International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Oct-2022)

https://dx.doi.org/10.12785/ijcds/120187

A Survey on Crowd Anomaly Detection

Harshadkumar S. Modi[1] and Prof. (Dr.) Dhaval A Parikh[2]

[1]*Research Scholar, Gujarat Technological University, Gujarat, India*
[1] *Lecturer, Computer Engineering, Government Polytechnic, Gandhinagar, Gujarat, India*
[2]*Professor and Head, Computer Engineering , Government Engineering College, Gandhinagar, Gujarat, India*

Received 22 Jan. 2021, Revised 15 Jul. 2022, Accepted 23 Jul. 2022, Published 31 Oct. 2022

Abstract: Automated crowd anomaly detection and crowd scene analysis is a novel and emerging field of computer science and engineering domain. The analysis of crowd behavior based on density, trajectory, and motion helps prevent abnormal and unwanted incidents. The analysis of crowd behavior is complex and challenging due to visual occlusions, clutter, ambiguities, dense crowd, and scene semantics. These days, researchers are focusing on developing machine learning-based approaches for "crowd behavior, activity analysis, motion pattering, and anomaly detection in real-time applications". Firstly this study presents insight on crowd anomaly detection, ways to achieve it, and its applications and importance today. Secondly, it presents a detailed analysis of conventional machine learning as well as deep learning approaches for serving the purpose based on features, methods, datasets, and shortcomings. Thirdly it presents a thorough analysis of datasets and performance parameters. Finally, it presents the current challenges and future work in this field.

Keywords: Crowd Anomaly, Motion Features, Performance Parameters, Abnormal Event Detection, Video Surveillance

1. Introduction

Due to the growth of the human population as well as the variety of "human activities, crowded situations" have become more familiar than ever in the actual world. It poses significant difficulties to "public administration, security, and safety" [1]. Figure 1 shows several instances of congested situations. Human beings have the capacity to extort relevant information on "behavior patterns in the surveillance area", watch the progress in real-time for unexpected circumstances, and respond quickly [2]. Yet, psycho-physical studies suggest that their capacity to detect multiple signals is severely limited [3]. Overly crowded situations need a substantial amount of monitoring of a large count of people and their behaviors, which is a big issue.

Automatic scene comprehension or investigation has gotten a lot of attention in the computer field in the last decade [4]. Though numerous algorithms for tracking, recognizing, and understanding the actions of diverse objects in the video have been created [5], they were mostly intended for typical settings with low population density [6]. When dealing with a high-density crowd, the problem becomes more difficult to solve since the huge count of individuals included causes detection to fail and increases computing complexity. In such situations, crowd analysis becomes an important topic and requires special attention. It's becoming a popular study topic, and it's already attracting a lot of attention [5], [7], [8].

(a) (b)

Figure 1. Illustrations of crowded status (a) Normal scene (b) Abnormal scene [9]

Surveillance cameras are now deployed in nearly every area of modern society, resulting in a tremendous volume of video data. Video analysis has been a popular and important academic topic as processing power and hardware has improved, security cameras have become less expensive, and enormous amounts of video data have become available [10]. "Human behavior identification, traffic monitoring, and violence detection" are just a few of the real-time uses. Investigating massive amounts of video represents a time-consuming process. As a result, intelligent monitoring is critical, with human operators being immediately informed when there is an anomaly in the recorded footage. For intelligent surveillance cameras, detecting video anomalies is a painstaking process.

E-mail address: harshadsmodi@gmail.com, daparikh@gecg28.ac.in

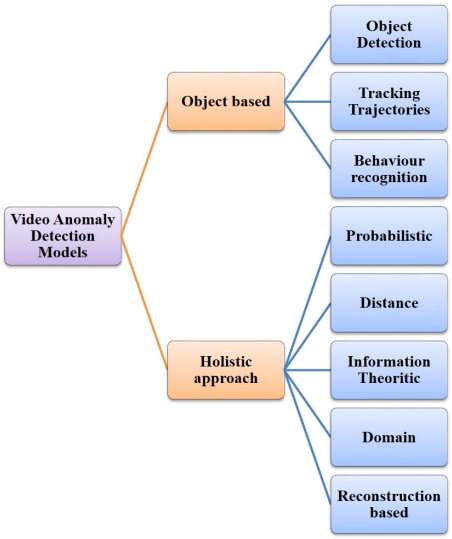

Figure 2. Categorization of video anomaly detection methods [12]

In both crowded and uncrowded situations, video anomaly detection may be performed. Due to occlusions, detecting an anomalous occurrence in a crowded situation is more difficult. Anomaly detection is further divided into local and global anomaly detection. The method of detecting global abnormality behavior in individuals in surveillance footage is known as global anomaly detection [11]. When an unexpected event occurs, like "a bomb blast, an accident, or violence", the majority of people flee in separate directions. The technique of detecting local anomaly behavior in a person or a group of individuals is known as local anomaly detection. For example, a person riding a bicycle on a route when everyone is walking [12].

Object-centric or trajectory-oriented models of video anomaly detection are distinguished from holistic or non-trajectory-oriented approaches. The crowd is considered as a group of people in this technique. Image frames are segmented, and every item is tracked using trajectory-oriented models [13]. However, models on the basis of holistic methods consider the audience as a whole [14]. In this type of detection method spatiotemporal characteristics are derived from frames to detect unexpected events. The primary disadvantage of trajectory-oriented video anomaly detection is that it performs worse as crowd density rises. When the crowd density is large, it is impossible to follow all persons, and the trajectories may appear sloppy, making anomaly identification harder. In crowded situations, a non-trajectory-oriented or holistic approach is better for detecting video anomalies. Non-trajectory-oriented techniques may be further categorized as "probabilistic based, distance-based, information theoretic-based, domain-based, and reconstruction based" depending on the techniques used, as illustrated in figure 2.

Normal occurrences contain a lower likelihood than abnormal events, according to the probabilistic-oriented approach. The technique is trained on the data, and its probability is calculated. An anomaly is defined as a probability that is less than the threshold [15]. The distance-oriented model posits that typical occurrences take place in a concentrated location, whereas abnormal events occur far away. Domain-oriented models are used to train data and understand its classifications, forming a region [16]. The test data position is computed, and the worst event is identified in relation to this boundary. Information-theoretic methods imply that the anomalous occurrence contains a significant influence on the dataset's varying information. As a result, the data is regarded anomalous if its eradication causes a significant variation in the dataset's content.

Methods on the basis of reconstruction error presume that anomalous occurrences contain a higher reconstruction error than regular events. Generally, any particular model encodes and decodes image sequences. The "reconstruction error" is then computed. If it is greater than the specified threshold, abnormal events are taken into account.

Due to its real-time uses, video abnormal event detection is becoming increasingly important. The monitored surroundings are almost always public spaces. These regions are frequently congested. Owing to" "high clutter, severe occlusions, and ambiguities", traditional techniques without additional considerations are ineffective in crowded situations. As a result, developing video anomaly detection in crowded situations is a critical challenge [12].

A. *Real world applications*

Several applications rely on efficient crowded scene analysis. Such applications are summarized in figure 3.

- **Automated Video Surveillance:** The goal of video-oriented crowd behavior identification is to solve difficult challenges like automating and recognizing varying crowd behaviors in real-life settings [17], [18]. Crowd behavior analysis outcomes may be applied to "crowd flow statistics and congestion analysis, anomaly detection and alerting, and other applications" [19], [20].
- **Public Events Management:** Crowd scene analysis and monitoring is applied to events like "concerts, religious places, political rallies, sports events" etc. to avoid specific disastrous situations [21], [22].
- **Virtual environments:** Virtual environments are essential to develop the mathematical description of crowd investigations for augmenting the simulation of the crowd and human life experience [22], [23].
- **Public Space Design:** Concerned with the construction and renovation of public areas such as retail malls, stadiums, railway stations, rail tracks, and airports [24].

