International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Mar-2022)

https://dx.doi.org/10.12785/ijcds/110181

Comprehensive Analysis Of Crowd Behavior Techniques: A Thorough Exploration

Safvan Vahora[1], Krupa Galiya[2], Harsh Sapariya[2] and Srijyaa Varshney[2]

[1]*Government Engineering College, Modasa, Gujarat, India*
[2]*Vishwakarma Government Engineering College, Chandkheda, Gujarat, India*

Received 15 Feb. 2021, Revised 12 Jan. 2022, Accepted 12 Feb. 2022, Published 31 Mar. 2022

Abstract: In recent years, one of the most important issues for public security is "Automated analysis of a crowd behavior" using surveillance videos. Vision-based crowd behavior analysis methods can be divided into three categories, namely, people counting, people tracking and identification of crowd anomalies. The deployment of such an automatic system is very complex since it requires complex algorithms to detect context-sensitive uncommon behaviors of people. With this perception, we have presented an extensive review of the different methods for crowd counting, crowd tracking, and crowd anomaly detection along with the advantages and challenges associated with crowd behavior techniques. Based on the feature descriptors used to analyze the behavior of the crowd, different methods are sub-categorized into traditional feature descriptor based approaches which use handcrafted features like PCA, HOG, SIFT, optical flow, GMM, spatiotemporal filter, etc. and the self-learned feature descriptor based approaches which use deep learning models like CNN, RNN, GD-GAN, etc. Besides, in this paper, we have also presented the performance of different methods on different datasets in each class along with details of implementation. The reviews are helpful for various applications related to human activity analysis, which mainly includes crowd behavior. The methods described here can be useful in different applications of crowd behavior, For example, anomaly detection at public places. Moreover, the review helps the beginners and developers as the benchmark and the researchers of this domain to study the challenges of the crowd behavior techniques, analyze the research gap and further enhancement in these techniques.

Keywords: Crowd Counting, Crowd Tracking, Anomaly Detection, Traditional Approach, Deep learning Approach

1. Introduction

Analysis of crowd behavior is constantly gaining importance with the continual growth of the human populace and increase in the human need to socialize. The world's total population of 7.7 billion people is rising at a level of 1.07 percent, an incredible 82 million people a year, according to the author [1]. With such a vast population, it is not unusual to see people gathered for one reason or another. With the crowd becoming such a commonplace, the subject of analyzing crowd behavior is gaining importance among the computer vision community. Analyzing the individual and collective behavior of the crowd has become a key area of research these days.

Crowded scenes can be classified as structured and unstructured scenes [2]. Structured crowds move in a common direction and the direction of movement is eternal, due to bad construction stage collapsing happens.

Large groups of people gathering together promote numerous issues regarding safety, such as violence, stampedes, terrorist attacks, theft, and harassment which often lead to health issues like injury, heat exhaustion and death. Surveillance systems are usually used for security and monitoring of crowded areas, which have a well-understood need for automation. Many kinds of research are being conducted to tackle these problems using computer vision.

The present study has been taken up with a vision to detect anomalies in crowd behavior so that preventive actions can be taken to avoid disastrous events. The previous systems intended to put control over incidents such as the repeated events of "crush and stampede" during the annual Hajj pilgrimage in Mecca, which causes massive deaths every year. For instance, upon hearing the word "bomb", the crowd is panic, on March 4th, 2010, which led to a stampede injuring several people [3]. One of the starting shoves to the field of crowd behavior analysis was given by the King's Cross underground fire in London in 1987.

Human agents ubiquitously use CCTV cameras to record and monitor scenes. But the limited availability of human resources is not feasible and sometimes not cheap. With the era of new technology, "Intelligent Video

E-mail address: safvan465@gmail.com, krupagaliya@gmail.com, sapariya.2108@gmail.com, srijyaa99@gmail.com

surveillance systems" are intended to monitor and capture the flow of the scene, Estimate crowd density in a scene, detection of abnormality in a scene. This can also help to reduce manual tasks.

One of the prominent computer vision research areas deals with understanding activities and human behavior from images and videos and is having a large impact on many real-world applications. The key techniques used are object-based and holistic approach [4]. Object-based methods, consider a crowd to be a collection of individuals. It detects and tracks each person to understand the overall behavior of the crowd. This approach becomes considerably complex to detect objects, track trajectories, and recognize activities in dense crowds having a high amount of occlusions. Alternatively, the holistic approaches treat the crowd as a global collective entity and extract features to represent the state of motion in the entire frame for analyzing it at a higher level.

The generic architecture of the crowd behaviour analysis is presented in figure 1. A method used to count the number of individuals in a crowd is Density Estimation of Crowd or providing the total count of the Crowd. Crowd counting counts people at public places to know what number of persons are around. The high-density crowd containing potential danger at a place that is usually sparsely populated.

Crowd Tracking is a method that follows the emerging path forwards from a starting point to wherever the object currently is in the crowd scene or locating a mobile object(s) over time. Crowd tracking involves the Detection of each object individually, giving them a unique id and applying algorithms that give a variety of tools for identifying the moving objects. Crowd Tracking becomes challenging due to the occlusion and distortion of the view of the camera.

Crowd Anomaly Detection is a technique that classifies abnormal or suspicious activities in the crowd scene. Abnormal events like the panic in the crowd, activities like walking or jogging differ from running vehicles in the garden, having a different moving flow than the normal moving flow of direction.

The purpose of this research, therefore, is represented by two objectives:

- Introduce a set of methodologies for analysing crowd behaviour at various levels, along with the benefits and challenges that each level presents.
- Determine the optimal strategy for presenting a model interpretation for crowd behaviour analysis models, using a comparative experimental study.

2. Literature Survey

The crowd behavior methods are mainly categorized into the traditional methods and deep learning methods based on the approach used by the model. The traditional approach is to find dissimilar features from the problem known as handcrafted features. Various operations can be performed on the input feature sets using these handcrafted features, which can be used to tackle the problem through traditional methods [5]. Artificial neural networks are the foundation of deep learning technologies. In order to improve the efficiency of training, these ANNs continuously provide learning algorithms and consistently increase the data size. A greater volume of data makes this method more efficient. The ability to learn efficient feature representations for localization and calculation of the crowd density is also one of the advantages of deep neural networks [6]. The overall categorization of crowd behavior analysis along with subcategories is presented in figure 2.

A. Traditional Approach

The various traditional approach based methods of crowd behavior analysis sub-categorized into crowd counting, crowd tracking and crowd anomaly detection. The various methods of crowd counting use regression, detection, density estimation, etc. Additionally, part-based, patch-based, optical flow, point-based, etc. the crowd tracking methods are described in this section. The supervised and unsupervised crowd anomaly detection methods are presented in brief along with the advantages and challenges of the methods.

1) Crowd Counting

The procedures for Crowd Counting and Density Estimation are characterized as Direct approaches and Indirect approaches [7]. The Direct approaches are based on object detection methods that attempt to detect every single person in a crowded area followed by counting using some classifiers. These methods lead to high complexity for a dense crowd with occlusions. In the Indirect approaches, counting techniques discover to map the low-level features to density estimation of the crowd, and therefore explicit object detection and segmentation of objects in cluttered scenes can be avoided.

Direct Approach

The Direct Approach uses two methods, Crowd Counting by Detection and Crowd Counting by Clustering. The first method uses a model/appearance of human shapes to count people after segmenting and detecting each individual in the input. In the second approach, independent motions in the crowd scene are detected for crowd counting by analyzing clusters of feature points over time on people tracked.

