Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

A comprehensive review on deep learning-based methods for video anomaly detection

Rashmiranjan Nayak *, Umesh Chandra Pati, Santos Kumar Das

Department of Electronics and Communication Engineering, National Institute of Technology, Rourkela, Odisha 769008, India

ARTICLE INFO

Article history:
Received 28 August 2020
Received in revised form 12 November 2020
Accepted 25 November 2020
Available online 30 November 2020

Keywords:
Deep learning
Deep regenerative models
Deep one-class models
Hybrid models
Spatiotemporal models
Video anomaly detection

ABSTRACT

Video surveillance systems are popular and used in public places such as market places, shopping malls, hospitals, banks, streets, education institutions, city administrative offices, and smart cities to enhance the safety of public lives and assets. Most of the time, the timely and accurate detection of video anomalies is the main objective of security applications. The video anomalies such as anomalous activities and anomalous entities are defined as the abnormal or irregular patterns present in the video that do not conform to the normal trained patterns. Anomalous activities such as fighting, riots, traffic rule violations, and stampede as well as anomalous entities such as weapons at the sensitive place and abandoned luggage should be detected automatically in time. However, the detection of video anomalies is challenging due to the ambiguous nature of the anomaly, various environmental conditions, the complex nature of human behaviors, and the lack of proper datasets. There are only a few dedicated surveys related to deep learning-based video anomaly detection as the research domain is in its early stages. However, state of the art lacks a review that provides a comprehensive study covering all the aspects such as definitions, classifications, modelings, performance evaluation methodologies, open and trending research challenges of video anomaly detection. Hence, in this survey, we present a comprehensive study of the deep learning-based methods reported in state of the art to detect the video anomalies. Further, we discuss the comparative analysis of the state of the art methods in terms of datasets, computational infrastructure, and performance metrics for both quantitative and qualitative analyses. Finally, we outline the challenges and promising directions for further research.

Contents

* Corresponding author.
E-mail addresses: rashmiranjan.et@gmail.com, rashmiranjan_nayak@nitrkl.ac.in (R. Nayak), ucpati@nitrkl.ac.in (U.C. Pati), dassk@nitrkl.ac.in (S.K. Das).

https://doi.org/10.1016/j.imavis.2020.104078

1. Introduction

An Intelligent Video Surveillance Systems (IVSS) is capable of detecting the anomalous activities such as crimes, fighting, road accidents, riots, and stampede as well as the anomalous entities such as weapons at sensitive places, and abandoned objects automatically in real-time. When an observation significantly deviates from the other observations in the same context and evoke an intuition that a different mechanism generates it, then that particular observation can be called as anomaly corresponding to the specific environment [1]. In other words, outliers diverging significantly from the trained model can be considered as the anomalous events [2]. The various synonyms used for the anomalies in literature are abnormalities, deviants, outliers, and unusualness.

Interpretation about "What is an abnormality in a scene to a particular context and time?" leads to the video anomaly detection. Most of the time, it can be assumed that for any specific context in a scene, there is a notion of what constitutes normal or regular activity and conversely, abnormal or anomalous activity. For an example, while it is "normal" to observe crowds in markets in usual days, such type of crowd gathering is treated as "anomalous or abnormal" if it is found in the same market during curfews. Hence, video anomalies are context and subject dependent. Subsequently, an anomalous activity can be defined as "an activity executed at an unusual location, at an unusual time," or "activities that are fundamentally different in appearance and motion" [3,4]. In case of the IVSS, the process of automatic identification of the abnormal video patterns such as anomalous activities or entities in the spatiotemporal dimensions is known as video anomaly detection.

Further, the video anomalies are equivocal, new, unknown, rare (less frequent), irregular, unexpected, atypical [5] and out-of-the-dictionary in nature [3,6–8]. Video anomalies can be localized or distributed in spatiotemporally in complex scenarios. These typical characteristics of uncertainties further complicate the modeling process for video anomaly detection.

Video anomaly detection and video anomaly localization are very much interrelated. Video anomaly detection focuses on finding whether the given video frame exhibits an anomaly or not by addressing the question "Does the given frame contain an anomaly or not?". Video anomaly localization focuses on the localization of anomalies by determining the actual location of the anomalies in the given video frame with the help of the bounding box, corresponding to the question, "Where is anomaly occurring in the given frame?" [9]. The recent trend in state of the art suggests that deep learning-based methods take care of both video anomaly detection and localization in a single end-to-end pipeline.

The research areas of the video anomaly detection and human activity detection or recognition are closely related, not the same. Anomaly detection can be defined as an unsupervised learning technique meant for identifying the abnormal patterns or trends present in the data [10]. Generally, human activity recognition is a supervised learning technique used to classify various human activities. Broadly, the video anomaly detection is different from human activity recognition as well as other supervised video analysis problems such as action recognition, event detection, object detection, etc. in three crucial aspects. Firstly, the video data is inherently unbalanced one between the positive and negative classes, i.e., generally, the positive examples (anomalous events) are fewer than the negative examples (regular events). Secondly, the high variance within the positive classes as anomalous events may contain a large variety of different classes [2], and it deters the applicability of supervised learning-based techniques for the detection of video anomalies. Thirdly, the classes of activities are well-defined during human activity recognition. However, video anomalies are vaguely defined due to its equivocal nature and may cover a large variety of activities [11].

Moreover, there are various domain challenges such as environmental conditions (illumination variations, shadow effects of the objects, object occlusions, complex background, and so forth.), crowd density, noisy data, complex nature of human behaviors, recording camera setting, spatiotemporal variations, difficulty in accessing good computational infrastructure, the trade-off between intra-class and inter-class variations. In addition to the inherent challenges (ambiguous nature, data imbalance problem, and the high variance within the positive classes) of the video anomaly detection, these domain challenges make the video anomaly detection as one of the tedious computer vision tasks.

Further, video anomalies can be detected by using stationary (or fixed) surveillance cameras (e.g., CCTV cameras installed in smart cities) and dynamic (or moving) cameras (e.g., vehicle dashboard cameras). However, there is scanty research for the video anomaly detection using video streams form the moving cameras. Hence, this survey will be limited to the deep learning-based video anomaly detection methods corresponding to only the fixed surveillance camera networks.

In this section, we first discuss the related surveys. Then, we outline the contributions made in this work and organization of the article.

1.1. Related surveys

There are only fewer deep-learning approaches that have been reported for video anomaly detection despite the substantial advancements achieved by deep-learning methods in various other domains. Hence, there are several existing surveys related to video-based human activity recognition; however, only a few dedicated surveys related to the deep learning-based video anomaly detection methods such as anomalous activity detection or anomalous entity detection as presented in Table 1. However, there is no single survey that provides an inside-out study covering all the aspects such as definitions, classifications, modelings, performance evaluation methodologies, open and trending research challenges of video anomaly detection to our best of knowledge. Hence, there is a requirement for a comprehensive survey on the deep-learning-based video anomaly detection methods from the foundation to advanced levels. This is one of the essential contributions of the present survey.

Table 1
Summary of the recent important related surveys.

Year	Reported paper	Main focus/contribution
2010	Ko [12]	Challenges to automatically detect abnormal behavior
2010	Poppe [13]	Challenges in visual surveillance, video retreival and human-computer interaction
2010	Candamo et al. [14]	Understanding abnromal activities such as loitering at transit scenes
2011	Weinland et al. [15]	Methods for action representation, segmnetation and recognition
2011	Weinland [15]	Representation, segmentation and recognition of human actions using vision-based techniques
2011	Buch et al. [16]	Computer vsison techniques for analysis of urban traffic
2012	Popoola and Wang [3]	Video-based contextual abnormal human behavior detection
2012	Sodemann et al. [17]	Anomaly detection in automated surveillance
2013	Bengio et al. [18]	Representation learning
2013	Vishwakarma and Agrawal [19]	General frameworks for human activity recognition
2013	Wang [20]	IVSS using multi-camera network
2014	Gowsikhaa et al. [21]	Semantically enhanced analysis for human behavior
2014	Pimentel et al. [22]	Techniques for novelity detection
2015	Chong and Tay [23]	Challenges involved in modeling for video anomaly detection
2018	Pawar and Attar [9]	Deep learning approaches such as CNN, LSTM, AEs, etc.
2018	Tripathi et al. [24]	Methods for recognition of sucpicious or abnormal human activities
2018	Ahmed et al. [25]	Video surveillance using object trajectories
2018	Kiran et al. [10]	Unsupervised and semi-supervised learning based video anomaly detection
2018	Mabrouk and Zagrouba [26]	Feature extraction and description techniques for abnormal behavior recognition
2019	Chapapathy and Chawla [27]	Deep-learning-based anomaly detection techniques for various domains
2019	Santosh et al. [28]	Video anomaly detection in road traffic

1.2. Contributions

The present survey is initially build on few recent surveys specific to video-based anomaly detection [9,10,24,27,29]. Subsequently, the survey presents a detailed and structured analysis of video anomaly detection using deep learning methods, specifically focused on state of the art reported in last decade. The main contributions of the research work can be outlined as follows.

- A structured, comprehensive review of the state of the art research in video anomaly detection using deep-learning methods both for the accuracy and real-time processing approaches is presented.
- A graphical taxonomy for the video anomaly detection using deep-learning methods has been put forth.
- A comparative analysis of the deep learning methods for anomaly detection has been put forth so that a researcher can easily decide which method will be potentially more suitable for the corresponding application.
- A thorough analysis of the performance evaluation methodologies in terms of datasets, computational infrastructure, various evaluation criteria, performance metrics for both quantitative and qualitative analysis is presented.
- A short overview regarding uncovering the decision-making process using explainable deep learning is introduced in the qualitative analysis section.

- The open and trending research challenges of the video anomaly detection methods have been explored.

1.3. Organization

The rest of the article is organized as follows. Classification of video anomalies is presented in Section 2. Training and learning frameworks are briefly explained in Section 3. Section 4 deals with various approaches that are used for video anomaly detection. Problem formulation for the modeling of the video anomalies is outlined in Section 5. Further, a structured and comprehensive review of the state-of-the-art involving deep learning-based methods for video anomaly detection is put forth in Section 6. A comparative analysis of deep learning-based methods used for the detection and localization of the video anomalies is presented in Section 7. An in-depth analysis of performance evaluation methodologies in terms of bench-marked datasets, computational infrastructure, evaluation criteria, and performance metrics for quantitative as well as qualitative analysis is presented in Section 8. Section 9 offers the research challenges present in deep learning-based methods for video anomaly detection. Finally, the conclusions and potential applications are presented in Section 10.