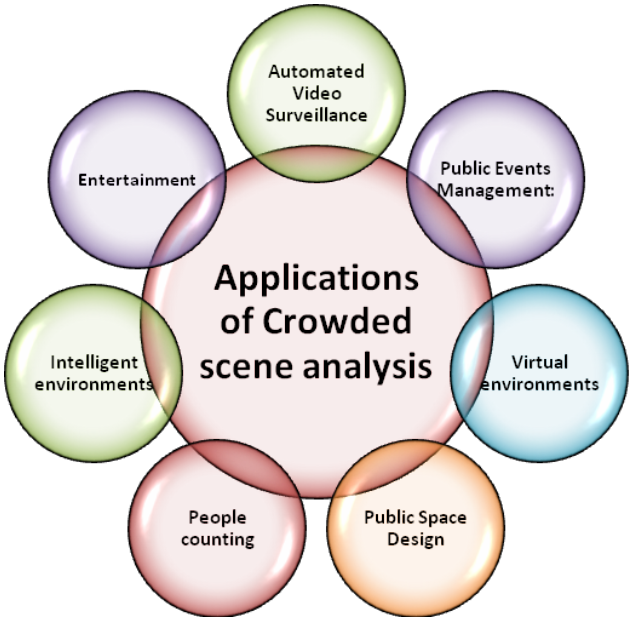

Figure 3. Applications of crowded scene analysis [21]

- **People counting in densely populated areas:** Increases in the count of people cause issues, especially in smaller regions [21]. For example, deaths, bodily injuries, and so on. These issues can be avoided if such a gathering is detected early.
- **Intelligent environments:** If crowd scene investigation is done intelligently, then it can assist the whole crowd or a person to follow the existing and desired patterns of the crowd [22].
- **Entertainment:** The creation of mathematical methods based on a deeper knowledge of crowd dynamics can give more realistic experimentation that is to be utilized in "computer gaming, film, and television industries" [1]. Synthesizing crowd recordings with realistic micro-scale behavior has been proposed in recent work [25].

B. Motivation

In a country like India where stampede occurs very frequently as we have a large number of people gathering in different scenarios at many places, such as any religious place, during the strike, political or non-political rally, etc. The recent examples of such stampede are the protest against the new law of Act 370, National Register of Citizens (NRC) or Citizenship Amendment Act (CAA), violation of Lockdown, Anomalous activities at Border areas, etc. Above stampede during pandemic motivates us to do research on anomaly detection in video surveillance.

C. Organization of the paper

The rest of the paper is organized as follows: Section 2 outlines related work in crowd anomaly detection. In section 3 we have compiled popular "benchmark crowd video datasets. Section 4 provides the study of various performance metrics employed to track the efficiency. In section 5 we have listed the challenges faced by the researchers in the area of crowd anomaly detection. We conclude the paper in Section 6 and end this review by providing some promising future directions in section 7.

2. Literature Survey

The general scenario of any type of anomaly detection is to first extract the frames from the video, then extract the different features available in the frames. The various features are like texture descriptor, color features, shape dimensions, optical flow, motion, histogram, tracklet, etc. Following feature extraction, the data will be differentiated with some ground truth to see if an abnormality happened.

There are certain physical crowd models as well as background information accessible are studied. They might be included in the techniques of analysis. The schematic of the crowded scene analysis is shown in figure 4. As stated in the papers surveyed, several numerous strategies are categorized into two major sets on the basis of the technology employed in anomaly detection: conventional or machine learning-oriented and deep learning-oriented approaches [26], [27], [28], [29], [30]. Work done by the authors is surveyed and summary tables, table I, and table II respectively for both the approaches are prepared. For the evaluation parameter, the values listed in the tables are based on the results produced with the respective method using the first dataset in case the method was applied to multiple datasets. All the values written for parameter are in percentage.

A. Conventional/Machine learning based methods

Li *et* al. [31] have suggested a global-frame scale technique to find "abnormal events in crowded scenes" that included two major procedures: the first is to calculate the "Histogram of Maximal Optical Flow Projection (HMOFP)" of the input on the basis of the saliency map in the OFF, and the second is to use the "online dictionary learning" technique to attain the "optimized dictionary based on the optimal dictionaries". It is obtained using the "Scale Invariant Feature Transforms (SIFT)" technique [32]. The real-time requirement may be satisfied using an improved process. The SIFT technique lowers the count of optical flow points [33], requiring less computation for computing HMOFP.

Chong *et* al. [34] have developed an unique method for detecting anomalous occurrences by modelling the spatiotemporal distribution of crowd movements. The suggested technique uses trajectory investigation with "Hierarchical Dirichlet Processes (HDP)" to identify the Regions of Interest (ROIs) that explain behavioral patterns, allowing the major crowd motions to be represented. Following the ROIs, a series of global as well as local histograms were created as "templates for the observed movement distribution",

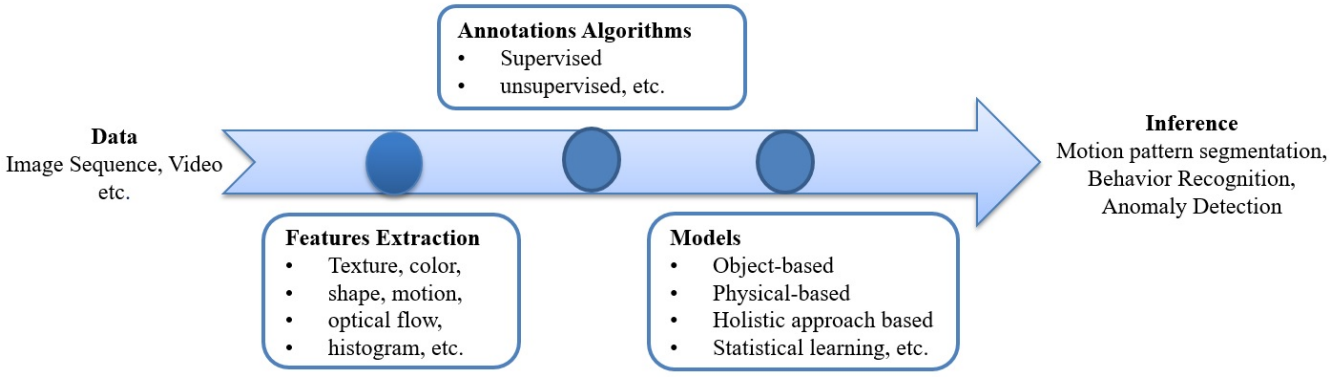

Figure 4. Crowded scene analysis general model [1]

which quantitatively represents "time-correlated crowd occurrences". After being constructed in a hierarchical manner, the down-sampling method is implemented to minimize the dataset. This can evaluate and extort information from recorded datasets, as well as to detect anomalies at different levels. The key feature of this technique is that it can understand the moving patterns of crowds both globally and locally, as well as statistically define crowd behaviors. The benefit of the introduced method is that just the density, as well as flow data streams are analyzed during the identification phase, eliminating the requirement to follow individual trajectories [35]. Rather than being explicitly defined, HDP may accurately discover the behaviors that correlate with the recorded dataset. HDP can understand more appropriate outcomes than the clustering techniques by doing co-occurrence analysis. Trajectories do not have to be aligned in time for HDP to work. The study has a flaw in that because HDP is a statistically-oriented technique, the findings aren't always correct, thus low-importance areas are considered as useless output and rejected [36]. This technique cannot incorporate the underlying temporal logic in crowd behaviors and activities owing to HDP's restriction of not being able to explain the temporal logic in co-occurrence analysis.

Ojha *et* al. [37] have proposed an approach which first detects the moving crowds. After that, moving sub-objects were segmented so that they could be handled as unique identities. The information of the detected objects was utilized for crowd tracking [38]. In order to the final step, the speed estimation was performed, to determine the anomalous crowd's susceptible motions. The weakness of this approach is that the detection is based on only speed, so it is limited to detect only over-speeding suspicion, also it is dependent on orientation and illumination.

