The big obstacle to identifying the anomalies efficiently in crowds is to use function sets and strategies replicable in any crowded situation. | Proposed dataset UCSD Ped-1, UCSD Ped-2. Accuracy state-of-the-art on UCSD Ped-1 (0.946), and UCSD Ped-2 (0.959). |
| Zeng et al. [82]          | Proposed a DSPNet Framework based on both backend and frontend CNN. DSPNet conducts the entire RGB-based image data analysis to promote model learning and minimize the lack of contextual details. | They proposed a novel DSPNet (Depth Scale Cleanser Network) that encodes multiple features and reduces the loss of context information for dense crowd counting to resolve this issue. | Three public datasets are used UCF-QNRF, ShanghaiTech, and UCF_CC_50 datasets. Comparison of DSPNet with state-of-the-art methods on the UCF-QNRF dataset (MAE 107.5 and RMSE 182.7) Comparison of DSPNet with state-of-the-art methods on the UCF_CC_50 dataset (MAE 243.3 and RMSE 307.6) Comparison of DSPNet with state-of-the-art methods on the ShanghaiTech dataset Part A (MAE 68.2 and RMSE 107.8) and ShanghaiTech dataset Part B (MAE 8.9 and RMSE 14.0) |
| Gao, Wang, and Li, [31]   | Proposed a Perspective Crowd Counting Network (PCC Net), consisting of DME based on local map estimates and Random R-HDC. | Due to high presentational similarity, shifts in vision, and the severity of congestion, the number of people in one image is challenging. | Public ShanghaiTech dataset is used for this experiment. Accuracy MAE of 11.0 (6.2-point improvement) and MSE of 19.0 (8.4-point improvement). |
| Liu, Salzmann and Fua,    | Devised a large, end to end trainable architecture                        | This creates an algorithm that goes beyond modern crowd counting methods,               | ShanghaiTech dataset Result: Part A (MAE 62.3 and RMSE 100) and Part B               |
| [51]                      |                                                                           |                         |                                                                                        |
Study | Objective | Issues and Challenges | Dataset and Accuracy
--- | --- | --- | ---
Hossain et al. [34] | • Consider the approximation of image density map, where each pixel size of the camera corresponds to the density of the crowd at the corresponding position of the image. | • The disparity in the scale of images is an obstacle for crowd counting. • This research suggests a new, size-conscious network of commitment to this task in this research. | • Several datasets used. • UCF-QNRF Dataset Result: (MAE 107 and RMSE 183). • UCF CC 50 dataset Result: (MAE 212.2 and RMSE 243.7) • World-Expo’10 dataset Result MAE 7.2 on average
Liu et al. [49] | • Considered an end-to-end crowd counting system | • This system is based on three demanding crowd counting data sets to provide state-of-the-art efficiency. | • Using Mall dataset DecideNet Result: MAE 1.52 and MSE 1.90. • For the five scenes tested, an average MAE of 9.23 was found. • This is the best result of all the comparisons done in this section, which was shown to provide 0.17 better than the second best “Switching-CNN technique.
Li, Zhang and Chen, [45] | • Considered an information-driven, in-depth learning framework to include a congested scene recognition network known as CSRNet | • The proposed CSRNet consists of two main components: the frontend neural network (CNN) for 2D object extraction, and a dilated, backend CNN for broader receiving areas and eliminating pooling operations by using dilated kernels. Owing to its pure convolutional form, CSRNet is an easily trained model. | • Used four datasets (ShanghaiTech dataset, the UCF CC 50 dataset, the WorldEXPO’10 dataset, and the UCSD dataset) • The results reveal that with 15.4% less MAE than the prior state-of-the-art methodology, CSRNet greatly increases quality in output.
Amirgholipour et al. [4] | • Considered an A-CCNN that captures system size variability to enhance accuracy. | • There are many attempts reported, challenges in the real world, such as large variations in image size and extreme occlusion among individuals, make this a challenge quite difficult. • The trained Resnet Crowd model will also be tested using some more | • Used two datasets (UCSD Dataset, UCF CC 50) • ACCNN performs favorably in upscale and minimum subsets versus the other techniques with the lowest MAE ever, 1.04 and 1.48. • Used UMN dataset on the performance of Crowd Anomaly detection and
Marsden et al. [55] | • Considered Resnet Crowd, a profound residual architecture for concurrent
| Study                  | Objective                                                                 | Issues and Challenges                                                                 | Dataset and Accuracy                                                                 |
|-----------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Sindagi and Patel, [65] | crowd counts, aggressive device identification, and classification of crowd mass. | • A Multi-Task Crowd data collection technique was proposed.                           | obtained accuracy (AUC 0.84).                                                          |
|                       |                                                                           | • Considered to jointly learn the multitude count classification and density map estimation proposed a new end-to-end cascade network of CNNs. | Result comparison: Error estimations on the UCF CC 50 dataset and got the accuracy, MAE 322.8 and MSE 397.9. |
|                       |                                                                           | • To divide the crowd count into different categories, the total numbers to the image are measured ground and integrated into the density estimation network at a high point. | Result comparison: Error estimations on the ShanghaiTech dataset and got the accuracy, Part A (MAE 101.3 and MSE 152.4) and Part B (MAE 20.0 and MSE 31.1). |
|                       |                                                                           | • It uses a reliable and functional crowd count estimator using computer vision techniques. | Used two datasets (ShanghaiTech dataset, UCF CC 50).                                   |
|                       |                                                                           | • Employed the concept of a completely convolutions-based paradigm for crowd counting posed in high-density scenes | Result achieved: MAE 126.5 and MSE 173.5 using ShanghaiTech dataset Part A and ShanghaiTech dataset Part B, MAE 23.76 and MSE 33.12. Method to improve the cutting-edge status for using UCF CC 50 dataset, obtain accuracy MAE and MSE by 11% and 13%. |