2. Classification of video anomalies

Video anomaly detection and localization mainly depend on the two factors such as the complexity of the environment and types of the anomalies. Further, the complexity of the environment is dependent on the density of the moving targets involved in the scene. Hence, based on the density of the moving targets, the environment can be classified into three types such as sparsely crowded environment (loose crowd, i.e., 10 sqft/person), moderately crowded environment (more dense crowd, i.e., 4.5 sqft/person), densely crowded environment (very dense crowd, i.e., 2.5 sqft/person) [30]. Video-based anomalous activity detection involving single entity such as loitering detection, intrusion detection, etc. and involving two persons or interactions such as fighting can be considered as the cases of sparsely crowded environment. Video anomaly occurring in groups such as violence, riots, and so forth are considered as the cases of the moderately crowded environment. Finally, video anomaly occurring in dense crowd situations such as stampede, crowd dispersion due to explosion, and so on can be considered as the cases of densely crowded environment. Many times this type of anomalies are also called crowd anomaly. Further, the video anomalies may coexist in a single scene. Broadly, video anomalies may be classified as follows.

2.1. Local and global anomalies

Local anomaly or local anomalous activity significantly deviates from its neighboring spatiotemporal activities, e.g., a car moving in the wrong direction [9,27,31–33]. Here, the behavior of the individuals significantly differs from the that of the neighbors [34]. Inversely, global anomaly or global anomalous activity is referred to activities that interact with each other globally in an abnormal, suspicious or unusual manner, even if individual activities may be normal or anomalous in isolation [9]. Global anomalies may be considered as collective or group anomalies. Collective or group anomalies are corresponding to a set of individual data points in which the individual samples separately behaves as normal data instances and collectively in a group exhibit abnormal behaviors [27]. Various examples of global anomalies are an irregular mixture of image pixels causing an unusual distortion in an image, car accidents, crowd dispersion due to bomb explosion, etc.

2.2. Point and interaction anomalies

Point anomaly is corresponding to the data points, which significantly deviate from the rest of the data points. In other words, the point anomaly corresponds to a random irregularity that can be further mapped into an anomalous activity exhibited by an individual, e.g., loitering [9,27]. The interaction between individual entities in an unusual way can be treated as an interaction anomaly, e.g., fighting between two persons [9,27].

2.3. Contextual or conditional anomalies

These are corresponding to the data points having significant deviation causing anomalies with respect to a specific context [9,27]. The context is identified by contextual features and behavioral features. Normally, time as well as space are considered as contextual features, and features used to describe normal behaviors are regarded as behavioral features. Most of the video anomalies such as fighting, riots, stampede, and so forth are falls in this category. Further, contextual anomalies may be classified as spatial and temporal anomalies [9,35,36]. However, usually contextual anomalies are best described by the spatiotemporal anomalies [37].

Generally, both space and time complexity required for the detection and localization of the video anomaly from the sparsely crowded environment to the densely crowded environment through a moderately crowded environment do increase, as well depicted in Fig. 1.

3. Training and learning frameworks

Generally, the training and learning frameworks of the deep learning methods used for video anomaly detection are classified into four categories: supervised, unsupervised, semi-supervised and active learning [3,27]. This classification is based on the utilization of human intervention and the amount of prior knowledge (labels) during the training process.

3.1. Supervised video anomaly detection

Supervised video anomaly detection methods involve the training of a binary classifier using the associated labels of normal and anomalous activities. Here, the anomalous activities must be clearly defined and the training dataset should be a balanced one. However, practically in most of the scenarios, it is not feasible to clearly define the video anomalies with the associated labels due to the equivocal nature, the rare occurrence, evolutionary quality, the high variance within positive samples, and the inherent data-imbalance problem [27] of the video anomalies. Hence, the recent trend shows the vast popularity of the data-driven approaches using unsupervised and semi-supervised training processes over the supervised video anomaly detection methods.

3.2. Unsupervised video anomaly detection

Unsupervised video anomaly detection methods detect the anomalous activities or entities using the co-occurrence statistical concepts from the unlabeled video data [3]. These methods are used in the auto-labeling of the unlabeled video data [27]. A general framework for the video anomaly detection using the clustering of low-level features such as color, Scale-Invariant Feature Transform (SIFT), velocity, etc. is proposed [38]. However, this technique provides more false positives in crowded scenes due to its dependency on trajectory-based classification. Further, the effectiveness of the unsupervised video anomaly detection methods mostly requires the large size of video dataset and massive computational resources. With the ease of availability of these two entities, unsupervised techniques have been outperforming the supervised video anomaly detection methods. In contrast, there is a lack of labeled data for the anomalous activities only, whereas massive

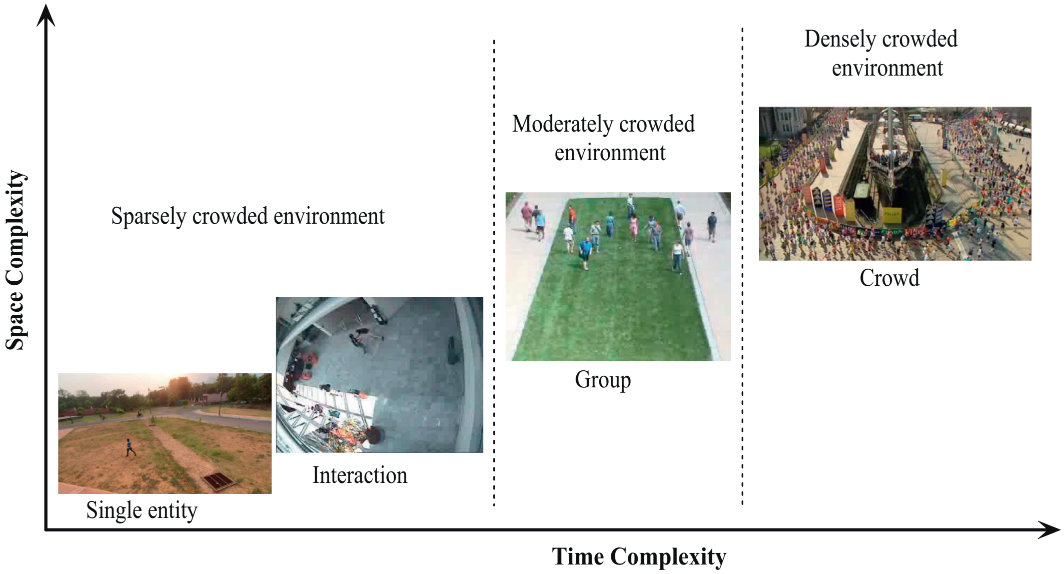

Space Complexity
Time Complexity
Sparsely crowded environment
Moderately crowded environment
Densely crowded environment
Single entity
Interaction
Group
Crowd

Fig. 1. Types of anomaly based on crowd density.

weakly-labeled normal data are available. However, the unsupervised video anomaly detection techniques do not merely utilize the full potential of this weakly-labeled normal data.

3.3. Semi-supervised video anomaly detection

Semi-supervised video anomaly detection methods detect anomalous activities or entities by training the model using only weakly-labeled normal instances of video. This type of video anomaly detection method is more widely used as it uses the benefits of both supervised and unsupervised techniques. Most of the time, the unsupervised video anomaly detection may be dealt with as semi-supervised video anomaly detection techniques due to the availability of the normal video or data with no anomalies [10]. Mostly, the semi-supervised video anomaly detection methods are modeled using deep autoencoders. Recently, deep-autoencoder based models are trained with sufficient training data comprising only normal events so that it produces minimal reconstruction error for the normal activities [27]. Conversely, the same model produces a high reconstruction error for the anomalous activities and hence, detects the video anomalies. However, the efficiency of the semi-supervised video anomaly detection methods can be further enhanced by keeping a domain expert in the loop.

3.4. Active learning-based video anomaly detection

In the case of unsupervised or semi-supervised video anomaly detection, the anomaly detection model is trained offline with normal training examples and not updated as and when new data has arrived. Hence, this results in ineffective video representations [39]. These issues may be addressed by using Active learning-based video anomaly detection, where the human (or domain experts) are kept in the loop for labeling the confusing decisions or samples in the online framework. Hence, active learning helps in minimizing the ambiguity nature of anomaly by introducing appropriate priors with the help of a domain expert. Recently, a deep active learning method suitable for unsupervised deep learning-based anomaly detection model is proposed [40]. Further, the generative feature of the Generative Adversarial Network (GAN) empowered with active learning is successfully applied for the detection of outliers [41]. Though active learning-based video anomaly detection methods are suitable for better accuracy in online applications, the model requires continuous intervention by the domain experts.

4. Approaches for video anomaly detection

Based on the desired detection accuracy and processing-time, there are two main approaches for the video anomaly detection: accuracy-oriented approaches and processing-time-oriented approaches [42].

4.1. Accuracy-oriented approaches

Accuracy-oriented (AO) approaches for video anomaly detection are aiming at detection and localization of video anomalies with high accuracy and low false alarms. Complex models trained with more number of attributes are used to achieve the desired high accuracy at the expense of high processing-time [42]. The objective of this category to make the video anomaly detection methods suitable for offline applications by using all the available training video datasets, fixed model parameters, and pre-defined or fine-tuned anomaly thresholds [27]. There have been significant contributions in this category over past decades for video-based anomalous activity detection. Some of the important research works are based on generative models [43,44], temporal regularity model [45], spatio-temporal / predictive models [46], hybrid models [2], and models using detection of the STIPs [47,48].

4.2. Processing-time-oriented approaches

Processing-time-oriented (PO) approaches for video anomaly detection are aiming at the detection and localization of video anomalies with minimum frame processing time and a competitive level of accuracy. The objective of this category to make the video anomaly detection methods suitable for real-time applications by attaining online performances such as high computational speed and less computation space [42]. Practically, the time required to process the current frame should be shorter than the inter-frame interval to attain the online performances. Hence, it is always desirable to use compact and robust features that require less computational time. This type of approach continuously updates the models by incrementally changing the model parameters based on the new training samples [27]. Complex models trained with more number of attributes are used to achieve the desired high accuracy at the expense of high processing-time [42]. Few low-level lightweight feature descriptors such as binary [49] and binary pair-based [50] video descriptors are used for video anomaly detection with online performances. However, these techniques are lack providing the abstract and complex information of human action. Subsequently, the deep learning models are based on Histograms of Optical Flow Orientation and Magnitude (HOFM) as well as lightweight convolutional LSTM autoencoder [51] and context online learning scheme [52] are used to detect the abnormal or suspicious behavior detection. Though the model is suitable for online applications, it lacks the descriptive power of the DNN [52]. Further, there have been significant contributions in this category over the past decades to detect video anomalies. Few important research works are based on generative models [53–55], temporal regularity model [56], spatio-temporal / predictive models [57,58], hybrid models, cell-based models [42,59–62].

There is a demand for an optimistic trade-off between detection accuracy and processing time for effective video anomaly detection specific to a particular surveillance application. However, the major challenge is to achieve this optimistic trade-off by developing a few highly descriptive features for achieving desired online performance and detection accuracy.