Yousefi *et* al. in [39] propose the algorithm which consists of three modules (i) "Low-Level Discovering (LLD)" to create "an activity pattern codebook, (ii) High-Level Discovering (HLD)" to describe salient characteristics in a dictionary, and (iii) identification of video input data normalcy. With the help of "Locality-constrained Linear Coding (LLC)", this study increases the accuracy of localized anomaly detection. The limitation of this technique is that it is generally intense, and a separate codebook is required to make for each video.

Li *et* al. in [40] have treated anomaly detection as a "Maximum A Posteriori (MAP)" issue, in which the knowledge is derived via background subtraction with the help of "Robust Principal Component Analysis (RPCA)" and the probability is calculated by a "trained global maximum grid template" [41]. It's worth noting that prior understanding isn't restricted to only the "foreground binary map"; it may be replaced with various other successful techniques. For anomaly identification, a simple motion feature is utilized, which is similar to the "Histogram of Optical Flow (HOF)" [42]. They tested their technique using publicly accessible datasets from "UCSD, UMN, and Avenue", and found that it successfully detects anomalous occurrences in complicated scenarios. The technique computes the "likelihood function" using the "global maximum grid template"; as a result, if there is a little amount of "training data or the motion patterns" are only focused on a few areas, the technique may fail. Anomaly identification may fail if the derived "global maximum grid template" loses the matching "motion patterns at the sites".

Qasim *et* al. [43] have developed a "modified Ant Colony Optimization (ACO) clustering algorithm" [44] that divides the "2D variance plane" into "salient and non-salient clusters". The high region is represented by the cluster of salient pixels. A "Histogram of Swarms (HoS)" is generated for each frame by a new predator-prey algorithm [45]. The greatest frame parts are retained by the ACO application. For anomaly detection, this results in a highly discriminative feature vector. However, the authors have not done a comparative analysis of the proposed novel HoS feature. More features may be added, and a fusion of features may be evaluated to enhance the performance in future work.

"Location-based Social Networks" (LBSNs) give geo-

TABLE I. CONVENTIONAL AND MACHINE LEARNING APPROACHES

Work	Features	Method	Dataset	Parameter	Open Issues
[31]	HMOFP	SRC	UMN	AUC-94.7	To provide rotation/illumination independent solution.
[34]	Motion	Optical Flow	Edinburgh	ROI Detection	Using the temporal logic, model and identify more complex crowd actions. The HMM may be used to fill the gap, and it's very good at describing temporal patterns.
[37]	Optical Flow, SIFT	Speed and Distance calculation	Local Video	-	Machine Learning approach can be analyzed.
[39]	HOG, HOF	Fuzzy C-Means clustering	UCSD	EER-23	Locality parameter is ignored.
[40]	Motion	Optical Flow	Avenue, UCSD, UMN	AUC-86.82	A small quantity of training data may lead to a failure.
[43]	Histogram of Swarms (HOS)	Optical Flow	UCF, UMN	AUC-98.54	More features may be added, and fusion of features may be evaluated
[46]	Geo-tagging	Entropy Analysis and Clustering	Instagram	Entropy	Other behavioral aspects like seasonality can be taken into account. Study how seasonality affects the findings and how to constantly adjust the size of the entropy search window to account for these variations and deliver correct outcomes.
[47]	Magnitude, Joint entropy, Variance, Optical Flow	SVM	UMN	AUC-99.96	Aside from the motion-oriented features utilized in this paper, the suggested technique may be enhanced by integrating "features that learn the body posture of walking individuals in typical footage in a surveillance camera".
[48]	HOF	SVM	UCSD, UMN	EER-20	Multiple data patterns can be updated (added and removed) at the same time.
[49]	Fusion among FBG FOS data and accelerometer data	SVM	Manual dataset	Accuracy-95	It does not operate with the publically available dataset.
[50]	Deep and hand crafted features	OCSVM	PETS2009, UMN	AUC-94	The anomalies are not detected in a broader range of applications using the combination of both appearance and motion deep features.
[51]	EKF, LBP, GLCM	KSVM	PETS2009, MIT	Accuracy-91, Precision-90.02, Recall-90	The usage of various classifiers is not explored.
[52]	Histogram of Magnitude	Automated approach	UCSD, UMN	AUC-82.31, EER-21.43	It becomes difficult to segregate the anomalous and connected normal objects in some cases.
[53]	Appearance and motion models	OCSVM	Avenue, UMN	AUC-84.52	The multiple information sources are not exploited.

positioned data, which was utilized in [54] to discover patterns of crowd dynamics in metropolitan locations. It has been observed that surprising characteristics are indicative of variations in city activities. Redondo *et* al. in [46] applied a hybrid approach to the data collected from LBSNs by combining entropy analysis as well as clustering approaches. In the smart city industry, LBSNs offer a highly appealing source of geo-located data that might be an intriguing to standard video sources for monitoring human activities [55]. Because of the widespread availability of LBSNs, the region under investigation may be readily altered without incurring additional costs. This method uses fewer computing resources than a video-oriented methods. Yet, several crucial data, such as the precise position of the discovered outliers and why these characteristics are labeled abnormalities, are not available using this innovative technique. For future work, other behavioral aspects like seasonality can be taken into account.

Qasim *et* al. [47] introduced an effective descriptor for extracting distinct characteristics of video sequences from optical flow (OF). The suggested descriptor is effective concerning fps rate and accuracy owing to the "low dimensionality and simplicity of the individual features". When utilized in combination, the three features: (i) is the sum of the optical flow field magnitude, (ii) the joint entropy of the OF magnitude of two consecutive frames, and (iii) the variance computed from a space-time cuboid, resulting in very strong discriminative power. The threshold is calculated directly, and it aids in the elimination of noise-induced fluctuations in the background region. This approach smooths out any minor changes caused by noise impacts and prevents the discovery of false outliers. The suggested descriptor produces some false positives because it is dependent entirely on changes in global motion patterns. Apart from the motion-oriented features utilized in this study, the suggested technique may be enhanced further by integrating features in a surveillance camera.

Lin *et* al. [48] developed a unique technique for "anomaly detection in crowd situations" using "online adaptive one-class Support Vector Machines (SVMs)". Incorporating this with a sliding buffer results in a main process that not only alters the technique in real time with little computational cost, but also eliminates outdated patterns. The recommended technique offers a unified framework for both global as well as local anomaly detection. This approach divided video segments into a collection of video events, and then used k-means clustering to create a visual vocabulary from a random subgroup of descriptors retrieved from the training group. Next, it will allocate every description to the vocabulary term that is nearest to it. To train one-class SVMs, histograms of video events are produced. The limitation of this approach is that it can only change (include or delete) a single pattern at a time.

B. Deep learning based methods

Kong *et* al. [56] presented the "Hierarchical Urban Anomaly Detection (HUAD) framework". This framework creates preliminary anomaly features that must be computed using data from subway and taxi traffic flows. The alternate anomalous areas were then retrieved. Next, to acquire the historical anomaly scores, the "Long Short-Term Memory (LSTM)" [57] is utilized to forecast the traffic. These features are then derived from "neighboring areas, adjacent periods, and previous anomalies". One-Class OC-SVM [58] was used to detect the last anomalous areas. To increase the method's accuracy, spatio-temporal data from taxis and subways are combined. To minimize overfitting, data augmentation and normalization must be performed on data to boost the training impact and enhance the accuracy of the training process.

Murugan *et* al. in [59] for quicker identification and to deal with scalability concerns, a technique dependent on region-oriented suggestions was presented. In this paradigm, abnormalities are characterized as readings with a probability under a certain threshold. The RS-CNN method was created with the help of the well-known Faster R-CNN detection framework. To recognize anomalies available in an image, the RS-CNN method gets the full image as well as the number of object proposals as input. The RS-CNN model's findings were compared to those of "Fast R-CNN, Minimization of Drive Testing (MDT), Mixtures of Probabilistic Principal Component Analyzers (MPPCA), and Social Force (SF)". The suggested method was put to the test on a variety of image sequences with various sizes and numbers of anomalies. Once the entire test images were used, the research result confirmed an acceptable detection rate. This method's ability to manage a wide range of anomalies considers it useful and paves way for improved detection characteristics. To identify various-sized abnormalities, the suggested "Region Proposal Network (RPN) with scalable feature" is utilized. It may be used in smart systems to swiftly consider aberrant behaviors in surveillance films. This model has a poor level of precision and necessitates additional computing.