4 Categories of CNN techniques

4.1 CNN techniques

CNN development categorization plays a significant role at their granular level. The techniques allow researchers to develop algorithms for remote control and tracking systems in military combat, emergency management, public events, etc. with the various crowd counting systems. Figure 7 presented the categories of CNN based crowd analysis techniques.

There are two forms of public and private dataset available. Public databases are freely accessible via the internet and private databases are normally owned by their respective authors/organizations. We list 5 most common and most recognizable dataset and their basic characteristics in Table 1.
4.2 Basic CNN techniques

This section covers crowd counting for an architecture that contains a simple CNN. Simple CNN approaches may be regarded as leaders in in-depth density analysis, utilizing the basic design in their network to produce a real-time crowd counting. Table 2 displays the basic CNN features, used databases, and architectures.

Fu et al. [29] proposed a formula for measuring two-tier density using a simple CNN model. Their first task was the estimation of the distribution of crowds (i.e. the insulation in various amounts). Through deleting related connections, calculation speed is improved. The second task was to identify racial traits by using a cascade. Likewise, an initial layer-based learning approach to divide the number of motorcycles into overlapped patches was proposed in Mundhenk et al. [56] for counting the image region. To minimize MSE, a change was made to distinguish between unallocated cars and contextual details. Wang et al. [74] developed a variant of the Argumentation Technique for the FCNN to enhance the system robustness training knowledge on diverse and varied processes. Zhang et al. [86] introduced the CNN model of tracking videos to count the number of people crossing a fence. To address the