5. Problem formulation for the modeling of the video anomalies

The primary objective of the modeling is always to immediately detect the video anomaly with its occurrence with minimum false alarms. Due to the equivocal nature of the anomaly, availability of huge data, and computational resources, data-driven statistical methods are preferred over the rule-based methods to model semantically meaningful scene behaviors [3]. In the case of video anomaly detection, the input data, i.e., video, is sequential in nature. Further, video is also considered as one of the high-dimensional data due to the involvement of several features. Hence, problem formulation is one of the crucial task in video anomaly detection. Once the target application and available resources are identified, the desired level of performance metrics such as expected accuracy and processing time are fixed. Subsequently, the problem formulation for the video anomaly detection can be carried out as follows.

- Assumptions: The frames corresponding to the anomalous activities are rare in occurrence with significant deviation in appearance (spatial properties) and motion (temporal properties).
- Inputs: Bench-marked video datasets covering different possible environmental challenges such as occlusion, variation in illumination as well as resolution, and different camera views are available for training. The training data sets should at least contain both normal samples (negative cases) for training and few anomalous samples (positive cases) for validation. A frame sequence of the training video segment, $X_{train} \in R^{N_{train} \times r \times c}$ comprises only normal motion (temporal) patterns as well as normal appearance (spatial) patterns and no anomalies. Further, a frame sequence of the testing video segment, $X_{test} \in R^{N_{test} \times r \times c}$ comprises both normal as well as anomalous frames. Here, N_{train} and N_{test} represent the number of frames present in training as well as testing data sets respectively. Further, the dimensionality of the each representation vector is represented by $d = r \times c$.
- Objectives: The critical task is twofold, i.e., video anomaly detection as well as localization in the spatiotemporal domain. Video anomaly detection involves the process of assigning an anomaly score to each frame corresponding to the temporal variation and find the anomalous frames having anomaly scores above the set threshold. Subsequently, video anomaly localization is the process of assigning a spatial score to each frame to localize the anomalous spatial region in the particular anomalous frames.
- Methods: Selection of appropriate deep learning-based modeling methods to detect video anomalies for an intended application is one important step. A systematic study of various important deep-learning based methods for the video anomaly detection will be discussed in Section 6.
- Field trails: Once the model for video anomaly detection and localization is developed, it can be validated practically both in offline mode (test samples are the stored video) and online mode (test samples are the live streaming video).

6. Deep-learning based methods for the video anomaly detection

Broadly, most of the current research work on video anomaly detection can be grouped into two steps, such as feature extraction and normal distribution learning [63]. Feature extraction is achieved either by hand-crafted techniques or auto-feature extraction techniques (representation learning or deep learning-based features). In normal distribution learning, a distribution is learned using only available normal samples of the training data. Subsequently, any test data samples containing the video anomalies will significantly deviate from the learned normal distribution and results in high reconstruction error.

The state-of-the-art of the deep-learning-based methods used for the detection and localization of the video anomalies can be categorized into 1) Trajectory-based methods, 2) Global pattern-based methods, 3) Grid pattern-based methods, 4) Representation learning models, 5) Discriminative models, 6) Predictive models, 7) Deep generative models, 8) Deep one-class deep neural networks, and 9) Deep hybrid models. These methods are depicted in a graphical taxonomy as shown in Fig. 2.

6.1. Trajectory-based methods

In the case of trajectory-based methods, firstly, the object or subject of interest is detected and then tracked across the frames to generate the object trajectory. The anomalous activities performed by the objects are inferred from the analysis of the generated object trajectory [64,65]. Few important reported works for anomalous activity detection using trajectory analysis are based on One-Class Support Vector Machine (OC-SVM) clustering [66], snapped trajectory [67], semantic scene learning as well as tracking [68], string kernel-based clustering [69], Hidden Markov Model (HMM)-based prediction from the time-series sequence of various features such as foot tracking and mean optical flow [70], deep learning-based classifiers [71], spatio-temporal path search [72], You Look Only Once (YOLOv2) [73,74] as well as Simple Online and Real-time Tracking (SORT) [75] for intrusion detection [76], You Look Only Once (YOLOv3) [77] as well as Deep Simple Online and Real-time Tracking (Deep-SORT) [78] for loitering detection [79,80], and anomaly detection in road traffic [28]. The effectiveness of trajectory-based methods significantly depends on object detection as well as tracking. Generally, tracking performances are affected by crowd density, low video resolution, sudden motion, and occlusion. Hence, the trajectory methods as found to be suitable for the sparsely crowded environment, not for the moderately or densely crowded environment. Further, these methods lack contextual information while deciding the abnormality.

6.2. Global pattern-based methods

In the case of global pattern-based methods, the frame sequences are analyzed as a whole entity by the computationally efficient low or meduim level features, extracted from the video [3,65]. These methods are found to be effective for the moderately crowded as well as densely crowded environments due to the absence of detection and tracking of indvidual object. Features such as spatial temporal gradients, kinetic energy, optical flow, etc. are widely used in these methods. A social attribute-aware force model based on the social characteristics of crowd behaviors is used to detect abnromal event [81]. Statistical hypothesis test as well as approximation of complex noise distribution using Gaussian Noise Model are used to detect anomalous event [82].

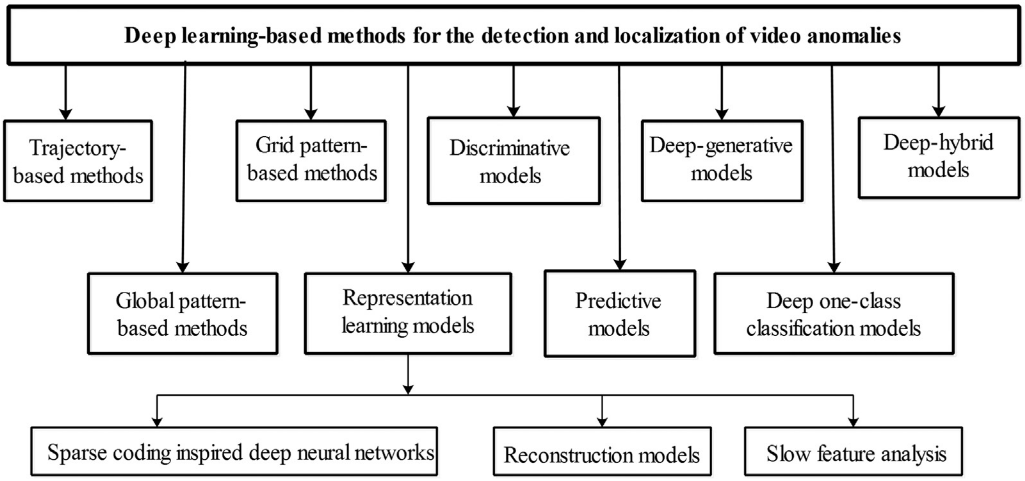

Deep learning-based methods for the detection and localization of video anomalies
Trajectory-based methods
Grid pattern-based methods
Discriminative models
Deep-generative models
Deep-hybrid models
Global pattern-based methods
Representation learning models
Predictive models
Deep one-class classification models
Sparse coding inspired deep neural networks
Reconstruction models
Slow feature analysis

Fig. 2. Classification of deep learning-based methods for the detection and localization of video anomalies.

A model based on the kinetic energy computed with the help of optical flow and crowd distribution index is proposed to anomalous crowd behavior [83]. Various models based on global motion map [84], stationary map [85], salient motion map [86], motion influence map [87], etc. are used to detect anomalous crowd activities. Further, a model based on both hierarchical feature representation and Gaussian process regression can simultaneously detect as well as localize the video anomaly [47]. These methods are helpful in the detection of video anomaly. However, the localization of the anomalies using global pattern-based methods is a tedious job.

6.3. Grid pattern-based methods

The grid pattern or cell-based methods are used to reduce the processing times by limiting the number of extracted features from the fixed spatiotemporal regions, i.e., anomalies of each grid or cells or sub-regions are evaluated separately, ignoring the connections among the objects [42,65]. More over, patterns are extracted from the splitted blocks of the frames instead of treating the frames as single entities [19,65]. Further, these techniques don't require STIPs and saliency detection [42]. Various approaches such as sparse combination learning [88], a probabilistic framework using local statistical aggregates [33], mixtures of dynamic texture comprise temporal anomaly detection using Gaussian Mixture Model (GMM) based background subtraction and spatial anomaly detection using saliency detection [32], etc. have been used to evaluate the extracted features or grid patterns for video anomaly detection.

Similarly, joint detection of spatial and temporal anomalies using spatial and temporal anomaly maps are evaluated for the crowded scenes [89]. Further, a cell-based analysis using speed, texture, and size of objects has been reported for the anomaly detection in crowded techniques [61]. The grid pattern based methods work in the spirit of local-region based methods where anomaly is detected using the probabilistic model that is capable of detecting the local spatiotemporal variation. A set of compact features based on Gaussian mixture models (GMMs), Markov Chains, and Bag-of-words are extracted from the grid with variable sized cells and offered online performances [42]. Recently, deep learning techniques such as deep Gaussian mixture model [90], spatio-temporal convolution neural networks [91], etc. are also applied to the grid-based methods for the video anomaly detection.

6.4. Representation learning models

The process of extracting useful features or learning good representations of the input video data by taking into account crucial prior information of the particular problem is known as representation learning. Further, representation learning helps in the reduction of computational complexity by converting a very high dimensional video data into important d-dimensional (low-dimensional) vectors and subsequently helps in eliminating the curse of dimensionality problem [18,53]. Models based on the representation learning can be used for building classifiers as well as predictors for various tasks such as action recognition, anomaly detection, and object detection. Moreover, the representation learning follows the principle of the No-Free-Lunch-Theorem, i.e., there exists no universal learner suitable for every training distribution [10]. Further, representation learning is helpful in extracting an effective video descriptor or representation that are generic (generalization ability of the descriptors over varieties of the videos), compact (requires less memory), efficient (faster processing), and simple (ease of implementation) [92]. Important methods of video anomaly detection based on representation learning can be briefly explained as follows.

6.4.1. Sparse coding inspired deep neural networks

Normal distribution learning is one of the widely used techniques for anomaly detection. Here, the model learns from the training video data sets comprises only normal events. Subsequently, the anomalous events do not conform to the trained model during inference. Based on this strategy, sparse coding based video anomaly detection is found to be promising one [46,93,94]. The basic assumption of sparse coding based video anomaly detection is that sparse linear combinations of normal patterns are capable of representing the normal activities with minimal reconstruction error and that of anomalous patterns with large reconstruction errors as anomalous activities are not present in the training data sets [46,65]. In other words, the sparse coding-based video anomaly detection model initially learns a dictionary from the training video data set comprising only normal activities and subsequently finds the anomalous activities that can not be successfully reconstructed by the atoms of the learned dictionary [46].