Chen *et* al. [60] presented a system on the basis of bidirectional prediction, in which the forward, as well as backward prediction sub networks, forecast the identical target frame. The loss function is then built using the real target frame as well as the bidirectional prediction frame. A technique for estimating anomaly scores on the basis of the sliding window strategy was also suggested. These are much more sensitive to anomalous occurrences since they generally pertain to only one or a few frames, making anomaly detection easier. Frame techniques often have simpler network architecture, "greater features, less reliance on domain expertise, and much more open access to temporal data (frame prediction techniques)". Although the offline training method used by P-GAN has no direct effect on testing efficiency, it does limit the model's ability to adapt to new settings and tasks [71]. Probability

TABLE II. DEEP LEARNING APPROACHES

Work	Input Data	Network	Classifier	Dataset	Parameter	Open Issues
[56]	GPS Data	LSTM	OC-SVM	Taxi GPS Trajectory Data, Subway Data	Recall-71	To increase training precision and strengthen the training impact, data enhancement, and normalization ned to be carried out on data to prevent over-fitting
[59]	Object Proposals	RPN, Fast R-CNN, RS-CNN,	R-CNN	UCSD Ped-1, UCSD Ped-2	Accuracy-86.7	The "RS-CNN model may be used in smart systems" to recognize anomalous activities in surveillance footage rapidly.
[60]	Frames	U-Net	PSNR	Avenue, UCSD	AUC-87.8	Model's given best configuration may not be appropriate for other datasets, and adaptable hyper-parameter adjustment techniques require further studies.
[61]	Ensemble of pre-trained CNNs	VGGNet, AlexNet, GoogleNet, ConvNets	SVM	Avenue, UCSD	Accuracy-92.7, AUC-89.3	Running time of a system may be high due to aggregation of ensembles.
[62]	Shape and motion features	Caffe reference model, FCN	Gaussian	UCSD, Subway	AUC-90.4, EER-17	High rate of false-positives.
[63]	Normal gradient and OFP	AAE	MGFC-AAE	Avenue, UCSD, UMN	AUC-84.2, EER-22.3	For extracting the temporal features of video sequences, the LSTM was combined with the adversarial auto-encoder (AAE).
[12]	Combination of Raw image and edge image sequences	AE, LSTM, ConvLSTM	HDLVAD	Avenue, UCSD	Accuracy-90.7	Investigating video streaming in big data.
[64]	Appearance and motion of an image	GAN	OC-SVM	UCSD, UMN	AUC-95.3, EER-11	It requires more training time as it uses Generative Model.
[65]	Spatio-temporal features of normal patterns	ConvLSTM	STAN	Avenue, UCSD	AUC-87.2	It requires more training time as it uses Generative Model.
[66]	Gaussian models	DL	SVM	UCSD Ped-1, UCSD Ped-2	TPR-83	It does not return appropriate solutions for monitoring the crowd characteristics.
[67]	Video processing features	CNN	ML	COCO train	-	It does not integrate various features for better detection models.
[68]	OF	CNN, bidirectional LSTM	Inception v3	UCSD Ped-1, UCSD Ped-2	Accuracy-85.64, AUC-94.83	It does not implement the DL-based OF estimation technique.
[69]	Hand crafted features-based method	DLADT-PW	DenseNet 169	UCSD	Accuracy-89.6	It cannot be used in several real time scenarios such as the vehicle detection in pedestrian walkways.
[70]	Spatial temporal features	Graph convolutional network	Graph neural network	Avenue, ShanghaiTech	AUC-87.3	The relationship of joints is not characterized over time.

estimation techniques generally operate with probability methods in the lack of domain information and restricted computer resources, which leads to greater generalization while reducing sensitivity to unknown abnormal events. The correct spatiotemporal scale of anomalous occurrences is difficult to identify, and the computation of "multi-scale features extraction and statistical model estimation" takes a long time [72]. The suggested method has an obvious description, but its underlying mechanism is unknown. The model's best configuration may not be appropriate for all potential datasets, and adaptable hyper-parameter adjustment techniques need more investigation. U-Net, as a backbone network, is simply not the right choice for collecting temporal sequence features.

Singh *et* al. [61] introduced the "Aggregation of Ensembles (AOE)" idea for identifying an abnormality in video data, which builds on the current capacity of "pre-trained ConvNets as well as a pool of classifiers". The hypothesized AOE approach trains versions of SVM classifiers, which are then merged to forecast the anomaly in crowd frame sequences. In various computer vision applications with minimal training data samples, the suggested approach can produce competitive results. The suggested AOE paradigm is extremely resilient, outperforming conventional approaches even when data is scarce, instead of using it as a feature extractor. As a result, this transfer learning technique is a deployable and potential countermeasure in situations when the data supplied is incredibly sparse and the complexity of the categorization issue statement is fair. AOE needs a certain quantity of acceptable data in order to simplify the implementation of CNN fine tuning. It is extremely likely that the system will not function properly if the described fine-tuning method is not used in an AOE installation.

Sabokrou *et* al. in [62] converted a "pre-trained supervised FCN" into an "unsupervised FCN" utilizing "fully convolutional neural networks (FCNs) and temporal data", enabling the identification of (global) abnormalities in images. "Feature representation and cascaded outlier detection" are the two major objectives addressed by this FCN-oriented architecture. The following is the suggested method: Initially, input frames are sent to an FCN that has already been trained. The "output of the k^{th} layer" then generates regional feature vectors. The "Gaussian classifier G1" is used to verify these feature vectors. Patches that deviate considerably from G1 are classified as aberrant. On a typical GPU, the authors were able to reach a processing speed of 370 frames per second, which is around three times quicker than the fastest available technique. Deeper features are time-consuming, and going further cause's bigger activation functions, which raises the incorrect localization, which contains inverse consequences on characteristics, therefore deeper layers are disregarded in the suggested technique. Fully CNNs, which conduct feature extraction and localization simultaneously, are used in the suggested methodology. This feature reduces the number of computations required. The suggested approach simply uses two convolutional layers to detect anomalous regions, while certain regions are classified by a sparse auto-encoder. This results in the less computations required to process these shallow layers. The suggested method produces a high incidence of false positives in situations such as overcrowding and persons walking in opposite directions.

Li *et* al. [63] suggested "Multivariate Gaussian Fully Convolution Adversarial Auto-encoder (MGFC-AAE) framework" based on only normal data and learns the hidden representation as a multivariate normal curve. An anomaly is defined as test samples whose hidden form is not connected with the "multivariate normal curve". The adversarial auto-encoder decoder is ignored in order to clarify the model. The proposed technique employs a common two-stream network to identify "appearance and motion anomalies" using "3D gradient and optical flow maps", accordingly. This employs data obtained from "local gradient and optical flow patches" as inputs to identify abnormalities happening in small areas. A "multi-scale patch approach" is developed to cope with the viewpoint issue in certain video situations, which specifies the video patch size based on their location to the camera [73]. For anomaly identification, "normal gradient and optical flow patches" of various sizes are independently fed into the "two-stream MGFC-AAE". The MGFC-AAE was unable to detect the abnormal optical flow patches due to their tiny size.