| Dataset | USCD [5] | Mall [5] | UCF [52] | WE [55, 61, 68, 56] | STA [69] | STB [69] |
|---------|----------|----------|----------|---------------------|---------|---------|
| Number of images | 2000 | 2000 | 50 | 3980 | 482 | 716 |
| Resolution | 158 × 238 | 320 × 240 | Varied | 576 × 720 | Varied | 768 × 1024 |
| Number of headcount (Minimum) | 11 | 13 | 94 | 1 | 33 | 9 |
| Number of headcount (Maximum) | 46 | 53 | 4543 | 253 | 3139 | 578 |
| Total amount of headcount | 49,885 | 62,325 | 63,974 | 199,923 | 241,677 | 88,488 |
complexity of the principal problem, the original issues were split into two sub-problems
(estimated crowd density and crowd speed). In Hu et al. [35], the author suggested a
fundamental research approach to approximate in images the mid to the high-level audience.
To approximate total density, a regression was applied to calculate the number of people in an
area by including the average local densities. Throughout the process of their research,
ConvNets software was used to approximate crowds throughout their respective local regions
to know a function vector. In many uses, including novel and outdoor counting, the writers in
[73] used a simple CNN. A layer raising or selective sampling (i.e., reducing inferior sample
effects of quality) reduces computation time and improves the accuracy of counting by
increasing the number and the iteratively generating of the new classifier to correct faults in
the previous classifier. Four ensemble networks, based on prior errors, make every network
possible.

It should be noted that the majority of strategies within this sub-category rely primarily on
the calculation of density rather than crowd count. Thanks to an over-simplified design, these
methods cannot be effective in strongly occult and diverse viewpoints. The level of the density
estimation in these techniques redundant samples can be enhanced by elimination. The
probability of errors can also be minimized by iteratively reducing errors in various network
layers.

4.3 Context CNN technique

This sub-category covers crowd counting tools to boost the precision of counting, leveraging
local and global contextual information. In a localized area, the spatial knowledge of a picture
means an overview change of adjacent pixels (i.e., contextual information). The company
makes very useful technologies in companies that contain figures, such as the number of flying
drones or cars in parking areas. These approaches also help to achieve resolution and
distribution of various distance-dependent images. The context-CNN with its characteristics,
data sets, and architectures is seen in Table 3.

It ought to be noted for dilated convolution in real-time, qualitative data may be used.
Deeper CNNs are primarily used to improve the density map performance and to maximize the
estimation accuracy by an adaptive distribution network. But at the cost of greater network
sophistication, this qualitative knowledge is collected. In this sub-category, techniques for real-time applications with low complexity requirements would not be feasible.

### 4.4 Scale-CNN technique

Basic-CNN techniques that have developed in terms of size variations (in order to increase robustness and accuracy) are termed Scale-CNN techniques. Size variance means that the resolution affected by multiple viewpoints varies. Contextual knowledge of the image indicates a relation to the general place for the total creation of neighboring pixels (i.e., adjacent images). Strategies within this framework are often very beneficial for projects that need quantitative statistics, such as the number of flying drones or parking vehicles [64, 67]. These methods may also be used to achieve depth ranges and distribution based on the distance of such images. Table 3 describes its features, common databases, and structures in the history of CNN.

In Chattopadhyay et al. [16] for example, authors suggested the concept of a regular item count to consider the novel concept of associative subtlety (the capacity of humans to provide rapid counting estimates/evaluations for small objects). Zhang et al. [87] suggested a focus model (high probability indicates the head position) for head position recognition. Multiscale divisions also marginalized the non-head area. Li et al. [45] merged the CNN and distributed convolution to increase the accuracy of a density map (extension kernels to bypass pooling). In different congested scenarios, an extended convolutional layer was often used to integrate contextual knowledge (Table 4).

Han et al. [33] suggested in silent images a spontaneous CNN Markov multitude counting system. The entire image was divided into small overlapping patches to strip functionality from patches and to reverse patch count by using fully connected NN’s. Because of overlaps, the next patches are strongly linked. To increase the overall precision of the crowd count, MRF used the association to smooth counts across the adjacent local patches. In Wang et al. [75] the

| Table 3 Context-CNN algorithms description |
|--------------------------------------------|
| Technique                    | Features                      | Datasets         | Architecture     |
|-----------------------------|-------------------------------|------------------|------------------|
| Chattopadhyay et al. [16]   | Associative subtilizing       | PASCAL VOC, COCO | “ConvNets”       |
| Zhang et al. [87]           | Attention model for head detection | UCF, STA, STB   | AM-CNN           |
| Li et al. [45]              | Dilated convolution and multiscale contextual information | UCF, STA, STB, WE | CSRNet           |
| Han et al. [33]             | Combination of correlation and MRF | UCF            | ResNet           |
| Wang et al. [75]            | Density adoption network      | ST, UCF         | DAN, LCN, HCN    |
| Liu et al. [50]             | Spatially aware network       | ST, UCF, WE     | Local Refinement Network |