During the past decades, well-established features such as Histogram of Oriented Gradients (HOG) [95], Histogram of Oriented Flows (HOF) [96], 3D spatio-temporal gradient [97], sparse combination learning [88], Low-rank and Sparse Decomposition (LSD) [98] with the help of feature learning and sparse coding have been implemented successfully in sparse coding based video anomaly detection [5,99–101]. Sparsity is found to be useful prior information for model coefficients. However, there are two important challenges in sparse coding based video anomaly detection. Firstly, difficulties in the detection of video anomalies involving multiple objects and secondly, the capability of the sparsity-based linear model to represent the class separation effectively. These problems are addressed with the help of adaptive sparse representations, i.e., a combination of joint sparsity model for anomaly detection involving multiple objects and non-linearity for class separation [102].

Further, various robust sparse coding structures such as Hierarchical sparse coding method for classification [103], coarse-to-fine blocks (splitting of frames in multi-scale structures for computing the sparse representations) [88,89], and dynamic sparse coding with sliding window mechanism [93] have been applied to detect the anomalies. Similarly, dictionary learning-based anomaly detection methods such as adaptive dictionary learning [104], and online adaptive dictionary learning with weighted sparse coding [105] are used to generate better dictionary representations of reduced space complexity. These methods rarely considered the encoding of both spatial as well as temporal connections of various blocks and different frames, respectively [65]. Hence, a two-stage stacked sparse coding method capable of encoding both spatial and temporal connections using the Foreground Interest Point (FIP) descriptor of the blocks for video anomaly detection has been proposed [65]. Here, the anomaly detection accuracy is improved with the help of an intra-frame classification strategy. Recently, data-driven approaches such as deep learning techniques have been attempted to overcome the limitations of handcrafted features and utilize the representative capacity of the Deep Neural Networks (DNNs) for extracting the high-level feature representations from the sub-regions of the video. Few important DNN-based models such as Convolutional 3D (C3D) network [106], Genetic Adaptive Incident Detection (GAID) [107], Appearance and Motion DeepNet (AMDN) [108], PCANet [90,109], Adaptive Intra-frame Classification Network (AICN) [110], Temporally-coherent Sparse Coding (TSC) inspired Deep Neural Network [63], self-supervised representation learning approach [111], and so on are used for video anomaly detection.

6.4.2. Reconstruction models

Reconstruction models for the video anomaly detection learn only the normal behavior or activities using the training video data set comprising only normal data samples. Various methods based on Principal Component Analysis (PCA) and Auto-Encoders (AEs) have been used in reconstruction models to represent the appearance pattern and motion patterns with the help of the linear as well as nonlinear transformations. During the inference or testing phase, only the normal events conform to the learned model and are properly reconstructed (low reconstruction error). However, anomalous or abnormal activities do not conform to the learned model and cause poor reconstruction (high reconstruction error). Subsequently, frames having high reconstruction error (usually above the set threshold) are treated as the anomalous frames corresponding to the anomalous activities. Generally, anomalous activities or abnormalities evident themselves as deviations from the normal visual patterns, and the reconstruction error is a function of frame visual statistics. Hence, reconstruction errors are widely used in detecting the anomalous activities [3,45].

PCA [112] reduces the dimensionality problem by linearly projecting the high-dimensional data sets into a lower-dimensional space to segregate the signal from the noise. In the case of the video anomaly detection, PCA is used to model the spatial correlations among the pixel values corresponding to a frame [10]. A data point that resides far away from the projection plane is treated as an anomalous sample. However, PCA mainly suffers from two problems, such as linear projection and highly sensitive to data perturbation. Firstly, it is not always possible to separate the data points distinctly due to the linear projection mechanism. Secondly, most of the time, masking of the anomalies results as PCA is highly sensitive to data perturbation, i.e., only one extreme data point is capable of the orientation of the projection completely [113]. However, these problems are addressed by the methods using Auto-Encoders (AEs).

An AE is a neural network whose output follows the input and is trained by the back-propagation mechanism with objective to minimize the reconstruction error. However, the mapping from input to output is based on a non-linearity, introduced by the nonlinear activation functions used in the neurons. Hence, AEs helps in performing the dimensionality reduction as well as better feature representations as compared to PCA [10,114]. Any AE consists of feed-foward multilayer neural network having more than one number of hidden layers is known as Deep Auto-Encoder (DeepAE). The DeepAEs are capable of providing better feature representations due to the hierarchical feature extraction mechanism. Mathematically, an AE can be represented as

$$\widehat{X} = D(E(X)) \tag{1}$$

where, X is the input video data, $\widehat{X}$ is the reconstructed video data, E is the Encoder which encodes input data to hidden layer, and D is the Decoder which map from hidden layer to output layer. The objective is to train the Encoder and Decoder (ED) pair to minimize the reconstruction error $\|X-\widehat{X}\|$ by following the optimization problem

$$\min_{D,E} \|X-\widehat{X}\|. \tag{2}$$

Convolutional Auto-Encoders (CAEs) use the convolution operator instead of summation operation to preserve the spatial relationship among the pixels [56,115]. Generally, CAEs are learned to extract the features that can be used to reconstruct the same input, preferably the images. A deep CAE was trained to preserve the spatiotemporal properties during encoding dynamics in an end-to-end training framework [45], i.e., temporal regularity of the video sequence was learned and used to detect the video anomaly. Moreover, CAE accompanied by high-level spatial and temporal features has found to be suitable for detection of contextual video anomalies [116]. Further, a spatio-temporal autoencoder based on Convolutional Long-Short-Term-Memory (ConvLSTM) architecture [117,118] is proposed to learn the spatiotemporal patterns of the input video sequences for detecting video anomalies [56]. Here, convolution and LSTM are used in modeling due to the effectiveness in preserving the spatial as well as temporal properties, respectively. Though the model is semi-supervised, it is capable of detecting video anomalies while providing robustness to noise. However, these methods provide more false alarms for more complex scenes. A hybrid spatiotemporal autoencoder comprising ConvAE and LSTM Encoder-Decoder was proposed to extract better contextual spatiotemporal features for video anomaly detection [119]. Here, extrapolate ability to generate better frames of the decoder is improved using a shortcut connection for increasing the information propagation in the decoding step. A Robust Deep AE (RDAE) is capable of extracting high-quality nonlinear features and eliminating the outliers as well as noise without access to any clean training data [114]. In other words, RDAE can be considered as the combination of AE and Robust PCA (here, the input data X is segregated to the effective reconstruction by the DAE and noises as well as outliers present in the original data). Here, the anomaly is detected by allowing a sparse set of exceptions to the enforcement of the auto-encoder function on-the-fly. Robust PCA (RPCA) is unable to perform inductive anomaly detection, i.e., making predictions on test data samples. So, a robust deep AE with inductive (a generalization of the model to unseen data points of test data) ability is used to detect video anomaly in live setting [113]. A sparse AE based system capable of detecting and localizing the video anomalies dynamically is proposed to reduce the execution memory requirement and false positive rate [120]. Structured AE based on subspace clustering can learn nonlinear transformations to smoothly mapping the input to the output without disturbing the local and global subspace structures [121]. Though various types of the AEs are widely used for data representation to detect video anomalies; they suffer from the black-box nature, i.e., unable to explain the reason to justify why a particular data sample is anomalous [113].

In the case of reconstruction modeling, learning the distribution of the training data by using DNN is one of the crucial tasks. Various tractable distribution estimators such as Neural Autoregressive Distribution Estimator (NADE) [122], deep Auto-regressive Network [123], Contractive Auto-Encoder (CAE) [124] and Masked Autoencoder for Distribution Estimation (MADE) [125] have been proposed to efficiently model the distribution of high-dimensional vectors such as video in a hierarchical manner. However, there is a need for compact and robust distribution estimators for video anomaly detection in real-time applications.

Recently, inspired by the deep learning approaches, a spatiotemporal feature known as Convolutional 3D (C3D) is successfully used as a video descriptor [92]. Generally, high-level percepts (intermediate visual representations) provides better discriminative information with low-spatial resolution. Low-level percepts facilitate modeling of finer motion patterns by preserving higher spatial resolution at the cost of high dimensional video representations. Hence, Ballas et al. [126] proposed a convolution-based Gated Recurrent Unit (GRU) network that can reduce the dimensions of the feature vectors and the number of model parameters by enforcing sparse connectivity and sharing the model parameters across the input spatial locations. Here, the memory requirement of the model is reduced by introducing the sparsity and locality concepts into the RNN units. Further, an energy-based anomaly detection technique based on the reconstruction model using the Restricted Boltzmann Machines (RBM) is proposed by Vu et al. [127]. However, it could use a limited potential of hierarchical feature extraction as the RBMs are shallow generative networks of two layers only.

6.4.3. Slow feature analysis

Slow Feature Analysis (SFA) [128] is based on the slowness principle, which extracts slowly varying representations of the rapidly varying high dimensional input using unsupervised representation learning technique [10]. Initially, SFA based video feature representations are used in human activity recognition [129,130]. Later, Deep Incremental Slow Feature Analysis Network (D-IncSFA) [131] is used to detect video anomalies by combining the feature extraction and anomaly detection into a single step using the global video feature representations [34]. The SFA based video anomaly detection methods are found to be suitable for online applications due to their low computational complexity. However, there is a high chance of getting more false alarms due to a lack of generalization ability.

6.5. Discriminative models

Discriminative modeling methods try to learn the discriminant features existing among the classes. Further, discrimininative models are always effective when these are trained in supervised learning technique for the balanced datasets. These methods are not widely used for the video anomaly detection due to lack of balanced video anomaly datasets and clarity in the definition of the anomalous activities as well as anomalous entities. However, a discriminative framework based on the classical density estimation approach is proposed for the detection of video anomalies [132]. Here, there is no need of temporal ordering of the sequences and splitting of the training sequences. A discriminative model having a layer of statistical inference for fusing the information related to anomaly across the time, space, and scale in a global consistent manner is proposed [89]. However, discriminative models are not widely used due to lack of universally accepted definition for the anomaly.

6.6. Predictive models

Video can be viewed as a spatio-temporal signal where the particular ordering of the frames provides a certain pattern. In case of the predictive modeling, the objective is to model the conditional distribution $P(X_t/(X_{t-1},X_{t-2},X_{t-3},\ldots\ldots,X_{t-p}))$ for predicting the current frame from the past frames [10]. The predictive models or spatio-temporal models are widely used for video anomaly detection as these models utilize simultaneously both spatial (or appearance) and temporal (or motion) features. A comparative analysis of the existing deep-learning based predictive models reported for the video anomaly detection and localization are presented in Table 2. Here, comparison is carried out based on the various aspects such as approaches followed (App.), deep network architecture used (NW architecture), proposed technique, strategy used for the anomaly formulation (ano. formulation), datasets used, chosen performance evaluation criteria (Eval. criteria) or anomaly measures and performance metrics. However, most of the time, the predictive models are found to be more effective for the offline applications as compared to the online applications due to more computational complexity.