Subburaman et al. [7] detects the head area owing to its relatively clear visibility in a crowd. The head detector relies on a state-of-the-art cascade of increased part features. Heads are detected in the image using gradient information from the grayscale image to identify points of interest. Two separate context subtraction methods are tested, "Vibes" and "Idiap," to further decrease the search area. It is followed by placing a sub-window around interest points and based

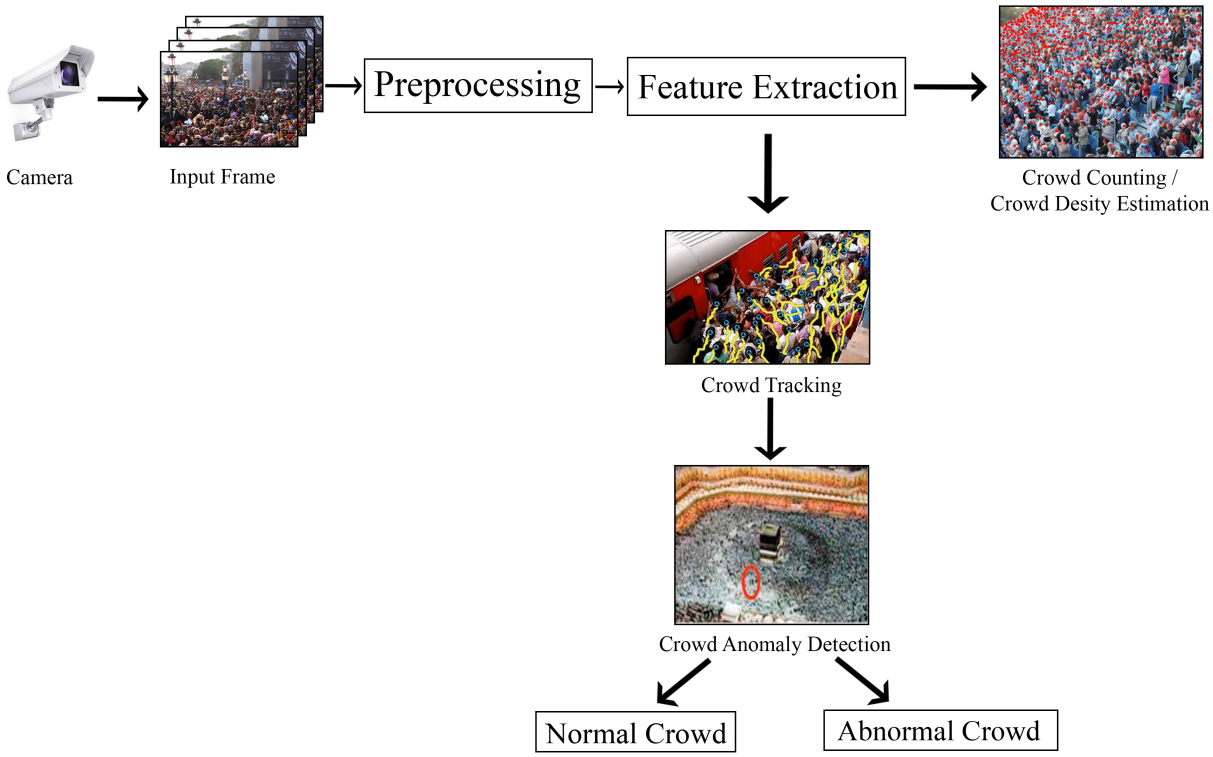


Figure 1. Generic Architecture of Crowd Behavior Analysis

on perspective calibration information, a classifier is used to classify it as head or non-head region. This technique does not work for occluded scenes. It relies on a Single-view approach which makes it difficult to analyze the crowd scene, due to a high possibility of severe occlusions.

Topkaya et al. [8] proposed a crowd counting method applying a clustering method depends on Dirichlet Process Mixture Models (DP-MMs). For each input frame, an individual detector is run to give a collection of detection areas as its output and uses abstraction, color and temporal details to define a collection of features for every detection. In the next step, the total number of people or groups is estimated using DPMMs and Gibbs sampling to cluster the detections with no limit concerning the quantity of clusters. A measure to calculate the particular range of people among every cluster is defined to infer the final count estimation. This method counts people in sparsely crowded scenes. Due to occlusion occurring in a dense crowd scene, the HOG detector cannot detect each individual in dense scenes.

Indirect Approach

The indirect counting approach is to extract the features from a group of people in an image. The Features of the foreground are extracted using the regression function. These methods have presented a way to map the low-level features to people count in the scene. They prove to be more efficient as feature detection is easier than person detection.

Gad et al. [9] recommended a new automatic density of the crowd estimation method for a single camera to overcome the issue of linearity and enhance the accuracy of counting prediction. A hybridization of features such as edges, texture, segmented regions, properties and SIFT feature vectors are extracted from the segmented foreground regions to resolve the linearity issue. The problem of perspective distortion is solved using these features in normalized form. Crowd count is then predicted accurately by training five regression models such as random projection forest (RPF), random forest (RF), Gaussian process regression (GPR), Least Absolute Shrinkage and Selection Operator (LASSO) and K-nearest neighbors (KNN) using normalized features which help to detect overcrowded situations.

A single feature extraction or detection system is unable to predict accurate counting in extremely dense crowds due to poor resolution, significant occlusion, subtle shading, and viewpoint. Idrees et al.[10] suggested a system that would estimate the number of persons in a single frame

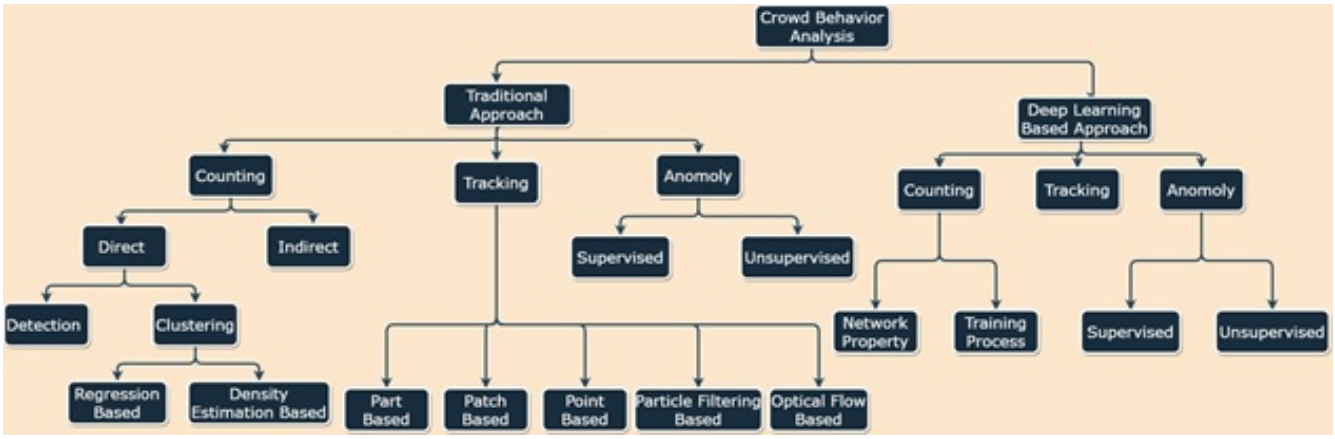


Figure 2. Categorization of Crowd Behavior Methods

of an extremely dense crowd using multiple sources of knowledge. The multiple sources include low-confidence head detection, frequency-domain analysis, and repetition of texture elements (using SIFT) to estimate the count in the image region. In localized patches, individual counts are computed, which are then controlled universally to find an approximation of the count for the whole area. Using Markov Random Field, a global consistency restriction is used on counts. This provides for inequality in local and across-scale community counts.