| Table 4 Scale-CNN algorithms description |
|------------------------------------------|
| Technique                  | Features                                | Datasets         | Architecture     |
|----------------------------|-----------------------------------------|------------------|------------------|
| Liu et al. [50]            | Geometry-aware crowd counting           | ST, WE, Venice   | Siamese          |
| Huang et al. [36]          | Exploits cross-scale similarity         | ST, WE           | Wide and Deep    |
| Kang et al. [41]           | Image pyramid to deal with scale variation information | ST, WE, USCD | VGG network      |
| Zeng et al. [82]           | Single multiscale column                | ST, UCF         | Inception        |
authors suggested a network based on density tolerance to count the number of objects correctly. A common system, educated on one dataset and then based on another, was suggested. The degree of density was determined by choosing a network educated in numerous data sets. The system contained three networks: a medium or high-density network that changed and the two other networks that counted. Liu et al. [50], a profoundly spatially aware replicated network, proposed using a space transformer feature as a tracker for both size and rotational changes.

It ought to be noted that for dilated convolution, real-time spatial information can be used. In particular, it is possible to use a larger, dilated CNN to increase the map performance and to optimize measurement accuracy by using an adaptive density net. These conditions, however, come at the cost of increased reliability of the network. Therefore, for modular structures with low complexity specifications, approaches in this article could not be feasible.

4.5 Multi task-CNN techniques

CNN methods, which consider not just crowd counting but others, such as sorting, segments and uncertainty calculation, and crowd behavior monitoring, are the multiple-task CNN techniques. We analyze the association between these different tasks and their impact on the results in the sense of multitasking CNN.

In [6], the authors suggested an architecture from ConvNet to count the number of penguins. Because of the occlusion and the various sizes, a multitasking computing system was recommended to address foreground segmentation and depth prediction uncertainty. The multitasking methodology has been studied by Idrees et al. [37] three major problems are connected: crowd count, an estimate of density, and location. The density estimation and location enabled the counting process. The deep and shallow FCN is being suggested by Zhu et al. [89]. Features taken from a profound FCN were merged with two deconvolution layers to render the output image identical in resolution to the input image. Huang et al. [36] suggested a CNN-based crowd processing methodology instead of relying on the visual properties scaling. The problem with crowd counting was broken down into a multitask problem in their study. The multitasks included the extraction and counting of significant quantities of semantic knowledge and the mapping to the semantic model (body map and density map) of the input point. Yang et al. [81] proposed a multi-column neural network (MNCN) for resolving significant differences in scale. The multi-columns have been used with three main improvements. Second, up and down scanning was used to assess multiscale functionality. Second, for sampling errors, deconvolution has been used. Thirdly, costs per size were reduced to enhance the understanding of applications. Liu et al. [50] suggested a self-managed system of improving data for better accuracy testing.

It should be noted firstly, where a smaller patch is cut (with a lesser or equivalent number of items compared to the larger patch), training results may be enhanced. Second, interplaying tasks can improve counting precision. Third, the deconvolution should be used to render the density diagram more accurate. Finally, certain activities lead to growing the network’s overall efficiency, increased network reliability and decreased program usage in real-time.

4.6 CNN techniques image view

This group focuses on the study of an input image and also on the network architecture to increase network accuracy, which is very useful for medical imaging, drone surveillance in
particular areas, and CCTV tracking. Since the slope, tilt, and location of the camera in support
of the target performance play a crucial role in the development of every algorithm, we
specifically split up the CNN picture view into two sub-categories: CNN and Perspective-
CNN.