6.7. Deep generative models

In case of the generative modeling, the objective is to learn the joint probability $P(X,Y)$ and subsequently calculate the conditional posterior probability $P(X/y)$ [10,133]. Generally, the generative model models the actual distribution of each class whereas the discriminative model models the decision boundary between the classes [133]. Deep generative models can learn through the maximum likelihood principle irrespective of the representation of the likelihood [10]. Recently, the deep generative models have been widely used for video anomaly detection as these models are capable of addressing the data scarcity as well as data imbalance problems of the video anomaly detection. A comparative analysis of the existing deep learning-based predictive models proposed for the video anomaly detection and localization is presented in Table 3. However, further research and development are required for using more deep generative models for the video anomaly detection.

6.8. Deep one-class classification models

The development of multi-class classification based deep learning models for the detection of video anomaly is restricted due to the dearth of anomalous ground truth data and ambiguous nature of the anomalies. Hence, detection of the anomaly with no labels for the anomalous samples can be treated as a One-Class Classification (OCC) or unary classification problem [53]. When DNNs are used to implement the OCC problem for video anomaly detection, then it is known as the Deep OCC models. Further, the Deep OCC models combine the hierarchical feature representation capability of the DNN with the one-class classification objectives such as hyperplane or hypersphere [27]. Recently, a limited number of researches have been proposed for the detection of video anomalies using Deep OCC models. The state of the art for the detection of the video anomalies using the OCC-based DNN models is given in Table 4. The Deep OCC models are capable of jointly training the DNN while optimizing the data enclosing hypersphere or hyperplane. However, these models require more training hours, specifically for the high dimensional input data such as video [27].

6.9. Deep hybrid models

The video anomaly detection suffers from the incompleteness of methodology, i.e., not a single method can do the complete job correctly. Hence, researches capable of taking advantage of multiple methods have been proposed. The hybrid models are comprising multiple models such a way that the overall performance of the end-to-end

Table 2
Predictive models for video anomaly detection.

Ref.	App.	NW architecture	Proposed technique	Ano. formulation	Datasets	Eval. criteria			Performance metrics				
						FL	PL	DPL	ROC	AUC	EER	DR	FPS
[134], 2013	PO	Spatiotemporal compositions	Bag of Video words	Similarity map construction	UCSD Ped1 & Ped2 [32], Subway [135], Anomalous Behavior/ York [136]	✓	✓	–	✓	–	–	–	–
[91], 2016	AO	CNN	Spatial–temporal CNN	Binary classification	UCSD Ped1 & Ped2 [32], UMN [137,138], Subway [135], U-turn [139]	✓	✓	–	✓	✓	–	✓	✓
[140], 2016	PO	CNN	Combination of CNN and background models	Nomality model and outlier detection	UCSD Ped1 & Ped2 [32], Subway [135], Anomalous Behavior/ York [136]	–	✓	–	✓	✓	–	–	✓
[58], 2017	PO	CNN (pretrained VGG Net)	Unmasking	Training accuracy resulted from classifying the consecutive frames	UCSD Ped1 & Ped2 [32], UMN [137,138]	✓	✓	–	✓	✓	–	–	✓
[141], 2017	AO	Fast R-CNN	Generic models and environment-specific model	Anomaly score of object proposal	UCSD Ped2 [32], Avenue [88]	✓	✓	–	✓	✓	✓	–	–
[142], 2017	AO	CNN	Dense Optical flow and One Class SVM with RBF kernel	Deviation of pixels from hyper plane	UCSD Ped1 & Ped1 [32]	✓	✓	–	✓	✓	–	✓	–
[11], 2018	AO	Conv-LSTM	Prediction of future frame from small number of input frames	Regularity score based on reconstruction loss	UCSD Ped1 & Ped2 [32], Subway [135], Avenue [88]	✓	–	–	–	–	✓	✓	–
[143], 2018	AO	Conv-LSTM Autoencoder	Spaitial and temporal feature extractions by CNN and LSTM, respectively	Regularity score based on reconstruction loss	UCSD Ped1 & Ped2 [32]	✓	–	–	–	✓	✓	✓	–
[144], 2018	PO	YOLO, CNN (VGG Net), LSTM	Combination of enity separation, posture classification and abnormal behavior detection	Classification	UT interaction data [145]	–	✓	–	–	–	–	✓	–
[57], 2018	PO	C3D	Bag of videos used for deep anomaly ranking model	Multiple Instant Learning (MIL) ranking loss	UCSD Ped1 & Ped2 [32], UMN [137,138], Subway [135], Aveune [88], BOSS [146], UCF crime [57]	✓	✓	–	✓	✓	–	–	✓
[147], 2019	AO	Two Stream CNN	Feature fusion from the two separate CNNs corresponding to spatial and temporal features	Binary classification	UCF crime [57]	–	–	–	–	–	–	✓	–
[148], 2019	AO	U-Net, ConvLSTM	Spatiotemporal frame prediction	Regularity score	UCSD Ped1 & Ped2 [32], CUHK Avenue [88]	✓	✓	–	–	✓	✓	✓	✓
[149], 2020	AO	Residual STAE using 3D CONVLSTM	Residual blocks are used to mitigate the vanishing gradient problem	Reconstruction error	UCSD Ped1 & Ped2 [32], CUHK Avenue [88], LV [150]	✓	–	–	✓	✓	–	–	–

Table 3
Generative models for video anomaly detection.

Ref.	App.	NW architecture	Proposed technique	Ano. formulation	Datasets	Eval. criteria			Performance metrics				
						FL	PL	DPL	ROC	AUC	EER	DR	FPS
[151], 2015	PO	Variational AE	Learning the normal distribution	Reconstruction probability	MNIST [152], KDD [153]	–	✓	–	✓	✓	–	–	–
[154], 2017	AO	GAN	GAN is trained to generate only normal distribution	Detection of the outliers w.r.t the learned distribution	UCSD Ped1 & Ped2 [32], UMN [137,138]	✓	✓	–	✓	✓	✓	–	–
[155], 2017	AO	Adversarial AE	Combination of Convolutional AE and GAN	Regularity score	UCSD Ped1 & Ped2 [32]	✓	✓	–	✓	–	–	–	–
[54], 2017	PO	Gaussian Classifier, 3D CNN	Cascading of shallow networks for detection of the normal patches and deep network to detect the complex normal patches	Mahalanobis distance between the test patch and Gaussian models	UCSD Ped1 & Ped2 [32]	✓	✓	–	✓	✓	✓	–	✓
[1], 2018	AO	Weighted Convolutional AE - LSTM network	Unmasking	Segmented moving foregrounds is used as a priori information	UCSD Ped1 & Ped2 [32], CUHK Aveune [88]	✓	–	–	–	✓	–	–	–
[43], 2019	AO	Stacked Variational AE	$S^2 - VAE$ comprising of $S_F - VAE$ and $S_c - VAE$ is used to detect the video anomaly	Training accuracy resulted from classifying the consecutive frames	UCSD Ped1 & Ped2 [32], UMN [137,138], Avenue [88], PETS [156]	✓	✓	–	✓	✓	✓	✓	–
[157], 2020	PO	Modified form of UNet	GAN is used to achieve better performance of prediction	PSNR	UCSD Ped2 [32], CHUCK Avenue [88], ShanghiTech [158]	✓	–	–	✓	✓	✓	–	–
[159], 2020	AO	U-Net	Bidirectional prediction of the same target frame by the forward and the backward prediction subnetworks	PSNR	UCSD Ped1 & Ped2 [32], CUHK Avenue [88]	✓	✓	–	–	✓	–	–	–

Table 4
One-Class DNN models for video anomaly detection.

Ref.	App.	NW architecture	Proposed technique	Ano. formulation	Datasets	Eval. criteria			Performance metrics				
						FL	PL	DPL	ROC	AUC	EER	DR	FPS
[53], 2019	AO	CNN	DeepOC capable of jointly optimizing the representation learning and one-class classification using CNNs	Reconstruction cost	UCSD Ped1 & Ped2 [32], Avenue [88], LV [150]	✓	✓	–	✓	✓	✓	✓	✓

Table 5
Hybrid models for video anomaly detection.

Ref.	App.	NW architecture	Proposed technique	Ano. formulation	Datasets	Eval. criteria			Performance metrics				
						FL	PL	DPL	ROC	AUC	EER	DR	FPS
[46], 2019	AO	SC2Net, LSTM	AnomalyNet is implemented using joint neural processing of feature learning, sparse representation, and dictionary learning	Reconstruction cost and sparsity loss	UCSD Pedestrian [32], CUHK Avenue [88], UMN [137,138]	✓	✓	–	✓	✓	✓	✓	–
[160], 2020	PO	STAE	Incremental Saptio-Temporal Learner (ISTL) is used with active learning and fuzzy aggression	Reconstruction cost	UCSD Ped1 & Ped2 [32], Avenue [88]	✓	✓	–	–	✓	✓	–	✓

pipeline for video anomaly detection increases significantly. Here, representative features extracted by the DNN-based models are provided as the inputs to the traditional ML algorithms such as Support Vector Machine (SVM), one-class Radial Basis Function (RBF), and so on [27]. The state of the art for the detection of the video anomalies using the hybrid models is given in Table 5. The deep hybrid models are more scalable and computationally efficient. Further, these models are effective in dealing with the “curse of dimensionality” specifically for high dimensional input data such as video. However, the hybrid model approach is sub-optimal due to its zero influence on the representational learning within the hidden layers [27].

7. Comparative analysis of deep learning-based video anomaly detection methods

Each of the DNN-based video anomaly detection techniques as discussed in the previous section has its unique advantages as well as disadvantages. The selection of the best suitable approaches and methods for a particular video anomaly detection problem is the most crucial aspect. The researches in the field of video anomaly detection are in a fast-changing and growing phase. Hence, it is not always feasible to give a clear cut answer to the question, “which video anomaly detection technique is the best suitable for a particular application?”. However, a summary of the relative strength and weakness of various video anomaly detection methods for setting a rough guideline is presented as follows.

- Learning: An appropriate learning technique should be selected based on the quality and quantity of the available datasets. However, semi-supervised video anomaly detection techniques are most suitable for video anomaly detection as these use the full potential of the normal data while addressing the issue of weakly-labeling and data imbalance. Recently, active learning-based video anomaly detection techniques are found to be most effective for real-time applications. However, one has to carry the initial expenses of the human expert.
- Approach: Either AO or PO approaches are selected based on the desired accuracy and targeted application. The AO approaches are most suitable for offline applications, whereas the PO approaches are most suitable for online applications.
- Method: The selection of the appropriate deep-learning method suitable for the particular video anomaly detection problem depends on available datasets, targeted applications, and expected performances. The deep-learning-based features must be compact, robust, efficient, and descriptive in nature. The predictive models are useful in video anomaly detection as they consider both spatial and temporal patterns. However, these models suffer from high time complexity for achieving high detection accuracy. The deep OCC models utilize the hierarchical feature learning ability of the DNN simultaneously with the optimization of the one-class objective. These models are found to be computationally effective [27]. Moreover, the hybrid models use the best of deep learning as well as traditional machine learning methods. In other words, in the case of hybrid models, the deep learning techniques are used for feature extraction and best performing ML algorithms are used for the classification. However, these models are sub-optimal as it is unable to influence the representation learning in the hidden layers [27].