Pham et al. [11] proposed a patch-based method to estimate the crowd density in a crowd scene. This technique uses a random forest model to map the patch characteristics to the relative positions of all objects within each image patch non-linearly. This allows to produce a map of patch density using Gaussian kernels. In a coarse-to-fine fashion, two split node layers construct the forest. In addition, a prior crowdedness and an efficient method of forest reduction are recommended to increase the speed and accuracy of the evaluation. The advantage of COUNT forest is that much less memory is needed to create and maintain the forest relative to regression forest models with densely structured labels.

2) Crowd Tracking

Detecting and tracking people in crowded scenes is a crucial part of the Crowd Behavior Analysis Applications. The key challenges to crowd monitoring reside in three areas: (1) partial or complete occlusions due to repeated interactions within multiple objects; (2) a wide group of people's identical appearance and (3) significant variation in appearance caused by the perspective distortion of camera views.

Shu et al. proposed an effective technique to track multiple people by employing a part-based model and handling occlusions [12]. This approach captures the articulations of human bodies by studying part-based person-specific SVM classifiers in a dynamically changing appearance and context. With the part-based model, detection performance can be improved by selecting a subset of parts to maximize the detection probability for a crowd scene. Occlusions can be handled dynamically during tracking by distributing the score of the learned person classifier among its corresponding parts, thus preventing the degradation of performance of the classifiers and allowing them to detect and predict partial occlusions.

Ali et al. proposed an automatic algorithm for multiple-people detection in high-density crowds having extreme occlusion [13]. The state-of-art methods do not apply to scenes having heavy crowds where most people are in motion and are partially or fully occluded. This problem is dealt with by employing a single framework integrating human detection and tracing, and introducing a confirmation-by-classification mechanism that combines tracking and detection, tracks people through occlusions, and removes the false positive traces. A Viola and Jones AdaBoost detection cascade is used along with a specific filter to track and histograms of color to encode the appearance. The 3D head plane information can be utilized to preserve high detection rates alongside improving accuracy, reducing false-positive rates.

To further analyze the patterns in crowd motion, trajectories of a set of elements are generated in crowd tracking. Zhu et al. recommended a method for the identification and monitoring of distinctive and secure patches using dynamic hierarchical tree structures to track crowd motion [14]. For that, in one model, low-level key feature monitoring, mid-level patch tracking and high-level group transformation are combined. The KLT tracker is used to track the key points. Keypoint tracking stops encountering ambiguity arising due to occlusions or other factors. Thus, this technique provides accurate motion information for a short duration. For partial occlusion or appearance changes, patch analysis is more efficient and group evolution directs the updating of group structures.

Bera et al. proposed a unique approach to measure each individual's trajectory in crowd scenes having a moderate

density in the real-time environment [15]. Multiple confidence metrics are employed for particle filtering in which the count of particles assigned to each person is modified dynamically. The confidence metrics are estimated using a non-linear parametric multi-agent motion method Reciprocal Velocity Barriers (RVO), to consider the reactive activity of pedestrians into consideration in a dense environment. The inability of the RVOs to consider the physiological and psychological traits of pedestrians is a limitation. A similar sensitivity to gender and density is used to model all pedestrians.

Direkoglu et al. adopted a novel method of optical flow for the extraction of event characteristics. Panic people start running around in an abnormal situation, this not only increases the optical flow magnitude but also measures the angle difference between the optical flow vectors at each movement of the pixel position in the consecutive frames [16]. Initially, the angle difference between optical flow vectors is computed in the current frame and the previous frame at each pixel location. The findings, however, are influenced by minor measurements of noisy optical flow and their angle variations. The angle difference is multiplied by the optical flow amplitude in the current frame, in order to reduce the chances of these noisy measurements.

Ramakrishnan et al. provided a precise algorithm to track Object Boundary Feature Points [17]. The CoMaL feature point is tracked for Object Tracking, which is detected and matched on object boundaries. To achieve this, the line-level segment correlated with the corner is tracked using MSER identification and matched to level lines obtained in the following frame. To weed out not good matches, the Hierarchical Chamfer Matching Algorithm is used, and the selected matches are then checked using Part SSD matching to get the best match. This method is more efficient than the KLT technique.

3) Crowd Anomaly Detection

Marsden et al. put forward a crowd anomaly detection technique that yields a low-dimensional scene-level descriptor, holistic crowd features that can be easily interpreted to analyze tracklet information [18]. Two features from the literature are combined in this low-dimensional descriptor: crowd collectiveness [1] and crowd conflict [2], along with a couple of additional crowd attributes: the average movement rate and a new crowd density formulation. Two different solutions for anomaly detection are put forward with the use of these features. A Gaussian Mixture Model (GMM) is used for detecting outliers with the availability of only normal training data. Otherwise, Support Vector Machine (SVM) is employed for binary classification in cases where data are available for both normal and abnormal training.

Mousavi et al. proposed tracklet based crowd anomaly detection methods. A new spatiotemporal feature is estimated using Histogram of Oriented Tracklet (HOT) that creates the analyzed sequence and uses the found tracklets to represent the dominant motions over short durations within the region in view [19]. For classification, there are two approaches. i) Latent Dirichlet Allocation (LDA) used with the availability of only normal training data; ii) Support Vector Machines (SVM) used when abnormal training data is also available.

Bera et al. suggested an algorithm to detect anomalies in real-time videos of low to medium density crowds [20]. In this approach, online computer vision monitoring algorithms, non-linear crowd simulation pedestrian movement models, and Bayesian learning methods are combined to dynamically classify trajectory-level activities for each agent in the frame. First, a real-time algorithm to track multiple people is used to retrieve the trajectory of each agent from the sequence of images. The anomaly was detected by computing the trajectory behavior for each individual using a Bayesian inference technique. Wang et al. suggested a method for Spatio-temporal recognition of crowd anomalies [21]. Low-level statistical elements are deployed and complicated processes of machine learning and recognition are mitigated to design a real-time application algorithm. The algorithm starts with transforming a video clip into the STV structure. With the application of average flow field models derived from live video streams, highly complex areas from crowded scenes can be easily recognized, which can guide the further study by the auto-size selection, sampling STV slices for locations and directions. The Gaussian approximation model analyzes the sampled STT texture patterns to discriminate crowd "normalities" from abnormal behaviors. The local changes and global similarities of the described video volumes are summarized by the statistical STT features.

Rabiee et al. proposed more than one novel descriptor for the detection and localization of unusual behavior [22]. The first descriptor for the detection of abnormalities in crowded scenes is the simplified Histogram of Oriented Tracklet (sHOT) model. Multiple descriptors are often combined to reach a descriptor having a single feature containing both positioning and magnitude data. A novel abnormal behavior descriptor is introduced by using a combination of sHOT and Dense Optical Flow, which is used to localize the abnormalities present in the crowd. To classify behavior, single class SVM has been employed.

Chaker et al. suggested using a social network model (an unsupervised method) for detecting and localizing anomalies in a crowded place [23]. The unsupervised approach analyses data on crowd movement in a scene to extract dense tracklets, followed by the construction of spatiotemporal cuboids. The local behavior is then modeled based on exclusive attributes of tracklets clustered the object in each cuboid to develop a local social network. The global social network is modified progressively for each subsequent time window, by adopting its local social networks and global social network of the previous window. Social network models are used to identify normal and abnormal behavior.