4.7 CNN techniques patch-based

In a patch-based approach, the CNN shall be equipped with a sliding window over the test
image using the cut bits. This is especially helpful as density maps are more reliable and cannot
be compromised, e.g., in the treatment of cancer. Both the cell count and cell resolution
affected are important. The key objective is to create a device for better density maps at a high
computational cost.

Cohen et al. [19] suggested a CNN network that was inspired by Deep. A smaller network
than the measurement of the crowd count calculates the number of events in a given area. The
authors suggested in the article that DecideNet should be employed by overestimating the
number of crowds in scarce areas by regression-based approaches and by underestimating the
number of crowds in density-dependent compact zones [31, 49]. The authors in [80] suggested
the optimized approach for the flow of data through various convolution and disengagement
rates, inspired by a skip link system for crowd counting. Edges and colors with convolution
layers were identified, but this low-level knowledge from an early stage may or may not have
led to improve the network’s means of absolute error (MAE) output. U-Net was employed to
evaluate how much information was sent to the final layer (convolution or fully linked) to
provide a more efficient feature-selection mechanism. Similar to the principle of [31, 49, 81]
proposed to work with very complex, varied images, a deep information-oriented crowd-
counting approach (Digcrowd). Segmentation took place on an image to break it into one part:
far-off parts. In the area of the near-view, detections are counted, and in the far-sights zone,
Digcrowd maps are used to map individuals to their density index. In Shami et al. [84] the
authors used an ahead detector to determine a human head’s specific size. The SVM classifier
classified crowded and uncrowded patches after dividing it into several patches. On each
patch, the head scale regression was done. If the head size has been calculated, a separating
region of the head size decides the overall number of the heads in a specific patch. In Zhang
et al. [87] by filtering the history on the headline, the authors proposed a count-net-technique.
Attributes have also been collected and measured at the same time. Zhang et al. [87]
introduced a density approach methodology for patch dependent CNN crowds with depth
approximation-accessible kernels. The different receptor field sizes used for each CNN column
are used for the treatment of various scale items (heads). However, the addition of the density
map at the end may have diminished the exactness of the density estimation map. In another
work, the Skip-connection CNN (SCNN) was suggested by Wang et al. [75] for crowd
counting. For the extraction of different functions, the whole network used four multi-units.
Three convolution layers were each composed of a multiscale array. Multi-sized units for
removing the various size characteristics were used. Besides, the two patches (with separate
scales) from each input image were applied to an incrimination technique (with no redundan-
cy) [79]. In these two scales, CNN was personally equipped to tackle any dramatic shifts in
size. In consideration, three returners specialized in low-, medium- and high-density pictures,
Sam et al. [62] suggested a CNN transfer technology. To address some density variation
issues, the entry patch was led to a separate regression by using a default (classifier).
It ought to be noted that the identification and regression on targeted image patches can be used sequentially to increase the estimation of the network accuracy. Besides, low-level network edge and color information can be filtered iteratively to reduce network computing costs.

5 Conclusion

This paper has reviewed different approaches, techniques and frameworks used for crowd analysis in video monitoring and surveillance specifically in crowd analysis based on Hajj video surveillance. First, the paper provides a brief discussion of the existing deep learning frameworks. Second, the paper presents a review of selected FCNN and CNN techniques on density estimation. The CNN techniques were categorized into network, image, and training-based CNN. In addition, the categories were subdivided into two main branches. Third, we critically reviewed selected research works related to crowd analysis. Lastly, we presented a review of the works in each category, with focus on the key characteristics, data sets, and architectures used. We believe that this work will contribute towards bridging the research gaps in this field of study.

Acknowledgements

Multimedia University, Cyberjaya, Malaysia fully supported this research.