Ramchandran *et* al. [12] suggested a framework for unsupervised deep learning that is successful. "Raw image sequences are merged with edge image sequences" and fed into the ConvLSTM model of a convolutional autoencoder. To detect anomalous occurrences, a reconstruction difference is calculated. Unsupervised "Hybrid Deep Learning Framework for Video Anomaly Detection (HDLVAD)" was developed, which efficiently identifies video abnormalities without the use of class labels. The technique provides greater accuracy with less computing cost by integrating handmade features and feature learning with the DL method. In the suggested approach spatial information is given manually to the model to lower the number of convolution layers. Also, to save the space all the frames are turned into the grayscale. The suggested HDLVAD technique does neither overfit nor underfit the data. The suggested approach merely recognizes only the frames in which the abnormal event happens and does not locate the aberrant object in those frames.

Ravanbakhsh *et* al. in [64] suggested to use "Generative Adversarial Nets (GANs)", which are imparted to exclusively create the data's "normal distribution" [74]. A "discriminator (D) is employed as a supervisor for the generating network (G)" during adversarial GAN training, and conversely. Anomalies are detected during the testing time (D). Furthermore, a cross-channel method is utilized to prevent (G) from learning a trivial identity function,

requiring (G) to transform raw-pixel input into motion information and vice versa. With this method, there is no "need to train one-class SVMs or other classifiers". It keeps track of local data. There is no set of data hyper-parameters to tweak in the suggested method. This makes the suggested approach highly resilient.

Lee *et* al. in [65] presented a new Spatio-Temporal Adversarial Network-based anomalous event detection technique (STAN). Using bidirectional ConvLSTM, a "spatio-temporal generator" is created. With 3D convolutional layers, it evaluates if an input sequence is real-normal or not. These two networks are trained to efficiently encode spatio-temporal characteristics of normal patterns in an adversarial manner. Following the learning, the "generator and discriminator" can be employed separately as detectors. This approach achieves robust anomaly detection since the "generator and discriminator complement" each other's detection findings. Only the generator was taught during training, which reduces the pixel-wise loss. At every frame, the method also identifies the place where the anomalous occurrences happen. The suggested technique had a flaw in that damaged frames were recognized as an unusual activity even though they were labeled as regular events in the ground truth.

C. Advancement in Crowd Anomaly Detection Techniques

Most of the Crowd Anomaly Detection techniques are complex in nature and categorize into two broad areas: Conventional/Machine Learning Approaches and Deep Learning Approaches. The statistics of various methods of conventional and machine learning approaches reviewed in this paper are shown in figure 5.

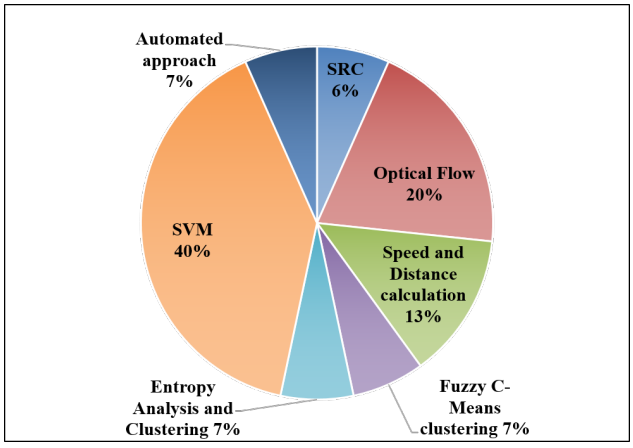

Figure 5. Statistics on Conventional and Machine Learning Approaches

Nowadays, there is a rapid advancement in Crowd Anomaly Detection with new capabilities models such as RPN, U-Net, VGGNet, AlexNet, GoogleNet, Graph convolutional network, LSTM, CNN, Caffe reference model, FCN, AAE, AE, GAN, DLADT-PW, etc. In order to identify anomalies and to classify abnormality of images, different

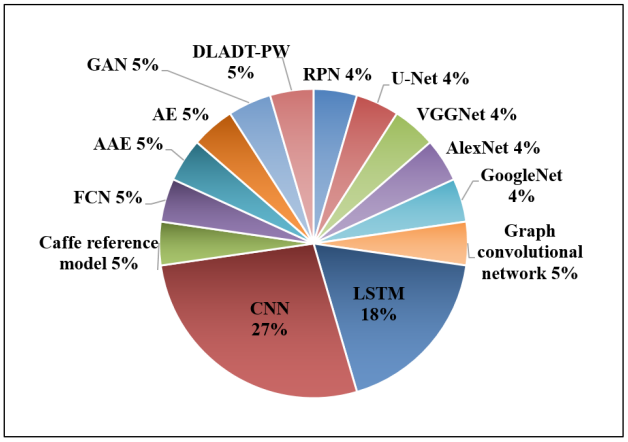

Figure 6. Statistics on Deep Learning Approaches

methods/techniques have been proposed and investigated by engineers and researchers. From the methods reviewed, the current trend of advancement in crowd anomaly detection is moving towards the deep learning approaches including CNN and LSTM, as shown in figure 6.

3. Crowd Video Datasets

Since researchers are concentrating on crowded scene investigation, several crowd datasets are now accessible. We've compiled a list of the most popular "benchmark crowd video datasets". In table III, we've summarized the descriptions, and anomalies found in every "datasets, database size, labeling degree, and accessibility".

- **UCF Crowd Dataset [75]:** "University of Central Florida, center for research in computer vision" provides crowded video data for crowded scene analysis. It provided different videos like UCF-CC-50, UCF-QNRF, UCF aerial action, UCF sports action etc. for crowd counting, action recognition etc. type of purposes.

- **UMN Crowd Dataset [9], [76]:** A dataset of University of Minnesota contains unusual crowd activities. It is open to the public, and every video starts with typical characteristics and finishes with the progress of aberrant characteristics.

- **UCSD Anomaly Detection Dataset [77]:** A CCTV camera mounted on pedestrian paths was used to collect the data set. The population density ranges from low to extremely dense. The circulation of non-pedestrian items such as "cyclists, skaters, small carts, or individuals strolling" over a path resulted in abnormal occurrences. The whole recording was carried out at the "University of California, San Diego (UCSD) in the United States".

- **CUHK Avenue Dataset for abnormal event detection [78], [79]:** The films were shot on the campus avenue of "CUHK (The Chinese University of Hong

TABLE III. ANALYSIS OF DATASETS

Dataset	Example Anomalies	Size	Label	Accessibility
UCF	Traffic, accidents, crimes or illegal activities	38 Videos	Partial	Yes
UMN	Run, panic escape	11 Videos	All	Yes
UCSD Ped 1	"Bikers, skaters, small carts, and people walking across a walkway"	70 Videos	All	Yes
UCSD Ped 2	"Bikers, skaters, small carts, and people walking across a walkway"	28 Videos	All	Yes
CUHK Avenue	Run, throw, strange action, wrong direction, non-human objects	37 Videos	Partial	Yes
Violent-Flows	Crowd violence	246 Videos	Partial	Yes
QMUL Dataset	"Illegal U-turn, vehicles did not follow the typical temporal order"	4 Videos	Partial	Yes
Subway Dataset	"Commuters walking in the wrong direction, loitering and avoiding payment"	2 Videos	Partial	Yes
BOSS Dataset	Harass, Disease, Panic	70 Videos	Partial	Yes

Kong)" and include "16 training and 21 testing video clips totaling 30652" (15328 training, 15324 testing) frames.

- **Violent-Flows Crowd Violence and Non-violence Dataset [80], [81], [82]:** A library of "real-world video footage of crowd violence", as well as established benchmark methods for determining violent/non-violent categorization and detecting violence outbreaks.
- **QMUL Dataset [83], [84]:** QMUL Dataset of Queen Mary University of London contains 3 videos of dense traffic flow and one video of a shopping mall. This dataset contains over 60000 labeled pedestrians and is useful for the crowd counting and profiling research [85].
- **Subway Entry/Exit Dataset [86], [87]:** This dataset has two categories, 1 hour 36 minute and 10 second long video of entrance and 43 minute 16 second long video of exit. It contains the unusual events like commuters "walking in the wrong direction, loitering and avoiding payment".
- **BOSS Dataset [88], [89]:** This dataset was initially acquired as part of the Eureka's Celtic Initiative project "BOSS: On Board Wireless Secured Video Surveillance BOSS". It's a compilation of videos shot by various cameras within a moving train. The data has difficult illumination circumstances when the train is traveling. Since the cameras are positioned on the vehicle, the data has difficult viewing circumstances (involving occlusions).