B. Deep Learning based Approach

Deep Learning is a subfield of Machine Learning that learns to represent the input as a hierarchical nesting of perceptions exhibiting great power and high flexibility. Each new concept in the hierarchy is defined concerning low-level simpler concepts and is computed as more abstract representations in terms of less abstract ones [24]. Deep Learning uses hidden layer architecture to incrementally learn categories. Extracting information from high-dimensional data is one of the key purposes of using deep learning techniques. The different Deep Learning based methods of crowd behavior analysis sub-categorized into crowd counting, crowd tracking and identification of crowd anomalies.

1) Crowd Counting

The objective of crowd density estimation or crowd counting techniques is to calculate the amount of people in crowded scenes. There is a vast collection of literature, classical and involving computer vision, on the challenges of crowd counting. The key challenges are occlusion, non-uniform distribution, perspective distortion, clutter, scale variation and complex enlightenment. Convolutional Neural Networks (CNNs) such as switching networks [25], multi-column CNNs [26], scale-aware regression models [27] are being used on a wide scale to conquer the crowd counting problem.

Basic CNNs are simple and can be trained at less computational cost, but result in lower accuracy. To achieve a significant performance boost, different scale-aware models and context-aware models are combined. However, it results in high computational complexity.

Zhang et al. suggested a model for cross-scene counting by mapping image frames to crowd counts and adapting the collected mapping to new target scenes [28]. The network is initially trained by alternating on two related objective functions for crowd counting and density estimation. Alternatively, these goal functions are optimized to produce a stronger local maximum. Training samples similar to the target scene are used to make the structure adaptive to new scenes. The evaluation of the network is done for single-scene and cross-scene crowd counting. Perspective maps for both training and testing scenes are required, which are not available freely.

The recent methods are focused on scale-awareness owing to large density variations in different images. Zhang et al. proposed architecture for images having arbitrary crowd density and perspective, called multicolumn CNN (MCNN) [26]. This model uses varying kernel sizes in the network to capture varying densities during training. It makes use of a vital characteristic of the images of high-density crowd scenes that head size is proportional to the separation between the centers of two adjacent individuals. This requires all regressors in the multi-column network to be trained on all the input patches.

By utilizing crowd density variations within an image, Sam et al. claim improved results by modeling regressors using a specific collection of learning patches [25]. The proposed approach called switch-CNN intelligently chooses an optimum regressor for a specific input patch. In several CNN regressors, the underlying functional and structural differences are used to resolve the large scale and viewpoint variations using a differential training regime. The drawback of the approach is the balancing of selection among multiple columns.

Liu et al. suggested a self-supervised technique that significantly improves performance using unlabeled crowd images for training [29]. The proposed technique generates a ranking of sub-images leveraged to train a network to estimate people count in relation to another image. A network is trained to compare images and rank them based on people count in images. The current practice to exploit self-supervised learning is training a self-supervised task followed by fine-tuning the resulting network on the testing part having minimum data.

A deep spatial regression model (DSRM) for arbitrary resolution and arbitrary viewpoint relying on CNN and LSTM has been proposed by Yao et al. [30]. Initially, the images are put into a pre-trained CNN for extracting a set of high-level features. Then, local counts are obtained by regressing using an LSTM structure, by considering the spatial information. To obtain the final global count and local patches are summed. Here LSTM is used to learn the spatial constraint relation of local counts in adjacent regions.

Olmschenk et al. compare the frequently accepted crowd density map labeling scheme to train deep neural networks with the less efficient alternative inverse k-nearest neighbor(ikNN) maps, although the directly current traditional networks show the supremacy of the latter [31]. A novel structure MUD-ikNN is provided, that utilizes multi-scale upsampling with transposed convolutions to use provided ikNN labels. Besides, this upsampling within ikNN maps provides better performance to the current traditional techniques.

2) Crowd Tracking

People Tracking is a learning problem that deals with the estimation of location and object scale, previous individual location, size, current and preceding picture frames. False-positive matches contributing to the erroneous association of tracks [32] is the biggest challenge to conventional methods for learning and/or tracking-by-detection.

Fan et al. designed a CNN-based tracking approach having shift-variant architecture [32]. Discriminant spatial and temporal characteristics are extracted from the discriminative model for specific object-level tracking, which learns features from a parametric feature group via extensive degrees of freedom. Multiple pathways in CNN's to the better fusion of local and global information are presented.

For combining local and global information better, several channels are implemented on CNN. Besides, CNNs are used to use the precise location of certain key points for size approximation. For offline training, Standard stochastic gradient descent (SGD) is utilized. A trained model is later fixed while tracking.

Chen et al. introduced an online multi-people tracking framework, the unreliable detection is tackled by picking candidates collectively from detection and track output [33]. The scoring function is devised with the aid of an efficient R-FCN for candidate selection, sharing computations on the whole image. ReID(re-identification) features are used to improve the identification ability while handling intra-category occlusions to associate data. Trained using a data-driven methodology, the use of ReID features greatly outperforms conventional hand-crafted applications.

Redmon et al. presented an advanced object detection methodology that also works fine with a unified object detection model, object tracking named YOLO (You Only Look Once) [34]. The architecture is easy to formulate and full images can be directly used for training. The whole system is jointly trained on a loss function that correlates directly to the detection output. Fast YOLO has proved to be the fastest general-purpose object detection approach in the literature that pushes state-of-the-art object detection in a real-world environment.

Gordon et al. proposed a real-time, recurrent, regression-based tracker, or Re3. The appearance model is simultaneously tracked and updated using a single forward pass [35]. This tracker incorporates temporal information into its model. A new, compelling method of tracking is offered by recurrent models owing to their offline learning ability from numerous examples and too quick online updating while tracking a specific object.

Fernando et al. proposed a crowd tracking by prediction approach, based on lightweight sequential Generative Adversarial Network architecture for person localization [36]. This is a robust lightweight algorithm used for multi-person tracking problems for data association, leveraging trajectory prediction. The proposed methodology expands the recent advances in the estimation of pedestrian trajectories and offers a novel scheme for trajectory-based data association.

Carraro et al. proposed a scheme to use calibrated RGB Depth camera networks to predict and track the 3D poses of multiple individuals [37]. A central node is used to gauge the multi-view 3D pose of each individual that works on the single-view results from each network camera. To estimate 2D poses, a CNN is used which is extended to #D utilizing sensor depth for computing each single-view outcome.

3) Crowd Anomaly Detection

The identification of crowd activity can be done as normal or abnormal. The unusual behavior of people who break public security is known as abnormal behavior. Abnormal events are also classified as the local abnormal event (LAE) and global abnormal events (GAE) for video surveillance purposes [38][39].

Zhou et al. put forward a technique called spatiotemporal CNN to analyze crowded video frames to detect and locate anomalous activities [40]. In the proposed spatiotemporal CNN, the spatial-temporal features are automatically extracted and dynamic regions are localized in crowded scenes. For spatial-temporal volumes of moving pixels, the algorithm performs spatial-temporal convolutions for noise sensitivity.

Marsden et al. suggested a novel deep residual network ResnetCrowd which is based on Resnet18 architecture [41]. The proposed multitask CNN model was learned for three different tasks such as crowd counting, crowd density level classification and crowd violence behavior detection. The network outputs the number of persons in the image, a heatmap to estimate pedestrians' density and a binary classification label to present the detection of violence in the image. This model uses only images as an input to the system.