Declarations

Conflict of interest The authors hereby declare that there are no conflicts of interest in this research.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. “True Islam”. Quran-Islam.org. Archived from the original on 2013-03-13. Retrieved 2013-07-31.
2. “A history of hajj tragedies | world news”. London: theguardian.com. January 13, 2006. Archived from the original on August 29, 2013. Retrieved 2013-07-31.
3. Albattah W et al (2020) Hajj crowd management using CNN-based approach. Comput, Mat Continua 66(2):2183–2197. https://doi.org/10.32604/cmc.2020.014227.
4. Amirgholipour S et al (2018). ‘A-CCNN : Adaptive CCNN for Density Estimation and Crowd Counting Global Big Data Technologies Centre, University of Technology Sydney, Australia Quantitative Imaging, CSIRO Data61, Australia Institute for Sustainable Futures, University of Technology’. 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 948–952. https://doi.org/10.1109/ICIP.2018.8451399.
5. Anjum N, Cavallaro A (2008) Multifeature object trajectory clustering for video analysis. IEEE Trans Circ Syst Video Technol 18(11):1555–1564. https://doi.org/10.1109/TCSVT.2008.2005603
6. Arueta C, Lempitsky V, Zisserman A (2016) Counting in the wild. In Proceedings of the European Conference on Computer Vision, Amsterdam, the Netherlands, 8–16; pp. 483–498.
7. Bansal M, Kumar M, Kumar M (2021) 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. Multimed Tools Appl 80(12):18839–18857. https://doi.org/10.1007/s11042-021-10646-0
8. Basharat A, Gritai A, Shah M (2008) Learning object motion patterns for anomaly detection and improved object detection, in: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–8.
9. Bendali-Braham M et al (2021) ‘Recent trends in crowd analysis: a review’, machine learning with applications. Elsevier Ltd., 4(October 2020), p. 100023. https://doi.org/10.1016/j.mlwa.2021.100023
10. Boominathan L, Kurthiventi SSS, Venkatesh Babu R (2016) ‘CrowdNet: a deep convolutional network for dense crowd counting’. MM 2016 - Proceed 2016 ACM Multimedia Conf pp. 640–644. https://doi.org/10.1145/2964284.2967300
11. Brostow GJ, Cipolla R (2006) ‘Unsupervised bayesian detection of independent motion in crowds’, Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 1, pp. 594–601. https://doi.org/10.1109/CVPR.2006.320
12. Candamo J, Kasturi R (2010) Understanding transit scenes: a survey on human behavior-recognition algorithms. IEEE Trans Intell Transp Syst 11(1):206–224. https://doi.org/10.1109/TITS.2009.2030963
13. Chan AB, Vasconcelos N (2009) ‘Bayesian poison regression for crowd counting’. Proceed IEEE Int Conf Comput Vision, (Iccv). 545–551. https://doi.org/10.1109/ICCV.2009.5459191
14. Chan AB, Liang ZSJ, Vasconcelos N (2008) ‘Privacy preserving crowd monitoring: counting people without people models or tracking’, 26th IEEE conference on computer vision and pattern recognition. CVPR. https://doi.org/10.1109/CVPR.2008.4587659
15. Chan TH, … Ma Y (2015) PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017–5032. https://doi.org/10.1109/TIP.2015.2475625
16. Chattopadhyay P et al (2017) ‘Counting everyday objects in everyday scenes’, proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017, 2017-January, pp. 4428–4437. https://doi.org/10.1109/CVPR.2017.471
17. Chen K et al (2012) ‘Feature mining for localised crowd counting’, BMVC 2012 - electronic proceedings of the British machine vision conference 2012. https://doi.org/10.5244/C.26.21
18. Cheng HY, Hwang JN (2011) Integrated video object tracking with applications in trajectory-based event detection. J Visual Comm Image Represent. Elsevier ltd., 4(October 2020), p. 100023. https://doi.org/10.1109/MC.2014.42
19. Cohen JP et al (2017) ‘Count-ception: counting by fully convolutional redundant counting’, proceedings - 2017 IEEE international conference on computer vision workshops, ICCVW 2017, 2018-January, pp. 18–26. https://doi.org/10.1109/ICCVW.2017.7