8. Performance evaluation methodologies

The effectiveness of the proposed algorithms for the detection and localization of the video anomalies are evaluated based on the datasets, computational infrastructure, performance metric of both qualitative and quantitative analysis.

8.1. Datasets

The publicly available datasets for the video anomaly detection are very less as this research area is comparatively new research field. Further, the fewness of the bench-marked datasets available for the video anomaly detection and location is due to rareness as well as infinite varieties of the anomalous activities in real-life scenarios [3,65]. The most commonly used datasets for the detection and localization of video anomalies are UCSD Pedestrian [32], UMN [137,138], CHUCK Avenue [88], Subway [135], PETS 2009 [156], Anomalous Behavior/ York [136], QMUL Junction [161–164], MIT Traffic [8,165], Violent flows [166], Weizmann [167], CAVIAR [168], BOSS [146], i-Lids challenge for detection of bags and vehicles [169], U-turn [139], Web [170], crowd [171–173], BEHAVE [174], etc. The majority of the publicly available datasets contain simulated abnormal behaviors, a limited number of realistic anomalous behaviors, videos that are recorded using predefined scripts, training and test samples from different camera setup, and videos of mostly on ideal environment [150]. Deep-learning based video anomaly detection methods require large datasets covering realistic anomalous behaviors. Recently, few important datasets such as the LV dataset [150], ShanghiTech dataset [158], and UCF-Crime datasets [57] are put forth for developing as well as testing deep-learning-based video anomaly detection methods. The datasets are different based on various important aspects such as dataset duration, size, resolution, surveillance environment, scenarios covered, challenges offered by the dataset, targeted applications, anomalous events involved, and availability of the ground truth (GT). The selection of a good combination of test data is one of the crucial components in deep learning-based developments [175]. A short review of video datasets used for video anomaly detection is presented in [176]. Furthers, there is a demand for good bench-marks of large size to evaluate the algorithms used for the detection as well as localization of video anomalies.

8.2. Computational infrastructure

The development and performance of the deep-learning-based algorithms meant for the detection as well as localization of the video anomalies significantly depend upon the computational infrastructure. Mostly, the data-driven techniques like deep learning methods are able to provide state-of-the-art results due to the availability of large datasets and better computational facility. Therefore, the performances of the developed algorithms should be compared under the same or equivalent set of experimental setups. The performances during training (learning) and testing (inference) of the developed models are dependent on both algorithms (software) as well as hardware performances. Hence, there is a requirement for the co-design of the hardware and software to get better performances during training and testing of the models [177,178]. Generally, computational systems having high-end system configurations such as higher RAM, GPU, processor, etc. are used for the development in a deep-learning environment for achieving better performances. So, based on the specifications of the training and inference devices, various performance parameters such as training time, fps, accuracy, etc. vary significantly.

Further, the efficiency of the system as an end-to-end pipeline depends upon the mode of computation. Broadly, the computing can be executed using the facilities of either centralized computing infrastructure or distributed computing infrastructure. The centralized computing can be achieved using either local central processing or cloud computing. The distributed computing can be realized by using any one of these configurations such as fog computing, edge computing, the combination of edge, fog, and cloud computing. In the case of local central processing, all the processing is carried out by the central processor locally. In the case of cloud computing, all the computation is carried out by the central processor located in remote place [179]. Cloud computing should be capable of providing on-demand self-service, broad network access, resource pooling, elasticity, and measured service [180]. In the case of cloud computing, all the uploaded data are processed at the cloud server and results are made available at the devices. Hence, cloud computing suffers from the inherent time delay in processing and requirement of the high bandwidth. In case of edge computing, all the computations is performed at the node of data acquisition, nearby user [181,182]. This helps in the reduction of bandwidth requirement, transmission cost, network traffic, and latency. However, the synchronization of the edge devices is difficult. Fog computing is a conceptual extension of cloud computing where computation, communication, and storage facilities are closer to the edge devices [183,184]. It tries to combine the benefits of both the traditional servers and cloud servers. However, fog computing requires more complex management and high investment. In the case of video anomaly detection, the lightweight and efficient anomaly detection models may be executed at the edge. Further, more computationally expensive models of activity detection, object detection, and so forth may be processed at the fog servers. Finally, all the processing and communication can be synchronized by the cloud servers. Hence, there is a high demand for the deep learning-based models having smartness (intelligent), low-latency, good privacy, better mobility, and low energy consumption [177].

8.3. Evaluation criteria

Generally, the performances of the detection as well as localization of the video anomalies are measured using the three evaluation criteria such as Frame-Level (FL), Pixel-Level (PL), and Dual-Pixel-Level (DPL) [46,54,65]. Recently, it has been reported that a better spatial localization can be ensured by Intersection Over Union (IOU) between a ground truth anomalous region and a detected anomalous region [185]. Further, three evaluation criteria, i.e., Object-Level (OL), Track-Level (TL) and Region-Level (RL) are proposed based on this IOU concept. All these evaluation criteria (scales of measurement) consider the matching between the evaluated results and corresponding Ground-Truth (GT).

8.3.1. Frame-level

In Frame-Level (FL) evaluation criterion, the anomaly is measured at the frame level. A frame is considered as an anomaly even if an anomaly is detected for at least one pixel of the particular frame [54]. In other words, the proposed algorithm predicts whether a frame contains anomalous events or objects. Subsequently, these predictions are compared with the frame-level GT labels to determine the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) [46]. In practice, it is advisable to find the starting and ending frames of the video anomaly quickly during real-world video surveillance [65]. This is because anomaly detection is a coarse-level understanding and the clipped anomalous video segment can be sent for further in-depth video analysis such as activity detection, object detection, etc. Hence, the FL evaluation criterion is a primary, relevant, and meaningful scale of measurement for the detection of the video anomalies.

8.3.2. Pixel-level

In Pixel-Level (PL) evaluation criterion, the anomaly is measured at the pixel level. A frame is considered as an anomaly when the 40% or more percentage of the evaluated anomalous pixels are matched with the pixel-level GT labels [54]. In other words, the pixels corresponding to the video anomalies (i.e., anomalous event or anomalous entity) are predicted by an algorithm first and subsequently compared with the pixel-level GT annotations for deciding the TP and FP frames [46]. The PL evaluation criterion provides correct detection of video anomalies as compared to that of FL evaluation criteria [65]. Hence, the PL evaluation criterion is not only helpful in anomaly detection but also useful in anomaly localization.

8.3.3. Dual-pixel-level

In Dual-Pixel-Level (DPL) criterion, a frame is considered as anomaly only if the following two conditions are satisfied [54]. Firstly, the frame must satisfy the PL evaluation criteria, i.e., the frame is an anomalous one as per the PL evaluation criterion. Secondly, there are at least β % (say, 10%) of pixels detected as the anomaly are common to the anomaly GT. Here, if the video anomaly detection algorithm not only identifies the anomaly region but also irrelevant regions as the anomaly, then the particular frame is not considered as anomalous one. Hence, this evaluation criterion reduces false alarms significantly. Further, the DPL evaluation criterion is found to be more effective as compared to the PL criterion during the performance evaluation of anomaly localization [54,186]. However, the effectiveness of the DPL evaluation criterion for counting the TPs and FPs is significantly reduces in the frames having multiple anomalies [185].

8.3.4. Object-level

In Object-Level (OL) evaluation criterion, the anomaly measurement is more concerned with the detection of the anomalous object. Sometimes, an anomalous frame as per the PL evaluation criterion may contain a large number of false-positive pixels. More than 40% of true anomalous pixels must be correctly detected when all the pixels of an anomalous frame are detected as anomalies. Hence, the OL evaluation criterion is used to find the anomalous frames having the Detected Abnormality Area (DAA) is closed to the True Abnormality Area (TAA) for a given threshold θ_{th} [9,65]. Mathematically, the OL evaluation criterion is represented in Eq. (3) [65].

$$\frac{DAA \cap TAA}{DAA \cup TAA} \geq \theta_{th} \tag{3}$$

8.3.5. Track-level

In Track-Level (TL) evaluation criterion or Track-based detection criterion, the performance of the anomaly detector is evaluated using the similar criteria used in object tracking. This criterion is used to measure the Track-Based Detection Rate (TBDR) versus the number of false-positive regions per frame [185]. The TBDR can be defined as the ratio of the number of anomalous tracks detected to that of anomalous tracks. This evaluation criterion is found to be effective in the frames having multiple anomalies, preferably in the sparsely as well as moderately crowded environment.

8.3.6. Region-level

In Region-Level (RL) evaluation criterion or Region-based detection criterion, the performance of the anomaly detector is evaluated using the similar criteria used in object detection. This criterion is used to measure the Region-Based Detection Rate (RBDR) over all the frames in the test versus the number of false-positive regions per frame [185]. The RBDR can be defined as the ratio of the number of anomalous regions detected to that of the anomalous regions. This evaluation criterion is found to be effective in the frames having multiple anomalies, even if in the densely crowded environment.

However, there is a further requirement of research for investigating the effectiveness of the OL, TL, and RL evaluation criteria.

8.4. Performance metrics for quantitative analysis

Most of the performance metrics used in the computer vision for objection, tracking, activity recognition, and so forth are not suitable for measuring the performance of the video anomaly detection algorithms [187]. For example, the ground truth found to be suitable for the object detection algorithms are not adequate for the video anomaly detection algorithms. This is because the ground truth suitable for object detection may have contained the non-object related pixels, which may introduce biases in the performance measurement of the anomaly detection. Hence, various types of other or modified performance metrics are used for the evaluation of the algorithms used for the detection of the video anomalies. The strength, weakness, effectiveness, and applicability of the various algorithms proposed for the detection of video anomalies should be investigated and measured. Further, the performance of these algorithms should be quantified using the most suitable evaluation metric for comparative analysis. There are various performance metrics such as receiver operating characteristic as well as precision curves, the area under the curve, etc. used for the quantitative analysis.

8.4.1. Error matrix

Though there is an underlying difference between the video anomaly detection algorithms and other domains of the computer vision, there is a common fundamental goal to classify the data points either as normal or anomalies. So, video anomaly detection can be treated as a binary classification problem. Further, each decision of the binary classifier can be represented by anyone of the basic elements of the performance analysis such as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These four performance parameters can be arranged in a tabular form known as a confusion matrix or error matrix for supervised learning and matching matrix for unsupervised learning. Subsequently, other advanced performance metrics such as True Positive Rate (TPR) or Recall or sensitivity or hit rate, True Negative Rate (TNR) or selectivity or specificity, False Positive Rate (FPR) or fall-out, False Negative Rate (FNR) or miss rate, Precision or Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Discovery Rate (FDR), False Omission Rate (FOR), Accuracy and F_1 Score can be derived from this error matrix [10,187,188]. Typical representations for illustrating the evaluation process in video anomaly detection and confusion matrix are depicted in Fig. 3 and Fig. 4, respectively. A decision is called as TP only when the truly anomalous data samples are predicted as anomaly. A decision is called as TN only when the truly normal data samples are predicted as normal. A decision is called as FP only when the truly normal data samples are predicted as anomaly. A decision is called as FN only when the truly anomalous data samples are predicted as normal. Here, the data samples may be pixel or frame based on the selected evaluation criteria.