Ravanbakhsh et al. presented a measure-based approach that allows integration of semantic information with low-level optical-flow [42], adding a minimum cost for training [43]. They showed effective detection of local anomalies by tracking the changes in CNN features over time. This is a three-step process: 1) Using a sequence of input frames to extract CNN-based binary maps; 2) Utilizing the extracted CNN-binary maps for measuring temporal CNN pattern (TCP); 3) Finding refined segments of motion by fusing the TCP parameters with low-level motion (the optical-flow) characteristics.

Fully connected CNNs (FCNs) were used by Sabokrou et al. to extract the discriminatory characteristics of video regions [44]. As a Gaussian distribution, they modeled a standard event and labeled a test region that varied as an anomaly from the normal reference model.

A modified 3D CNN based violent video detection model [45] is introduced by Song et al. In this model, the input frames sequence is produced by using a random sampling method. A set of consecutive random frames between two keyframes are identified and provided as an input to the 3D CNN model. The spatiotemporal features extracted from the 3D CNN model are fed into the violence detection classifier to provide an overall label.

Sabokrou et al. proposed a technique based on cubic-patches, characterized by a classifier cascade, which utilizes a state-of-the-art approach for feature learning [46]. There are two key stages of the cascade of classifiers. The first employs a lighter and deeper three dimensional auto-encoder to detect "many" regular cubic patches in the early stages. As the first step, this deep network runs on small cubic patches before selectively resizing the remaining candidates of

interest and using a more robust and deeper 3-dimensional CNN to examine those at stage two. Fan et al. proposed a Gaussian Mixture Variational Autoencoder-based approach, using deep learning to learn function representations of regular samples as a GMM [47]. For the encoder-decoder structure, a Fully Convolutional Network (FCN) with no fully connected layer is used to maintain the corresponding spatial location among the input frame and the output function map. To integrate the irregularities of presence and motion, a two-stream network architecture is used.

Ravanbakhsh et al. used networks trained using normal frames, called Generative Adversarial Nets (GANs) [48]. Only normal data is used to train GANs, which are incapable to produce abnormalities. During testing, perceptions and movement information characterization recreated by GANs are compared to actual data. Local differences are measured to detect abnormal areas.

Luo et al. recommended an Auto-Encoder framework based on convolutional LSTM to detect anomalies [49]. The content of each frame can be well characterized by using CNN to encode each frame, and ConvLSTM is employed to characterize the motion information. Meanwhile, ConvLSTM preserves spatial information that is useful for reconstructing current and previous frames. Wang et al. used a self-supervised learning-based approach, S2-VAE to recognize abnormal local and global activities [50]. The algorithm proposes 2 networks: SF-VAE and SCVAE. The first level consists of a shallow generational network (SF-VAE) for efficient data definition. At this stage, some unnecessary normal samples are quickly filtered out. Then, in order to accurately locate the abnormal event using a deep generative network, SC-VAE is employed in the second stage.

A novel bidirectional LSTM [51] model for real-time anomaly detection is suggested by Dinesh et al. A huge amount of input real-time video streams are processed by the Spark engine. The HOG feature of a video frame is extracted and provided to three different models such as human part model, the anomaly model and the negative model. The three models are fed into the temporal bidirectional LSTM network, which can access information both forward and backward and provide an anomaly class label as an output.

Hou et al. proposed a dictionary selection model, an unsupervised anomaly detection method [52]. A concise feature space is trained in an unsupervised manner by using a stacked autoencoder network used for feature representation. The forward-backward greedy approach is adopted for model optimization, improving the dictionary collection model and the sparse reconstruction model.

An ensemble of the several CNN model for crowd anomaly detection has shown promising results over a single CNN model [53]. The proposed model uses three different pre-trained CNN models namely AlexNet, VGGNet and GoogleNet. The fine-tuned extracted feature of each CNN model is aggregated to form a concluding feature vector. Moreover, a final feature vector fed into an aggregation of ensemble classifier softmax classifier, linear SVM, quadratic SVM and cubic SVM classifier to provide the overall classification label.

3. COMPARATIVE ANALYSIS AND CHALLENGES

The comparative analysis of the state-of-the-art methods for crowd behavior analysis methods presented in this section. The advantages and challenges of each method of crowd counting, crowd tracking and crowd anomaly detection along with the approach used by the method are summarized. Table I presents the comparative analysis of the crowd counting methods. The advantages and challenges of each method approach wise of crowd tracking methods are presented in table II. At last, the comparative analysis of the crowd anomaly detection methods presented in table III.

4. EXPERIMENTS

The performance of the state-of-the-art methods of behavior analysis of the crowd is carried out by different evaluation parameters. This section also presents the experimental setup used for crowd behavior analysis used by the state-of-the-art techniques. The state-of-the-art methods of crowd behavior analysis with its evaluation metrics on different benchmark datasets are compared herewith. The crowd counting literature methods use MAE (Mean Absolute Error) and MSE (Mean Squared Error) parameters for performance measurement purposes. The MAE measure indicates the accuracy of the crowd estimation algorithm while the MSE parameter specifies the effectiveness of the estimation

TABLE I. Comparative Analysis of Crowd Counting Methodology

Method	Technique	Advantages	Challenges
Subburaman et al. [7]	Traditional	The foreground region is obtained using background subtraction techniques such as "Vibes" and "Idiap".	Relies on a Single-view Approach which presents difficulties in the analysis of the crowd scene, due to highly possible severe occlusion.
DPMM + HOG. Topkaya et al. [8]	Traditional	Using a generalized person detector, the clustering approach has been improved. This method can handle an unknown number of clusters.	Affected by High-density crowd and background clutter. Time-consuming and resource consuming depend on tracking techniques.
Gad et al. [9]	Traditional	Overcoming the linearity problem by perspective normalization. High error rates due to random samples selection can be reduced by sample selection using Cross-validation.	Robustness lost with cross-dataset evaluation. Need to retrain the model on each dataset before use.
FHSc + MRF Idrees et al. [10]	Traditional	Fourier analysis on various scales in local neighborhoods to prevent the issue of irregularity in the presumed textures resulting from dense crowd images. It depends on multiple estimation sources, such as low-confidence head recognition, replication of texture features (using SIFT) and frequency interpretation.	Inadequate with crowd scenes with scale and perspective distortions.
Zhang et al. [28]	Deep Learning	The network, without any extra-label knowledge, is adapted to new target scenes.	It needs testing scenes as well as perspective maps of the test scene, which are not readily accessible.
MCNN [26]	Deep Learning	Good generalizability, from an individual image with arbitrary crowd density and arbitrary perspective, will reliably determine the crowd count, allowing the input image to be of arbitrary dimensions or resolution.	Need to train all multi-column network regressors on all the input patches. Time Consuming.
Switch-CNN [25]	Deep Learning	Choose an optimal regressor suitable for a particular input patch., Switch-CNN is resilient on a wide scale. Effective to facing large differences of size and perspective.	The selection amongst the multiple columns is not balanced.
L2R [29]	Deep Learning	Faster to train, uses no alongside details, supports a fast, scale-aware, and multi-task inference. By automatically creating rankings from them, it allows the use of comprehensively accessible training information through the Internet	It is guaranteed that any sub-image of a crowded image of the scene has the same or less people as the super-image, using observation.
MUD-ikNN [31]	Deep Learning	A network is more transferable and at any point in the network can be expanded to any CNN structure. If any particular regression module shows more precise results, its results may be treated as being more important to the final inference individually.	
DecideNet [54]	Deep Learning	Adaptive to varying crowd densities. DecideNet automatically switches between the two modes - regression mode and detection mod depending on the real density at a position in the frame.	Performance is compromised due to training on partially correct ground truth.