20. Cui X et al (2011) ‘Abnormal detection using interaction energy potentials’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 3161–3167. https://doi.org/10.1109/CVPR.2011.5995558
21. Dargan S, … Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. Arch Comput Methods Eng. Springer Netherlands 27(4):1071–1092. https://doi.org/10.1007/s11831-019-09344-w
22. Deb D, Venturi J (2018) ‘An aggregated multicolumn dilated convolution network for perspective-free counting’, IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, 2018-June, pp. 308–317. https://doi.org/10.1109/CVPRW.2018.00057
23. Dollár P et al (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761. https://doi.org/10.1109/TPAMI.2011.155
24. Du D, Qi H, Huang Q, Zeng W, Zhang C (2013) Abnormal event detection in crowded scenes based on structural multi-scale motion interrelated patterns, in: proceedings of the IEEE international conference on multimedia and Expo, pp. 1–6.
25. Elbishlawi S, Abdelpakey MH, Eltantawy A, Shehata MS, Mohamed MM (2020) Deep learning-based crowd scene analysis survey. J Imaging 6(9):95. https://doi.org/10.3390/jimaging6090095
26. Feng Y, Yuan Y, Lu X (2017) ‘Learning deep event models for crowd anomaly detection’, Neurocomputing. Elsevier 219(July 2016):548–556. https://doi.org/10.1016/j.neucom.2016.09.063
27. Fiaschi T, Giannoni E, Taddei ML, Chiarugi P (2012) Globular adiponectin activates motility and regenerative traits of muscle satellite cells. PLoS One 7(5):e34782. https://doi.org/10.1371/journal.pone.0034782
28. Forsyth D (2014) Object detection with discriminatively trained part-based models’. Computer 47(2):6–7. https://doi.org/10.1109/MC.2014.42
29. Fu M, … Zhu C (2015) ‘Fast crowd density estimation with convolutional neural networks’, engineering applications of artificial intelligence. Elsevier 43:81–88. https://doi.org/10.1016/j.enganappai.2015.04.006
30. Gao, G. et al. (2020), 'CNN-based density estimation and crowd counting: a Survey', pp. 1–25. Available at: http://arxiv.org/abs/2003.12783.
31. Gao J, Wang Q, Li X (2019) PCC net: perspective crowd counting via spatial convolutional network. IEEE Trans Circ Syst Video Technol 30:1–1, 3498. https://doi.org/10.1109/tcsvt.2019.2919139
32. Gupta, S., Kumar, M. and Garg, A. (2019) ‘Improved object recognition results using SIFT and ORB feature detector’, multimedia tools and applications. Multimedia Tools and Applications, 78(23), pp. 34157–34171. doi: https://doi.org/10.1007/s11422-019-08232-6.
33. Han K et al (2017) Image crowd counting using convolutional neural network and Markov random field. J Advanc Comput Intell Inform 21(4):632–638. https://doi.org/10.20965/jaciii.2017.p0632
34. Hossain MA et al (2019) ‘Crowd counting using scale-aware attention networks’, proceedings - 2019 IEEE winter conference on applications of computer vision, WACV 2019, pp. 1280–1288. https://doi.org/10.1109/WACV.2019.00141
35. Hu Y, … Li T (2016) Dense crowd counting from still images with convolutional neural networks. J Visual Commun Image Representation Elsevier Inc 38:530–539. https://doi.org/10.1016/j.jvcir.2016.03.021
36. Huang S, … Han J (2018) Body structure aware deep crowd counting. IEEE Trans Image Process 27(3):1049–1059. https://doi.org/10.1109/TIP.2017.2740160
37. Idrees H et al (2013) ‘Multi-source multi-scale counting in extremely dense crowd images’. Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 2547–2554. https://doi.org/10.1109/CVPR.2013.329
38. Idrees H et al (2018) ‘Composition loss for counting, density map estimation and localization in dense crowds’, lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 11206 LNCS, pp. 544–559. https://doi.org/10.1007/978-3-030-01216-8_33
39. Ilyas N, Shahzad A, Kim K (2020) Convolutional-neural network-based image crowd counting: review, categorisation, analysis, and performance evaluation. Sensors (Switzerland) 20(1). https://doi.org/10.3390/s20010043