In case of anomaly detection, it is vary much important to minimize FN (Type-II error) as compared to FP (Type-I error) as there is high potential for losses of assets and lives in case of Type-I error.

8.4.2. Receiver operating characteristic curve

Receiver Operating Characteristic (ROC) is a plot between TPR or sensitivity (in Y-axis) and FPR or probability of false alarm (in X-axis) for measuring the performance of detection at various FPRs (or thresholds) [10]. The Area Under Receiver Operating Characteristic curve (AU-ROC) is used to quantify the accuracy of the anomaly detector for a given test set [189]. The value of the AU-ROC should be as large as possible within the range of zero to one.

8.4.3. Precision-recall curve

Precision-Recall (PR) is a plot between precision (in Y-axis) and recall (in X-axis) [10]. The Area Under Precision-Recall curve (AU-PR) is more useful for the anomaly detection problem as compared to the AU-ROC. This is because of the data imbalance problem of the anomaly detection where TNs are very large as compared to the TPs. Further, the PR curve is more focused on the predictions around the positive or anomaly. The value of the AU-PR should be as large as possible within the range of zero to one.

8.4.4. Equal error rate

Equal Error Rate (EER) is defined as the percentage of misclassified frames when the TPR is equal to the FNR. It is effective for the detection of video anomalies [53].

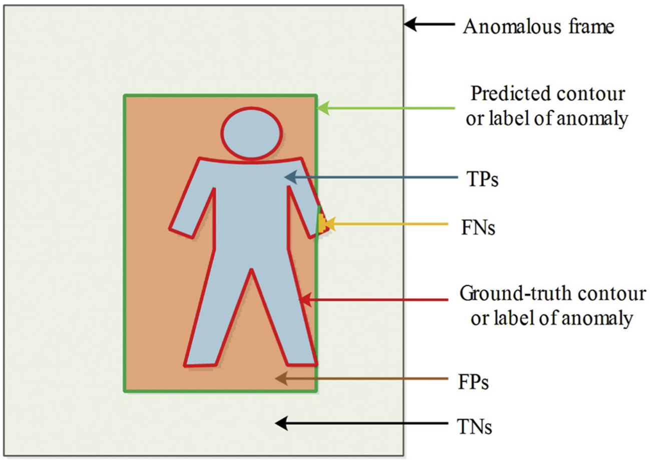
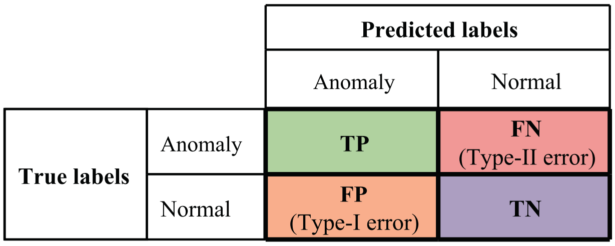

Anomalous frame
Predicted contour or label of anomaly
TPs
FNs
Ground-truth contour or label of anomaly
FPs
TNs

Fig. 3. Representation for illustrating the evaluation process used in video anomaly detection.



		Predicted labels	
		Anomaly	Normal
True labels	Anomaly	TP	FN (Type-II error)
	Normal	FP (Type-I error)	TN

Fig. 4. Confusion matrix.

8.4.5. Detection rate

Detection rate (DR) is defined as the ratio of the number of the anomalies detected to the total number of anomalies present in the data in percentage. DR or precision rate measured at the EER is effective for the localization of video anomalies [53].

$$Detection\ rate = \frac{TP}{TP + FP} \tag{4}$$

8.4.6. Reconstruction error

The reconstruction error $e_{reconst}(t)$ of a frame at time instant t corresponding to a video is calculated using Eq. (5) [45],

$$e_{reconst}(t) = \sum\nolimits_{(x,y)} e(x,y,t), \tag{5}$$

where $e(x,y,t)$ pixel level reconstruction error at a location (x,y) in the particular frame at time instant t for intensity level I. The $e(x,y,t)$ is calculated from the from the trained model f_W using Eq. (6) [45].

$$e(x,y,t) = \|I(x,y,t) - f_W(I(x,y,t))\|_2 \tag{6}$$

Higher values of the reconstruction error signifies the higher the probability of detection of the video anomaly.

8.4.7. Anomaly score

Anomaly score $S_{ano}(t)$ in the range of 0 to 1 is used to explain the degree of anomaly and is calculated using Eq. (7) [56].

$$S_{ano}(t) = \frac{e_{reconst}(t) - e_{reconst_{\min}}(t)}{e_{reconst_{\max}}(t)} \tag{7}$$

Higher values of the anomaly score signifies the higher level of the anomaly.

8.4.8. Regularity score

Regularity score $S_{reg}(t)$ can be considered as the opposite of the anoamly score and is calculated using Eq. (8) [56].

$$S_{reg}(t) = 1 - S_{ano}(t) \tag{8}$$

The higher values of the regularity score signifies the lower level of the anomaly.

8.4.9. Peak signal-to-noise ratio

The error between the generated frame and the ground truth frame can be used to measure the anomaly in terms of Peak Signal-to-Noise Ratio (PSNR). The higher values of the PSNR score signifies the lower level of the anomaly. Mathematically, the PSNR score in DB is defined as Eq. (9) [159],

$$PSNR = 10 \log_{10}\left(\frac{I_{\max}^2}{MSE}\right), \tag{9}$$

where $I_{\max}$ represents the maximum fluctuation of intensity permitted in the input frame, MSE is the Mean-Square-Error between the generated frame and original frame.

8.4.10. Computational complexities

Finally, it is always advisable to calculate the computational complexities such as time complexity and space complexity of the video anomaly detector. The time complexity is required to quantify the run time performance in terms of frame processing time and frame processing speed (number of frames processed in one second) of the video anomaly detector. The number of operations executed by the model can be measured in FLOPS, which is the number of floating-point operations [53]. Broadly, the time complexity is dependent on the input dimensions, size of the DNN (number of neurons and layers), experimental setup, code organization, and so forth. Ideally, it should be as small as possible for a desired level of accuracy.

Similarly, the space complexity of the final model should be calculated to decide the required hardware specification for deployment. The space complexity also depends upon the size of DNN, output feature maps, network parameters, etc. It is also advisable that the DNN model used for the detection of video anomalies should possess less space complexity for the desired level of accuracy. However, similar to processing-time and accuracy trade-off, there is a trade-off between the time and space complexities for the desired level of performance of the anomaly detector.

The selection of proper evaluation criteria and corresponding performance metrics is a very important task as the selection can favor (or oppose) the anomaly detector by introducing positive (or negative) bias for a particular application [187].

8.5. Performance metrics for qualitative analysis

There is a need of qualitative analysis to strengthen the decision taken based on the quantitative analysis. Here, it is advisable to uncover or visualize the decision-making process so that a proper justification can be made available for the corresponding decision. Followings are the important performance metrics used for the qualitative analysis of the video anomaly detection.

8.5.1. Visualization of frame regularity

A plot of the regularity score versus the corresponding frame numbers with proper annotations and anomalous inset frames is used to visualize the frame regularity as well as video anomaly detection. An example of the qualitative analysis for the video anomaly detection using visualization of the frame regularity can be found in [56].

8.5.2. T-distributed stochastic neighbor embedding

Proper visualization of the datasets is helpful in investigating the anomalous nature of the data samples. Since video is a high-dimensional signal, simply box-plots are not beneficial. T-distributed Stochastic Neighbor Embedding (t-SNE) is a data visualization technique suitable for high dimensional data (e.g., video) and is based on a non-linear dimensionality reduction mechanism [190]. Anomalies that are present in the datasets can be visualized in higher feature space by using the t-SNE as represented in [53].

8.5.3. Uncovering the decision-making process using explainable deep learning

The DNN-based models have been highly successful in the detection and localization of the video anomalies. However, one of the major bottlenecks in using these anomaly detectors in practical application is the opaqueness of the models or lack of interpretability. In other words, the DNNs work as a black-box, and most of the time, it lacks the proper justification for the decision being made [191,192]. Hence, explainable deep learning is used to uncover the opaqueness of the DNN-based models and provides a visual explanation to the user for “whether the DNN is doing the right things for the right reasons” [193]. Various explainable deep learning-based techniques such as Gradient-weighted Class Activation Mapping (GRAD-CAM) [194] and GRAD-CAM++ [195] can be used to generate the visual explanations in the form of the activation maps or heat-maps to highlight the anomalous region in the frame [194]. Recently, GRAD-CAM is used in addition to the spatio-temporal auto-encoder-based model to highlight the potential anomalous regions in the frame [196]. The explainable deep learning-based video anomaly detection is in its early stage and hence more attention is required to develop the video anomaly detectors capable of generating the visual explanations. The availability of an insight into the decision-making process during the practical applications will help in earning more trust from the end-users.

The selection and use of the proper performance metrics for the qualitative analysis significantly help the researchers to understand the strength and weaknesses of the models. Further, the qualitative analysis helps in understanding and perceiving the quantitative performances of the models.

9. Research challenges in deep learning approaches for the video anomaly detection

Based on the comprehensive literature survey carried out on the video anomaly detection, few important research challenges may be outlined as follows.

9.1. Need of better datasets

The publicly available datasets for video anomaly detection are significantly less as this research area is comparatively a new research field. Further, the data unbalance between positive samples (anomalous events) and negative samples (normal events) prevents the use of supervised learning-based models. The dearth of anomalous ground truth data (or insufficient labeled data) and ambiguous (or equivocal) nature of the anomalies hinder the development of end-to-end trainable deep learning models [127]. The high variance within positive samples (anomalous events may contain a great variety of different cases, though generally, only limited training data is available) makes the modeling further complicated [2]. Hence, there is a demand for good bench-marks to evaluate the algorithms used for the detection as well as localization of video anomalies [3,65].

9.2. Reduction in computational complexity

Generally, the costly step (in other words, computationally expensive as well as longer time-consumption) of feature representation during video anomaly detection, acts as a major bottleneck for the deployment of the video anomaly detector in practical applications [127]. Subsequently, most of the existing algorithms used for the detection of the video anomalies have high costs of time and space complexities. Hence, these methods are not suitable for the applications of the real world [65]. Therefore, there is a requirement of more compact as well as deeper DNN (networks having more numbers of layers and less number of neurons) for better feature representation. Further, there is a need for simple, lightweight, and efficient algorithms for the detection of video anomalies. However, the complexity of the algorithms used for the detection of video anomalies increases further with the high dimensional structure combined with non-local variation across the frames.