TABLE II. Comparative Analysis of Crowd Tracking Methodology

Method	Technique	Advantages	Challenges
Shu et al. [12]	Traditional	Correctly associate detections and tracking under partial occlusions and appearance changes. Better description of the articulated body is achieved through combination of parts, which leads to better detection.	Sensitive to scale and density of crowd. Performance degrades in high occlusions in highly dense crowd scenes.
Ali et al. [13]	Traditional	Detects and tracks using the automatically extracted three dimensional head plane data. High accuracy thus maintaining high detection rates with decreased false-positive rates. Works well for dense crowds with high occlusions.	Not suitable for long-term person tracking as any head trajectory is eliminated for a person near to the exit-zone (border of the image) and the motion model predicts the position outside the frame.
Zhu et al. [14]	Traditional	The static single-tree structure is outperformed by the dynamic hierarchical tree structure. Combines tracking of low-level feature points, mid-level patch tracking, and high-level group evolution.	Requires manual initialization of target(s) to track.
MLPF-RVO Bera et al. [15]	Traditional	At each time-interval step, k calculates adaptively for each pedestrian. Offers a good balance among accuracy and velocity.	RVOs do not consider psychological and physiological pedestrian characteristics. All individuals are modeled with a similar sensitivity towards gender and density. The technique does not consider heterogeneous features.
CoMal Tracking Ramakrishnan et al. [17]	Traditional	Superior and resilient performance at boundary regions. On the MSER boundaries found in the next image but not on the edge map, a match is carried out, which renders the process very stable.	Can handle partial occlusions, but fails to effectively accommodate high occlusions in the highly-dense crowd scenes.
Fan et al. [32]	Deep Learning	The method of scale estimation is focused on localizing key points that are independent of the object.	CNN model is not designed for distractors of the same object type to accommodate complete and long-term occlusions. For all learning-based approaches, this restriction usually exists.
Chen et al. [33]	Deep Learning	Tackles unreliable detection and intra-category occlusion. Uses deeply learned person ReID features along with spatial data.	Requires explicit encoding of spatial data into the score maps. Uses a Kalman filter for new location prediction, not appropriate for long-term tracking.
Fernando et al. [36]	Deep Learning	Overcomes occlusions and noisy detections in a multi-person environment. Uses data association scheme based on trajectory prediction to eliminate computationally expensive person re-identification.	
Carraro et al. [37]	Deep Learning	Marker-less, multiperson, independent of background and does not presume the presence and initial pose of individuals. Provides a real-time solution.	Requires exclusive Camera Network setup and cannot be integrated with existing security cameras at most places.

TABLE III. Comparative Analysis of Crowd Anomaly Detection Methodology

Method	Technique	Advantages	Challenges
Bera et al. [20]	Traditional	Capture pedestrians' constantly shifting movement patterns. Real-time detection.	In a very dense crowd, it may not work efficiently.
Chaker et al. [23]	Traditional	Free from various application videos for the identification of abnormal behaviors. Localizes the detected anomaly. Can be used in online mode via an incremental update mechanism.	Requires very large amount of training data.
Wang et al. [21]	Traditional	Better accuracy, increased efficiency, and versatility. When recognizing and discovering crowd anomalies, the combination of wavelet, moment and boundary of the texture feature space will improve sensitivity. Real-time detection.	Performance superiority and robustness are lost for low-density crowd scenes as individual behaviour takes more weight. Lacks adaptive feature selection and pattern recognition.
Marsden et al. [18]	Traditional	Present a low-dimensional scenic descriptor with easily interpretable, comprehensive features of the crowd.	Poor performance for data with high scale variations in crowd density.
sHOT Rabiee et al. [22]	Traditional	Simplified Centered Tracklets Histogram (sHOT), that is much smoother and has improved outcomes than other state-of-the-art systems. Localizes the detected anomaly. Works well with variable crowd densities.	Tracklets are sensitive to frame-rate and camera position. Therefore, shows varying results for different configurations. Cannot be used for very low-density crowd scenes due to the use of crowd-based anomaly detection.
Sabokrou et al. [44]	Deep Learning	To reduce computational complexities, CNN isn't really scratch-trained, rather simply fine-tuned. High processing speed, gives acceptable results in real-time.	High false-positive rates for crowds with high density and people walking in random directions.
Sabokrou et al. [46]	Deep Learning	Effective in run-time as well as accuracy. Using a smaller, deep network that performs on smaller patches, complex patches are found.	Holistic approach for anomaly detection, no separate module for feature extraction and classification makes results in computationally expensive training. Unable to correctly classify in crowd scenes with multiple anomalies.
Ravanbakhsh et al. [43]	Deep Learning	The benefits of the generative paradigm present that during the training time only normal samples are required. It is based on calculating the difference from the usual pattern learned when identifying what is abnormal.	Disregards the location-dependent nature of anomalies in the scene. Results assume that every test frame sequence has at least one normal and one abnormal frame, as per-video normalization is used.
Hou et al. [52]	Deep Learning	By using an adaptive greedy model based on 0 norm constraints, which is more robust, accurate and sparse in practice, an unsupervised anomaly detection method is scheduled. Optimize this non-convex issue of optimization.	Computationally expensive, in real-time, it cannot be used for anomaly detection. Requires a large amount of training data with anomalous crowd scenes.
Zhou et al. [40]	Deep Learning	The model extracts features from Spatio-temporal dimensions, which are used to encode motion and position information in the frame sequence.	Requires at least some anomalous frames in input video, else results in increased false-positive classification. Works only with static camera inputs.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - C_i^{GT}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i - C_i^{GT}|} \quad (2)$$

where, C_i is a Predicted count, C_i^{GT} is the ground truth count and N is the total number of images. The comparison of state-of-the-art crowd counting methods on various benchmark datasets presented in table IV.

In Subburaman et al. experiment was carried out in the Visual Studio 2008 platform, using a C++ with the open-source library for computer vision which is openCV [7]. Gad et al. implemented the model on a machine with 16 GB DDR3L RAM and Intel Core i7 2.50 GHz processor [9] using scikit-learn.

As a performance metrics measurement parameter, the crowd tracking literature methods mainly used MotA (multi-object tracking accuracy) parameter. Here,

$$MOTA = 1 - \frac{\sum_t FN_t + FT_t + IDS_t}{\sum_t GT_t} \quad (3)$$

where, FN_t is false negative count (missed targets), FP_t is false positive count (ghost trajectories), IDS_t is the number of identity switches at time t. In case the IoU (Intersection over Union) with the ground truth is inferior to the given threshold, the target is considered. Table V presents the comparison of state-of-the-art crowd tracking methods on different benchmark datasets.

In Shu et al. experiments are accomplished using MATLAB R2012b with 16 GB RAM on a 3.20 GHz Intel Core i5 processor machine [12]. Zhu et al. used Matlab tool and the tracking speed on the provided dataset is between 7 fps to 14 fps using Intel Core 2 Duo processor of CPU 3.0GHz [14]. Chen et al. used GTX1080Ti GPU for experiment purposes [33].

The performance measure metric used by the majority of the state-of-the-art methods of crowd anomaly detection are ACC, EER and AUC. The different methods have used either one or two parameters. Here, the AUC is Area Under the ROC Curve, which is the area under the ROC curve (receiver operating characteristic curve). EER is Equal Error Rate, it corresponds to the error rate of a method when the false positive and false negative rates are similar. Higher AUC and lower EER are better. ACC is accuracy, it is the selection of predictions when the model got right. The comparison of state-of-the-art crowd anomaly detection methods on different benchmark datasets described in table VI.