40. Ji S, … Yu K (2013) ‘3D convolutional neural networks for human action recognition’, IEEE transactions on pattern analysis and machine intelligence. IEEE 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59
41. Kang D, Chan A (2019) ‘Crowd counting by adaptively fusing predictions from an image pyramid’, British machine vision conference 2018. BMVC 2018:1–12
42. Kumar A, Arunnehru J (2021) Crowd behavior monitoring and analysis in surveillance applications: a survey crowd behavior monitoring and analysis in surveillance applications: a survey. Turkish J Comput Math Educ 12(7):2322–2336
43. Lempitsky V and Zisserman A (2010) “Learning To Count Objects in Images,” pp. 1–9
44. Liu C, Yuen J, Torralba A (2011) Sift flow: dense correspondence across scenes and its applications. TPAMI 33:978–994
45. Liu J et al (2018) ‘DecideNet: counting varying density crowds through attention guided detection and density estimation’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 5197–5206. https://doi.org/10.1109/CVPR.2018.00545
46. Liu L et al (2018) ‘Crowd counting using deep recurrent spatial-aware network’, IJCAI international joint conference on artificial intelligence, 2018-July, pp. 849–855. https://doi.org/10.24963/ijcai.2018.118.
47. Liu W, Salzmann M, Fua P (2019) ‘Context-aware crowd counting’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2019-June, pp. 5094–5103. https://doi.org/10.1109/CVPR.2019.00524
48. Lu C, Shi J, Jia J (2013) ‘Abnormal event detection at 150 FPS in MATLAB’, Proceedings of the IEEE International Conference on Computer Vision, pp. 2720–2727. https://doi.org/10.1109/ICCV.2013.338
79. Xu M et al (2019) ‘Depth information guided crowd counting for complex crowd scenes’, Pattern Recognition Letters. Elsevier B.V., 125, pp. 563–569. https://doi.org/10.1016/j.patrec.2019.02.026
80. Xu M et al (2019) ‘Depth information guided crowd counting for complex crowd scenes’. Patt Recogn Lett. Elsevier B.V., 125, pp. 563–569. https://doi.org/10.1016/j.patrec.2019.02.026
81. Yang B, … Zou L (2018) ‘Counting challenging crowds robustly using a multi-column multi-task convolutional neural network’, signal processing: image communication. Elsevier Ltd 64:118–129. https://doi.org/10.1016/j.image.2018.03.004
82. Zeng X, … Ye Y (2020) ‘DSPNet: deep scale purifier network for dense crowd counting’, expert systems with applications. Elsevier Ltd 141:112977. https://doi.org/10.1016/j.eswa.2019.112977
83. Zhang C et al (2015) ‘Cross-scene crowd counting via deep convolutional neural networks’, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, pp. 833–841. https://doi.org/10.1109/CVPR.2015.7298684
84. Zhang C et al (2015) ‘Cross-scene crowd counting via deep convolutional neural networks’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, 07-12-June, pp. 833–841. https://doi.org/10.1109/CVPR.2015.7298684
85. Zhang Y et al (2016) ‘Single-image crowd counting via multi-column convolutional neural network’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 589–597. https://doi.org/10.1109/CVPR.2016.70
86. Zhang Y et al (2016) ‘Single-image crowd counting via multi-column convolutional neural network’, proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2016-December, pp. 589–597. https://doi.org/10.1109/CVPR.2016.70
87. Zhang Y et al (2018) ‘Auxiliary learning for crowd counting via count-net’. Neuro Computing. Elsevier B.V., 273, pp. 190–198. https://doi.org/10.1016/j.neucom.2017.08.018
88. Zhou S, … Zhang Z (2016) ‘Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes’. Signal Process: Image Comm. Elsevier 47:358–368. https://doi.org/10.1016/j.image.2016.06.007
89. Zhu J, Feng F and Shen B (2018) ‘People counting and pedestrian flow statistics based on convolutional neural network and recurrent neural network’, Proceedings - 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation, YAC 2018. IEEE, pp. 993–998. doi: https://doi.org/10.1109/YAC.2018.8406516.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.