4. Performance Measures

We must track the effectiveness of every newly produced model by comparing it to other conventional or state of the art approaches once we construct it. Different performance metrics are employed for this aim. When assessing the model, selecting the appropriate metric is critical. Here's we provide the summary of various metrics used to assess anomaly detection methods.

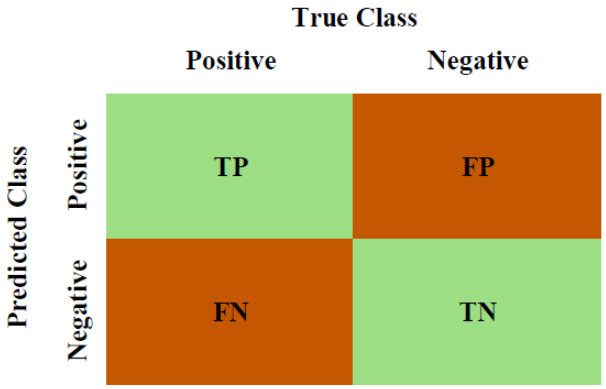

Figure 7. Confusion Matrix [90]

- **Confusion Matrix [90]:** It's a $N \times N$ matrix, where N shows the count of predicted classes. We have a 2×2 matrix if we divide the problem into two classes. The examples in a forecasted class are indicated by the rows of the confusion matrix, whereas the occurrences in an actual class are indicated by the columns. The Confusion Matrix is the foundation for all remaining measurements.
  As shown in figure 7, there are four important terms as discussed here:
  1) "True Positive (TP): Predicted positive and it's true".
  2) "True Negative (TN): Predicted negative and it's true".
  3) "False Positive (FP): Predicted positive but it's false".
  4) "False Negative (FN): Predicted negative but

it's false".

- **Classification Accuracy [91]:** It's a metric for how accurate a detection is.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Precision [92]:** When the classifier is unbalanced, i.e. one class is more common than the others, classification accuracy is not a useful measure of model success. In this scenario, we would achieve a high accuracy rate even if we predicted the entire sample as one of the most common classes, which makes no sense because our system isn't learning everything and is simply predicting something as the top class. As a result, we must consider class-specific performance metrics, one of which is precision.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- **Sensitivity [93]:** A "true positive rate, or recall", is another term for this. In comparison to the entire positive samples, sensitivity refers to the fraction of positive samples that are accurately classified as positive.

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

- **Specificity [93]:** This, also called as the "true negative rate", indicates to the percentage of negative samples that are accurately classified as negative when compared to the entire negative samples.

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

- **False positive rate [93]:** In comparison to all negative samples, this is the fraction of negative samples that are incorrectly assumed as positive.

$$Falsepositiverate = \frac{FP}{FP + TN} \quad (5)$$

- **False negative rate:** This is also known as the miss rate. While assessing the samples it shows the probability of missing the true positive by a test.

$$Falsenegativerate = \frac{FN}{FN + TP} \quad (6)$$

- **F-Measure [94]:** Depending on the application, it is required to give greater priority to either precision or recall, but in some applications both precision and recall are equally important. Therefore there's a need to combine these two measures. F-measure also known as f1-score combines these two measures that is the "harmonic mean of precision and recall".

$$f - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

- **Matthews correlation coefficient [95]:** It is a metric for evaluating "binary and multiclass classification quality".

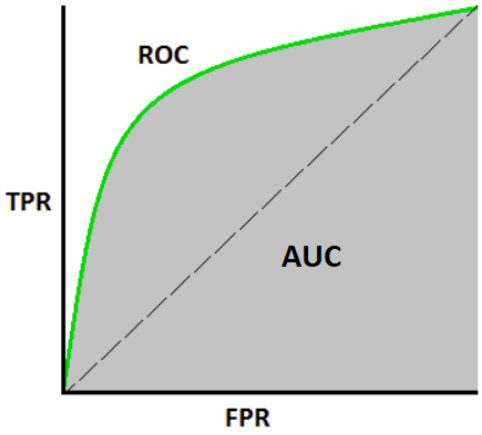

Figure 8. Area Under Curve [92]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (8)$$

A correlation coefficient value ranges in the interval of [1,+1], the extreme values –1 reached when there is a perfect misclassification and the extreme values +1 reached when there is a perfect classification.

- **Cohen's Kappa coefficient [96], [97]:** It's a number that expresses how well two raters agree on a categorization task. It's described as

$$K = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

Here, p_o represents the observed percentage of agreement on any sample's label, and p_e shows the anticipated agreement if both raters give labels at random. p_e is calculated over the class labels with the help of a per-classifier empirical prior. The kappa value can vary from 0 to 1. A score of 0 indicates that the raters' agreement is random, whereas a score of 1 indicates that the raters' agreement is comprehensive. A score less than 0 indicates that there is less agreement than by chance [98].

- **AUC-ROC Curve [99]:** The "AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve" is utilized to visualize the performance of multi-class categorization. "AUC represents the degree or measure of separability, while ROC shows a probability curve. It indicates how well the model can discriminate among classes. The AUC indicates how well the model predicts "true as true and false as false". The ROC curve is drawn with the TPR on the y-axis and the FPR on the x-axis" as shown in figure 8.
An effective model has an AUC close to 1, indicating

TABLE IV. VARIOUS EVALUATION PARAMETERS

Performance Measure	Equation
Classification Accuracy	Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Precision=$\frac{TP}{TP+FP}$
Sensitivity	Sensitivity=$\frac{TP}{TP+FN}$
Specificity	Specificity= $\frac{TN}{TN+FP}$
False positive rate	False positive rate=$\frac{FP}{FP+TN}$
False negative rate	False negative rate=$\frac{FN}{FN+TP}$
F-Measure	f-measure=$2 \times \frac{precision \times recall}{precision+recall}$
Matthews correlation coefficient	MCC=$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$
Cohen's Kappa coefficient	$K = \frac{p_o - p_e}{1 - p_e}$

that it has a high level of separability. AUC around 0 indicates a bad model, which implies it has the lowest measure of separability. In reality, it indicates that the outcome is being reciprocated. "When AUC is 0.5, the model has no capacity for class separation".

5. Challenges

From the various research paper reviewed on crowd anomaly detection, various challenges faced by researchers are listed below:

1) Versatility of anomalies with different scenes, unpredictable crowd behaviors, *i.e.* movements and individual appearance
2) Due to significant occlusion between individual objects, clutters, complex actions, posture changes, low-resolution recordings with dynamic backgrounds, and unpredictable variations in the density of people over time, analyzing crowd scenes is extremely difficult.
3) The crowd counting and anomaly detection is a challenging task with the distribution of irregular objects or density distribution variation in an image or a video under normal and abnormal occurrences of anomalous events.
4) Abnormalities last for a limited period, non-uniform object scales or pixel distribution of the same object and variation in camera angles/position also affects the estimations of actual density, ground-truth, and its behavior.
5) Human crowds are complicated because they have both dynamic as well as psychological features, which are frequently goal-oriented. This makes determining an acceptable degree of granularity to represent crowd dynamics extremely difficult.
6) Selection of different datasets for training that covers all types of normal and abnormal crowd behavior, various motion patterns, features, density levels, with low computational cost etc. is a challenging task.
7) To define or clear characterization of the crowd through its differentiation like some form of social interaction, repeated behaviors among neighbors, movement of a large group of people, etc., and to take decision for the various crowded scenes for such scenario adds new challenges to the emerging research areas.

The above points are observed from the literature which helps and motivates researchers in the development of efficient methods for crowd anomaly detection with improving the present scenarios.