Xu et. al [55] have used NVIDIA Quadro K4000 graphics card and a machine with 32 GB RAM along with a multi-core 2.1 GHz CPU to implement the proposed AMDN model. The implementation of the TCP model is carried out by using Matlab 2018 software on a Windows 7 operating system with Intel Core i7-6700 CPU, 2.60GHz and 16GB RAM [43]. Zhou et al. have used a 2.80GHz Genuine CPU, 128GB RAM, and Ubuntu 64-bit operating system for the experiment purpose [40].

5. CONCLUSION

This paper describes the various methods of crowd behavior analysis which can be categorized into three key domains such as crowd counting, tracking and crowd anomaly detection. Moreover, each category method is also analyzed based on the feature extraction methods used in the model such as traditional approach methods and deep learning approach methods. The challenges and advantages of the methods in each category are presented which will be helpful for the application developers to choose the appropriate method as per the requirements. Moreover, the presented category-wise result analysis on various benchmark datasets along with implementation details will be helpful to beginners as well as researchers of this domain. The beginners will get a quick start in this domain. The researchers will get research design and the new directions in this domain to gain performance and showing better results. The systematic review of crowd behavior techniques also presents that deep learning methods are showing better performance in each category on the majority of the benchmark datasets.

TABLE IV. Performance Comparison of state-of-the-art Crowd Counting Methods on benchmark datasets

Method	Technique	Dataset	MAE	MSE
Subburaman et al. [7]	Traditional	PETS2009 [56]	5.95	
Topkaya et al. [8]	Traditional	PETS2009	1.47	
GPR Gad et al. [9]	Traditional	UCSD [57]	0.69	1.13
FHSc+MRF Idrees et al. [10]	Traditional	UCSD_CC_50 [10]	419.5	487.1
Pham et al. [11]	Traditional	UCSD	1.61	4.4
		MALL	2.5	10
Zhang et al. [28]	Deep Learning	UCF_CC_50	467	498.5
		UCSD	1.6	3.31
		WorldExpo'10 [28]	12.9	
MCNN [26]	Deep Learning	ShanghaiTech Part A [26]	110.2	173.2
		ShanghaiTech Part B [26]	26.4	41.3
		UCF_CC_50	377.6	509.1
		UCSD	1.07	1.35
Switch-CNN [25]	Deep Learning	ShanghaiTech Part A	90.4	135
		ShanghaiTech Part B	21.6	33.4
		UCF_CC_50	318.1	439.2
		WorlExpo'10	9.4	
		UCSD	1.62	2.1
DecideNet [54]	Deep Learning	Mall dataset	1.52	1.9
		ShanghaiTech Part B	21.53	31.98
		WorldExpo'10	9.23	
L2R (Query-by-example) [29]	Deep Learning	ShanghaiTech Part A	72	106.6
		ShanghaiTech Part B	14.4	23.8
		UCF_CC_50	291.5	397.6
DSRM [30]	Deep Learning	ShanghaiTech Part A	74.4	114.7
		ShanghaiTech Part B	15.2	29
		UCF_CC_50	283	372
		AHU-CROWD	81	129
		WorldExpo'10	8.4	
MUD-i1NN [31]	Deep Learning	UCF-QNRF	104	172
		ShanghaiTech Part A	70.4	112.7
		ShanghaiTech Part B	14.4	20
		UCF-CC-50	237.76	305.69
		WorldExpo'10	9.4	
TEDnet [58]	Deep Learning	UCF-QNRF	113	188
		ShanghaiTech Part A	64.2	109.1
		ShanghaiTech Part B	8.2	12.8
		UCF-CC-50	249.4	354.5

TABLE V. Performance Comparison of state-of-the-art Crowd Tracking Methods on benchmark datasets

Method	Technique	Dataset	MotA (%)
Shu et al.[12]	Traditional	Town Center	72.9
DHT +AP Zhu et al. [14]	Traditional	Traffic	99
		Crowds	90
		Marathon	91
		Split	86
		Merge	84
		Cross	90
MLPF—RVO Bera et al. [15]	Traditional	IITF–1	69
		NPLACE–1	71
		NPLACE–2	73
		NPLACE–3	64
		NDLS–2	72
Chari et al. [59]	Traditional	PETS2009	85.5 (M)
Chen et al. [33]	Deep Learning	MOT16	35.7
APRCNN [60]	Deep Learning	TUD	61.3
		PETS	38.9
SiameseCNN [61]	Deep Learning	TUD	73.7
		PETS	34.5

TABLE VI. Performance Comparison of state-of-the-art Crowd Anomaly Detection Methods on benchmark datasets

Method	Technique	Dataset	Acc (%)	EER (%)	AUC (%)
Bera et al. [20]	Traditional	UCSD	85	20	
		ARENA	76		
SNM Chaker et al. [59]	Traditional	UCSD			86.7
HOT–BW Mousavi et al. [19]	Traditional	UCSD	82.3		
Marsdenet et al. [18] Single scene	Traditional	UMN [62]			92.9
Marsdenet et al. [18] Cross scene		UMN			86.9
sHOT Rabiee et al. [22]	Traditional	UMN			99.6
		Violent Flow Dataset	82.2		
Direkoglu et al. [16]	Traditional	UMN	96.46		
		PETS2009	96.72		
S 2 -VAE [50]	Deep learning	UCSD PED1		14.3	94.25
		Avenue [63]			87.6
AMDN (double fusion) [55]	Deep learning	UCSD PED1 (frame level)		16	92.1
		UCSD PED2 (frame level)		17	90.8
TCP [43]	Deep learning	UCSD PED1 (frame level)		8	95.7
		UCSD PED2 (frame level)		18	88.4
		UMN			98.8
Zhou et al. [40]	Deep learning	UMN		99.63	
Aggregation of ensemble [53]	Deep learning	UCSD PED1			93.2
		UCSD PED2			92.1
		Avenue			92.7

References

[1] worldometers.info, "worldometers.info," http://www.worldometers.info/world-population/, 2020, accessed: 2020-05-30.

[2] N. Sjarif, S. Shamsuddin, and S. Hashim, "Detection of abnormal behaviors in crowd scene: a review," *Int. J. Adv. Soft Comput. Appl*, vol. 4, no. 1, pp. 1–33, 2012.

[3] History, "History of stampedes," https://thoughtcatalog.com/brandon-gorrell/2010/08/a-history-of-human-stampedes/, 2020, accessed: 2020-05-30.

[4] K. Y. Joshi and S. A. Vohra, "Crowd behaviour analysis," *International Journal of Science and Research, ISSN*, pp. 2319–7064, 2014.

[5] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *The Visual Computer*, vol. 35, no. 5, pp. 753–776, mar 2018. [Online]. Available: https://doi.org/10.1007%2Fs00371-018-1499-5

[6] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks—counting, detection, and tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1408–1422, may 2019. [Online]. Available: https://doi.org/10.1109%2Ftcsvt.2018.2837153

[7] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. IEEE, sep 2012.

[8] I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, aug 2014.

[9] A. F. Gad, A. M. Hamad, and K. M. Amin, "Crowd density estimation using multiple features categories and multiple regression models," in *2017 12th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, dec 2017.

[10] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2013.

[11] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[12] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012.