6. Conclusion

We have presented a review of the crowd anomaly and scene detection techniques based on "motion pattern segmentation, crowd behavior recognition, feature representation, density estimation, crowd dynamics, and anomaly detection for real-world applications". It is observed that the "anomaly detection" in visual occlusions, clutter, ambiguities, dense crowd, and scene semantics scenario requires more attention and for such crowd scene conditions, there is still no directly accepted solution. This paper reviews conventional and machine learning approaches like "HMM, GMM, Optical Flow, STT, etc. as well as deep learning approaches like CNN, LSTM, R-CNN, etc." with performance measures and image processing framework.

Further, the highest reported accuracy for crowd anomaly detection is 95% as achieved by Mustapha *et* al. However, the data used is a manual one and not the benchmark one. Upon the benchmark UCSD dataset, Singh *et* al. reported the highest accuracy of 92.7%.

7. Future Work

Many areas are still untouched in the area of crowded anomaly and scene analysis like multi-sensor information fusion of multi-camera system, tracking-learning-detection of the real-time surveillance video system, and real-time processing using deep learning and machine learning approach to improve accuracy rates.

Possible approaches for future research in crowd analysis are:

- Massive crowd motion analysis and behavior recognition.
- Creating annotators dedicated to massively ground truthing datasets that can be used for crowded scenes analysis tasks such as crowd behavior recognition, crowd tracking, and motion prediction.
- Represent the spatiotemporal relationships between the object's activities and visible changes.

- Extend Geographical Information System (GIS) in crowd analysis to accommodate better dynamic observations.

References

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2014.

[2] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International journal of computer vision*, vol. 98, no. 3, pp. 303–323, 2012.

[3] T. Surasak, I. Takahiro, C.-h. Cheng, C.-e. Wang, and P.-y. Sheng, "Histogram of oriented gradients for human detection in video," *2018 5th International conference on business and industrial research (ICBIR)*, pp. 172–176, 2018.

[4] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2871–2878, 2012.

[5] J. C. Silveira Jacques Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.

[6] A. E. Gunduz, C. Ongun, T. T. Temizel, and A. Temizel, "Density aware anomaly detection in crowded scenes," *IET Computer Vision*, vol. 10, no. 5, pp. 374–381, 2016.

[7] R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," pp. 439–452, 09 2010.

[8] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012.

[9] "UMN - Unusual crowd activity dataset of University of Minnesota." [Online]. Available: http://mha.cs.umn.edu/proj_events.shtml

[10] M. Rai, A. A. Husain, T. Maity, R. K. Yadav, and A. Neves, "Advance intelligent video surveillance system (aivss): a future aspect," *Intelligent Video Surveillance*, p. 37, 2019.

[11] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320312005055

[12] A. Ramchandran and A. Kumar, "Unsupervised deep learning system for local anomaly event detection in crowded scenes," *Multimedia Tools and Applications*, vol. 79, 12 2020.

[13] K. K. Santhosh, D. P. Dogra, P. P. Roy, and B. B. Chaudhuri, "Trajectory-based scene understanding using dirichlet process mixture model," *IEEE transactions on cybernetics*, vol. 51, no. 8, pp. 4148–4161, 2019.

[14] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Holistic features for real-time crowd behaviour anomaly detection," *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 918–922, 2016.

[15] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[16] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," *Springer*, pp. 161–169, 2018.

[17] J. Wang and Z. Xu, "Crowd anomaly detection for automated video surveillance," *IET Seminar Digest*, 2015.

[18] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "The large-scale crowd behavior perception based on spatio-temporal viscous fluid field," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1575–1589, 2013.

[19] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 548–561, 2015.

[20] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942, 2009.

[21] K. Khan, W. Albattah, R. U. Khan, A. M. Qamar, and D. Nayab, "Advances and trends in real time visual crowd analysis," *Sensors*, vol. 20, no. 18, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5073

[22] Y. Balasubramanian and C. Nagananthini, "Computer vision based crowd disaster avoidance system: A survey," *International Journal of Disaster Risk Reduction*, vol. 22, 03 2017.

[23] A. Shendarkar, K. Vasudevan, S. Lee, and Y.-J. Son, "Crowd simulation for emergency response using bdi agents based on immersive virtual reality," *Simulation Modelling Practice and Theory*, vol. 16, pp. 1415–1429, 10 2008.

[24] M. Carmona, "Principles for public space design, planning to do better," *Urban Design International*, vol. 24, no. 1, pp. 47–59, 2019.

[25] M. Flagg and J. M. Rehg, "Video-based crowd synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 11, pp. 1935–1947, 2013.

[26] A. Afiq, M. Zakariya, M. Saad, A. Nurfarzana, M. Khir, A. Fadzil, A. Jale, W. Gunawan, Z. Izuddin, and M. Faizari, "A review on classifying abnormal behavior in crowd scene," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 285–303, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320318303146

[27] J. Ma, Y. Dai, and K. Hirota, "A survey of video-based crowd anomaly detection in dense scenes," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 21, pp. 235–246, 03 2017.

[28] M. Thida, Y. Yong, P. Climent i Pérez, H.-L. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," *Intelligent Multimedia Surveillance: Current Trends and Research*, pp. 17–36, 11 2013.

[29] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[30] K. Pawar and V. Attar, "Deep learning approaches for video-based anomalous activity detection," *World Wide Web: Internet and Web Information Systems (WWW)*, vol. 22, 03 2019.

[31] A. Li, Z. Miao, and Y. Cen, "Global anomaly detection in crowded scenes based on optical flow saliency," *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, 2016.

[32] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, 2014.

[33] D. G. Lowe, "Object recognition from local scale-invariant features. int," *Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[34] X. Chong, W. Liu, P. Huang, and N. I. Badler, "Hierarchical crowd analysis and anomaly detection," *J. Vis. Lang. Comput.*, vol. 25, no. 4, p. 376–393, aug 2014. [Online]. Available: https://doi.org/10.1016/j.jvlc.2013.12.002

[35] X. Wang *et al.*, "Learning motion patterns using hierarchical bayesian models," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[36] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011, special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314210002390

[37] N. K. Ojha and A. Vaish, "Spatio-temporal anomaly detection in crowd movement using sift," *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 646–654, 2018.

[38] E. Andrade, S. Blunsden, and R. Fisher, "Modelling crowd scenes for event detection," *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, pp. 175–178, 2006.

[39] H. Yousefi, Z. Azimifar, and A. Nazemi, "Locally anomaly detection in crowded scenes using locality constrained linear coding," *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pp. 205–208, 2017.

[40] S. Li, C. Liu, and Y. Yang, "Anomaly detection based on maximum a posteriori," *Pattern Recognition Letters*, vol. 107, 09 2017.

[41] T. Bouwmans and E. H. Zahzah, "Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance," *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314213002294

[42] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[43] T. Qasim and N. Bhatti, "A hybrid swarm intelligence based approach for abnormal event detection in crowded environments," *Pattern Recognition Letters*, vol. 128, pp. 220–225, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865519302478

[44] P. Shelokar, V. Jayaraman, and B. Kulkarni, "An ant colony approach for clustering," *Analytica Chimica Acta*, vol. 509, no. 2, pp. 187–195, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267003016374

[45] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2153–2166, 2015.

[46] R. P. D. Redondo, C. Garcia-Rubio, A. F. Vilas, C. Campo, and A. Rodriguez-Carrion, "A hybrid analysis of lbsn data to early detect anomalies in crowd dynamics," *Future Generation Computer Systems*, vol. 109, pp. 83–94, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19309859

[47] T. Qasim and N. Bhatti, "A low dimensional descriptor for detection of anomalies in crowd videos," *Mathematics and Computers in Simulation*, vol. 166, pp. 245–252, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378475419301788

[48] H. Lin, J. D. Deng, and B. J. Woodford, "Anomaly detection in crowd scenes via online adaptive one-class support vector machines," *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 2434–2438, 2015.

[49] S. Mustapha, A. Kassir, K. Hassoun, Z. Dawy, and H. Abi-Rached, "Estimation of crowd flow and load on pedestrian bridges using machine learning with sensor fusion," *Automation in Construction*, vol. 112, p. 103092, 2020.

[50] Z. Ilyas, Z. Aziz, T. Qasim, N. Bhatti, and M. F. Hayat, "A hybrid deep network based approach for crowd anomaly detection," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 24 053–24 067, 2021.

[51] N. Priyadharsini and D. Chitra, "A kernel support vector machine based anomaly detection using spatio-temporal motion pattern models in extremely crowded scenes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 5225–5234, 2021.

[52] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *The Visual Computer*, vol. 36, no. 3, pp. 609–620, 2020.

[53] Z. Aziz, N. Bhatti, H. Mahmood, and M. Zia, "Video anomaly detection and localization based on appearance and motion models," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25 875–25 895, 2021.

[54] D. Domínguez, R. Díaz Redondo, A. Vilas, and M. Ben Khalifa, "Sensing the city with instagram: Clustering geolocated data for outlier detection," *Expert Systems with Applications*, vol. 78, 02 2017.

[55] M. Ben Khalifa, R. Díaz Redondo, A. Vilas, and S. Servia-Rodríguez, "Identifying urban crowds using geo-located social media data: a twitter experiment in new york city," *Journal of Intelligent Information Systems*, vol. 48, 04 2017.

[56] X. Kong, H. Gao, O. Alfarraj, Q. Ni, C. Zheng, and G. Shen, "Huad: Hierarchical urban anomaly detection based on spatio-temporal data," *IEEE Access*, vol. 8, pp. 26 573–26 582, 2020.

[57] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[58] X. Kong, X. Liu, B. Jedari, M. Li, L. Wan, and F. Xia, "Mobile

crowdsourcing in smart cities: Technologies, applications, and future challenges," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8095–8113, 2019.

[59] B. Murugan, M. Elhoseny, K. Shankar, and J. Uthayakumar, "Region-based scalable smart system for anomaly detection in pedestrian walkways," *Computers Electrical Engineering*, vol. 75, pp. 146–160, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790618331847

[60] D. Chen, P. Wang, L. Yue, Y. Zhang, and T. Jia, "Anomaly detection in surveillance video based on bidirectional prediction," *Image and Vision Computing*, vol. 98, p. 103915, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885620300470

[61] 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, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121931197X

[62] 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, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314218300249

[63] N. Li and F. Chang, "Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219311828

[64] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1896–1904, 2019.

[65] S. Lee, H. G. Kim, and Y. M. Ro, "Stan: Spatio-temporal adversarial networks for abnormal event detection," *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1323–1327, 2018.

[66] F. L. Sánchez, I. Hupont, S. Tabik, and F. Herrera, "Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects," *Information Fusion*, vol. 64, pp. 318–335, 2020.

[67] J. Arunnehru *et al.*, "Deep learning-based real-world object detection and improved anomaly detection for surveillance videos," *Materials Today: Proceedings*, 2021.

[68] M. Sabih and D. K. Vishwakarma, "Crowd anomaly detection with lstms using optical features and domain knowledge for improved inferring," *The Visual Computer*, pp. 1–12, 2021.

[69] I. V. Pustokhina, D. A. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar, and K. Shankar, "An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety," *Safety Science*, vol. 142, October 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753521002009

[70] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220317720

[71] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

[72] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, pp. 320–329, 03 2012.

[73] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, 2017.

[74] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 7354–7363, 09–15 Jun 2019. [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html

[75] "Center for Research in Computer Vision, University of Central Florida (UCF)." [Online]. Available: https://www.crcv.ucf.edu/data/crowd.php

[76] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, "Real time, online detection of abandoned objects in public areas," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2006, pp. 3775 – 3780, 06 2006.

[77] N. V. Weixin Li, Vijay Mahadevan, "UCSD Anomaly Detection Dataset from University of California at San Diego (UCSD), USA," 2010. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[78] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," *2013 IEEE International Conference on Computer Vision*, pp. 2720–2727, 2013.

[79] "CUHK Avenue Dataset for abnormal event detection of The Chinese Univeristy of Hong Kong," 2013. [Online]. Available: http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

[80] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, 2012.

[81] L. B H, M. Aradhya, and D. Guru, "Violent video event detection based on integrated lbp and glcm texture features," *Revue d'Intelligence Artificielle*, vol. 34, pp. 179–187, 05 2020.

[82] "Violent-Flows - Crowd Violence Non-violence Database and benchmark," 2014. [Online]. Available: https://www.openu.ac.il/home/hassner/data/violentflows

[83] J. L. Russell, David, "QMUL Junction Dataset of Queen Mary University of London," 2012. [Online]. Available: https://personal.ie.cuhk.edu.hk/~ccloy/downloads_qmul_junction.html

[84] C. C. Loy, T. Xiang, and S. Gong, "From local temporal correlation to global anomaly detection," *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*, 2008.

[85] C. C. Loy, T. Xiang, and S. Gong, "Stream-based active unusual event detection," *Asian Conference on Computer Vision*, pp. 161–175, 2010.

[86] M. George, B. R. Jose, J. Mathew, and P. Kokare, "Autoencoder-based abnormal activity detection using parallelepiped spatio-temporal region," *IET Computer Vision*, vol. 13, no. 1, pp. 23–30, 2019.

[87] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," *arXiv preprint arXiv:1612.00390*, 2016.

[88] S. A. Velastin and D. A. Gómez-Lira, "People detection and pose classification inside a moving train using computer vision," *International Visual Informatics Conference*, pp. 319–330, 2017.

[89] V. M. Arceda, K. F. Fabián, P. L. Laura, J. R. Tito, and J. G. Cáceres, "Fast face detection in violent video scenes," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 5–26, 2016.

[90] S. Narkhede, "Understanding Confusion Matrix," 2018. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[91] A. Ratre and V. Pankajakshan, "Tucker tensor decomposition-based tracking and gaussian mixture model for anomaly localisation and detection in surveillance videos," *IET Computer Vision*, vol. 12, no. 6, pp. 933–940, 2018.

[92] S. Minaee, "20 Popular Machine Learning Metrics. Part 1: Classification Regression Evaluation Metrics," 2019. [Online]. Available: https://tinyurl.com/nbkvaf62

[93] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," 2018. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[94] T. Wood, "F-Score." [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score

[95] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[96] "Cohen's kappa score." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[97] H. Li and Q.-D. Phung, "Journal of machine learning research: Preface," *Journal of machine learning research: JMLR*, vol. 39, no. 2014, pp. i–ii, 2014.

[98] K. Pykes, "Cohen's kappa," 2020. [Online]. Available: https://towardsdatascience.com/cohens-kappa-9786ceceab58

[99] S. Narkhede, "Understanding AUC - ROC Curve," 2018. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Harshadkumar S. Modi Harshadkumar S. Modi has received his Bachelor of Engineering degree in Computer Engineering from U.V.Patel College of Engineering, Hemachandracharya North Gujarat University, Gujarat, India in 2006, and Master of Engineering degree in Computer Engineering from Government Engineering College, Gandhinagar, Gujarat, India in 2018. Previously, he worked as an Assistant Professor with Information Technology Department, Ganpat University, Gujarat, India. He is currently working as a Lecturer with Computer Engineering Department, Government Polytechnic Gandhinagar, affiliated with Gujarat Technological University Gujarat, India and has a teaching experience of more than 15 years. His research interest includes Computer Vision, Deep learning applications, Crowd Behavior Analysis and Visual Tracking.

Prof. (Dr.) Dhaval A Parikh Prof. (Dr.) Dhaval A Parikh has completed his bachelor degree from Gujarat University, Gujarat, India in 1991, master degree from Sardar Patel University, Gujarat, India in 2003 and PhD from C. U. Shah University, Gujarat, India in 2017. He has a teaching experience of more than 22 years. Currently he is working as a Professor and Head of the Department with Computer Engineering department, Government Engineering College, Gandhinagar, affiliated with Gujarat Technological UniversityGujarat, India.