[13] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Image and Vision Computing*, vol. 30, no. 12, pp. 966–977, dec 2012. [Online]. Available: https://doi.org/10.1016%2Fj.imavis.2012.08.013

[14] F. Zhu, X. Wang, and N. Yu, "Crowd tracking with dynamic evolution of group structures," in *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 139–154. [Online]. Available: https://doi.org/10.1007%2F978-3-319-10599-4_10

[15] A. Bera and D. Manocha, "Realtime multilevel crowd tracking using reciprocal velocity obstacles," in *2014 22nd International Conference on Pattern Recognition*. IEEE, aug 2014.

[16] C. Direkoglu, M. Sah, and N. E. O'Connor, "Abnormal crowd behavior detection using novel optical flow-based features," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, aug 2017.

[17] S. K. Ramakrishnan, S. K. Ravindran, and A. Mittal, "CoMaL tracking: Tracking points at the object boundaries," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jul 2017.

[18] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Holistic features for real-time crowd behaviour anomaly detection," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016.

[19] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, jan 2015.

[20] A. Bera, S. Kim, and D. Manocha, "Realtime anomaly detection using trajectory-level crowd behavior learning," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2016.

[21] J. Wang and Z. Xu, "Spatio-temporal texture modelling for real-time crowd anomaly detection," *Computer Vision and Image Understanding*, vol. 144, pp. 177–187, mar 2016. [Online]. Available: https://doi.org/10.1016%2Fj.cviu.2015.08.010

[22] H. Rabiee, H. Mousavi, M. Nabi, and M. Ravanbakhsh, "Detection and localization of crowd behavior using a novel tracklet-based model," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 12, pp. 1999–2010, apr 2017. [Online]. Available: https://doi.org/10.1007%2Fs13042-017-0682-8

[23] R. Chaker, Z. A. Aghbari, and I. N. Junejo, "Social network model for crowd anomaly detection and localization," *Pattern Recognition*, vol. 61, pp. 266–281, jan 2017. [Online]. Available: https://doi.org/10.1016%2Fj.patcog.2016.06.016

[24] DeepLearning, "Why deep learning over traditional machine learning?" https://towardsdatascience.com/why-deep-learning-is-needed-overtraditional-machine-learning-1b6a99177063, 2020, accessed: 2020-05-30.

[25] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[26] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[27] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, may 2018. [Online]. Available: https://doi.org/10.1016%2Fj.patrec.2017.07.007

[28] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *2015 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun 2015.

[29] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[30] H. Yao, K. Han, W. Wan, and L. Hou, "Deep spatial regression model for image crowd counting," *arXiv preprint arXiv:1710.09757*, 2017.

[31] G. Olmschenk, H. Tang, and Z. Zhu, "Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2020.

[32] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1610–1623, oct 2010. [Online]. Available: https://doi.org/10.1109%2Ftnn.2010.2066286

[33] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, jul 2018.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[35] D. Gordon, A. Farhadi, and D. Fox, "Re^3: Re al-time recurrent regression networks for visual tracking of generic objects," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 788–795, apr 2018. [Online]. Available: https://doi.org/10.1109%2Flra.2018.2792152

[36] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Tracking by prediction: A deep generative model for mutli-person localisation and tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018.

[37] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time marker-less multi-person 3d pose estimation in RGB-depth camera networks," in *Intelligent Autonomous Systems 15*. Springer International Publishing, dec 2018, pp. 534–545. [Online]. Available: https://doi.org/10.1007%2F978-3-030-01370-7_42

[38] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, jun 2014. [Online]. Available: https://doi.org/10.1016%2Fj.sigpro.2013.12.026

[39] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, nov 2012. [Online]. Available: https://doi.org/10.1109%2Ftsmcc.2011.2178594

[40] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, sep 2016. [Online]. Available: https://doi.org/10.1016%2Fj.image.2016.06.007

[41] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, aug 2017.

[42] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 25–36. [Online]. Available: https://doi.org/10.1007%2F978-3-540-24673-2_3

[43] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018.

[44] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, jul 2018. [Online]. Available: https://doi.org/10.1016%2Fj.cviu.2018.02.006

[45] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019. [Online]. Available: https://doi.org/10.1109%2Faccess.2019.2906275

[46] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, apr 2017. [Online]. Available: https://doi.org/10.1109%2Ftip.2017.2670780

[47] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder," *Computer Vision and Image Understanding*, vol. 195, p. 102920, jun 2020. [Online]. Available: https://doi.org/10.1016%2Fj.cviu.2020.102920

[48] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2017.

[49] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, jul 2017.

[50] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, and C. Choi, "Generative neural networks for anomaly detection in crowded scenes," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1390–1399, may 2019. [Online]. Available: https://doi.org/10.1109%2Ftifs.2018.2878538

[51] D. J. S. R., F. E, G. Manogaran, V. G.N, T. T, J. S, and A. A, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Computer Networks*, vol. 151, pp. 191–200, mar 2019. [Online]. Available: https://doi.org/10.1016%2Fj.comnet.2019.01.028

[52] D. Hou, Y. Cong, G. Sun, J. Liu, and X. Xu, "Anomaly detection via adaptive greedy model," *Neurocomputing*, vol. 330, pp. 369–379, feb 2019. [Online]. Available: https://doi.org/10.1016%2Fj.neucom.2018.09.080

[53] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, jan 2020. [Online]. Available: https://doi.org/10.1016%2Fj.neucom.2019.08.059

[54] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: Counting varying density crowds through attention guided detection and density estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[55] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *Procedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[56] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, dec 2009.

[57] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2010.

[58] J. Li, H. Yang, L. Chen, J. Li, and C. Zhi, "An end-to-end generative adversarial network for crowd counting under complicated scenes," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, jun 2017.

[59] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[60] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai, "Online multi-object tracking with convolutional neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2017.

[61] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2016.

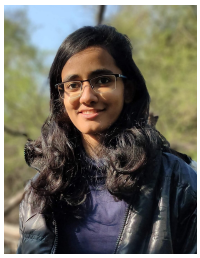
[62] U. crowd activity dataset, "Unusual crowd activity dataset of university of minnesota," http://mha.cs.umn.edu/movies/crowdactivity-all.avi, 2020, accessed: 2020-03-30.

[63] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.

Safvan Vahora Safvan Vahora, completed his Ph.D. from Charusat University, Changa, India. He has received M.E. degree in computer engineering from Gujarat Technological University, India, 2011 and B.E. degree in Information Technology from Sardar Patel University, India, 2009. He is at present working as an Assistant Professor at Department of Information Technology, Government Engineering College, Modasa, Gujarat, India. His research interest spans computer vision, image processing and machine learning. He has published research paper in prestigious conferences and journals in the field of computer vision and machine learning.

Krupa Galiya Krupa Galiya completed her B.E. degree in Information Technology from Vishwakarma Government Engineering College. She is at present working with Patter ai. She is a Public Speaker for Domain like Machine Learning, Deep Learning and Natural language processing. Her Research Area is in Computer Vision with Deep Learning Techniques.

Harsh Sapariya Harsh Sapariya earned his B.E. degree in Information Technology from Vishwakarma Government Engineering College. He is at present working as a Software Developer at Growth Source Financial Technologies Private Limited. His areas of interest are Machine Learning and Deep Learning.



Srijyaa Varshney Srijyaa Varshney completed her B.E. degree in Information Technology from Vishwakarma Government Engineering College. She is at present working as Systems Engineer at Tata Consultancy Services. Her areas of interest are Data Science and Deep Learning.