9.3. Solving incompleteness of the methodology

The existing methods for the detection of the video anomalies suffer from the incompleteness of the methodology, i.e., a single method is not capable of detecting all kinds of anomalies [65]. Hence, there is a demand for the efficient video anomaly detector that can address this incompleteness problem by using hybrid models.

9.4. Finding best evaluation methodologies

There is a requirement of more robust and efficient evaluation criteria as well as the performance metrics which can incorporate the context-aware spatio-temporal information.

9.5. Need for co-design of hardware and software

Deep learning-based video anomaly detection methods provide high detection accuracy at the expense of computational as well as space complexity. Hence, there is a need for hardware and algorithms (software) co-design [177,178] for attaining online performance without compromising on the detection accuracy.

9.6. Trade-off between accuracy and processing time

There is difficulty in finding the optimal trade-off between detection accuracy and processing times in video anomaly detection methods, by employing few, but highly descriptive features in order to attain online performance with competitive accuracy [42]. Further, the high accuracy of the detection as well as localization of the video anomalies using deep learning-based methods are achieved at the cost of high computational complexity and long processing time. So, there is always a trade-off between the required detection accuracy with the computational complexity as well as processing time.

9.7. Need to address the environmental challenges

The efficiency of the video anomaly detection algorithms is significantly affected by the change in scales because the scales of the objects in the surveillance camera change as per the viewpoint as well as the distance measured between the concerned object and the surveillance camera [65]. Though this problem is addressed by the use of grid pattern-based methods, there is still a requirement for further improvement. Further, there is a high demand for efficient techniques to tackle the challenges of the complex environment such as variation in backgrounds as well as illuminations, occlusion problems, noisy input, and working status of the surveillance camera [3].

10. Conclusions

The video anomaly detection is an essential component of the IVSS and is used to detect sufficiently interesting video outliers [197]. In this article, a comprehensive review of deep learning-based methods for the detection of video anomalies has been presented. The existing deep learning methods used for the detection of video anomalies have been categorized into different groups based on the modeling techniques. Subsequently, a comparative analysis of the existing deep learning-based video anomaly detection methods is provided for better selection of a particular method that works best for a particular application. Further, an in-depth analysis of the performance evaluation methodologies in terms of datasets, computational infrastructure, evaluation criteria, performance metrics for the quantitative and qualitative analysis is presented. Finally, the trending and open research challenges available in deep learning approaches for the video anomaly detection is briefly outlined. The video anomaly detection can be applied in many potential video surveillance application domains such as detection of crime activities, traffic violations, abnormal crowd behavior, abandoned objects, weapons at sensitive areas, and industrial production monitoring.

Declaration of Competing Interest

None.

Acknowledgement

This work is supported by the **IMP**ACTING **R**ESEARCH **IN**NOVATION AND **T**ECHNOLOGY (IMPRINT) INDIA (Grant No.-7794/2016), an initiative of Ministry of Human Resource Development, Government of India and Ministry of Housing and Urban Affairs, Government of India.

References

[1] B. Yang, J. Cao, R. Ni, L. Zou, Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention, Adv. Multimedia (2018) 1–8.

[2] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, X.-S. Hua, Spatio-temporal autoencoder for video anomaly detection, Proceedings of the 25th ACM International Conference on Multimedia 2017, pp. 1933–1941.

[3] O.P. Popoola, K. Wang, Video-based abnormal human behavior recognition-a review, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (6) (2012) 865–878.

[4] J. Varadarajan, J.-M. Odobez, Topic models for scene analysis and abnormality detection, IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE 2009, pp. 1338–1345.

[5] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, Semi-supervised adapted hmms for unusual event detection, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE 2005, pp. 611–618.

[6] H. Zhong, J. Shi, M. Visontai, Detecting unusual activity in video, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2004 , (pp. II–II).

[7] T. Xiang, S. Gong, Incremental and adaptive abnormal behaviour detection, Comput. Vis. Image Underst. 111 (1) (2008) 59–73.

[8] X. Wang, X. Ma, W.E.L. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2008) 539–555.

[9] K. Pawar, V. Attar, Deep learning approaches for video-based anomalous activity detection, World Wide Web 22 (2) (2019) 571–601.

[10] B.R. Kiran, D.M. Thomas, R. Parakkal, An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos, J. Imaging 4 (2) (2018) 36.

[11] J.R. Medel, A. Savakis, Anomaly detection in video using predictive convolutional long short- term memory networks, arXiv preprint arXiv:1612.00390 (2016).

[12] T. Ko, A survey on behaviour analysis in video surveillance applications, Video Surveillance, InTech 2011, pp. 279–294.

[13] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.

[14] J. Candamo, M. Shreve, D.B. Goldgof, D.B. Sapper, R. Kasturi, Understanding transit scenes: a survey on human behavior-recognition algorithms, IEEE Trans. Intell. Transp. Syst. 11 (1) (2009) 206–224.

[15] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vis. Image Underst. 115 (2) (2011) 224–241.

[16] N. Buch, S.A. Velastin, J. Orwell, A review of computer vision techniques for the analysis of urban traffic, IEEE Trans. Intell. Transp. Syst. 12 (3) (2011) 920–939.

[17] A.A. Sodemann, M.P. Ross, B.J. Borghetti, A review of anomaly detection in automated surveillance, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (6) (2012) 1257–1272.

[18] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[19] S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, Vis. Comput. 29 (10) (2013) 983–1009.

[20] X. Wang, Intelligent multi-camera video surveillance: A review, Pattern Recogn. Lett. 34 (1) (2013) 3–19.

[21] D. Gowsikhaa, S. Abirami, R. Baskaran, Automated human behavior analysis from surveillance videos: a survey, Artif. Intell. Rev. 42 (4) (2014) 747–765.

[22] M.A. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.

[23] Y.S. Chong, Y.H. Tay, Modeling video-based anomaly detection using deep architectures: Challenges and possibilities, 10th Asian Control Conference (ASCC), IEEE 2015, pp. 1–8.

[24] R.K. Tripathi, A.S. Jalal, S.C. Agrawal, Suspicious human activity recognition: a review, Artif. Intell. Rev. 50 (2) (2018) 283–339.

[25] S.A. Ahmed, D.P. Dogra, S. Kar, P.P. Roy, Trajectory-based surveillance analysis: a survey, IEEE Trans. Circuit Syst. Video Technol. 29 (7) (2018) 1985–1997.

[26] A.B. Mabrouk, E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: a review, Expert Syst. Appl. 91 (2018) 480–491.

[27] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: a survey, arXiv preprint arXiv:1901.03407 (2019).

[28] S.K. Kumaran, D.P. Dogra, P.P. Roy, Anomaly detection in road traffic using visual surveillance: a survey, arXiv preprint arXiv:1901.08292 (2019).

[29] Y.S. Chong, Y.H. Tay, Modeling representation of videos for anomaly detection using deep learning: A review, arXiv preprint arXiv:1505.00523 (2015).

[30] H. Jacobs, To count a crowd, Columbia J. Rev. 6 (1) (1967) 37.

[31] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L.J. Hadjileontiadis, M.G. Strintzis, Swarm intelligence for detecting interesting events in crowded environments, IEEE Trans. Image Process. 24 (7) (2015) 2153–2166.

[32] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE 2010, pp. 1975–1981.

[33] V. Saligrama, Z. Chen, Video anomaly detection based on local statistical aggregates, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 2112–2119.

[34] X. Hu, S. Hu, Y. Huang, H. Zhang, H. Wu, Video anomaly detection using deep incremental slow feature analysis network, IET Comput. Vis. 10 (4) (2016) 258–267.

[35] M.J. Leach, E.P. Sparks, N.M. Robertson, Contextual anomaly detection in crowded surveillance scenes, Pattern Recogn. Lett. 44 (2014) 71–79.

[36] A. Munawar, P. Vinayavekhin, G. De Magistris, Spatio-temporal anomaly detection for industrial robots through prediction in unsupervised feature space, Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE 2017, pp. 1017–1025.

[37] Y. Zhu, N.M. Nayak, A.K. Roy-Chowdhury, Context-aware activity recognition and anomaly detection in video, IEEE J. Select. Top. Signal Process. 7 (1) (2012) 91–101.

[38] H. Li, A. Achim, D. Bull, Unsupervised video anomaly detection using feature clustering, IET Signal Process. 6 (5) (2012) 521–533.

[39] J. Varadarajan, R. Subramanian, N. Ahuja, P. Moulin, J.-M. Odobez, Active online anomaly detection using dirichlet process mixture model and gaussian process classification, Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE 2017, pp. 615–623.

[40] T. Pimentel, M. Monteiro, A. Veloso, N. Ziviani, Deep active learning for anomaly detection, Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) 2020, pp. 1–8.

[41] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, X. He, Generative adversarial active learning for unsupervised outlier detection, IEEE Trans. Knowl. Data Eng. 32 (8) (2020) 1517–1528.

[42] R. Leyva, V. Sanchez, C.-T. Li, Video anomaly detection with compact feature sets for online performance, IEEE Trans. Image Process. 26 (7) (2017) 3463–3478.

[43] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, C. Choi, Generative neural networks for anomaly detection in crowded scenes, IEEE Trans. Inform. Forens. Secur. 14 (5) (2018) 1390–1399.

[44] D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, Comput. Vis. Image Underst. 156 (2017) 117–127.

[45] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, L.S. Davis, Learning temporal regularity in video sequences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 733–742.

[46] J.T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, R.S.M. Goh, Anomalynet: an anomaly detection network for video surveillance, IEEE Trans. Inform. Forens. Secur. 14 (10) (2019) 2537–2550.

[47] K.-W. Cheng, Y.-T. Chen, W.-H. Fang, Video anomaly detection and localization using hierarchical feature representation and gaussian process regression, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 2909–2917.

[48] K. Cheng, Y. Chen, W. Fang, Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation, IEEE Trans. Image Process. 24 (12) (2015) 5288–5301.

[49] R. Leyva, V. Sanchez, T.-L. Chang, Fast binary-based video descriptors for action recognition, Proceedings of the IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE 2016, pp. 1–8.

[50] R. Leyva, V. Sanchez, C.-T. Li, A fast binary pair-based video descriptor for action recognition, Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE 2016, pp. 4185–4189.

[51] M. George, B.R. Jose, J. Mathew, Abnormal activity detection using shear transformed spatio-temporal regions at the surveillance network edge, Multimed. Tools Appl. (2020) 1–22.

[52] D.M. Torres, H.L. Correa, E.C. Bravo, Online learning of contexts for detecting suspicious behaviors in surveillance videos, Image Vis. Comput. 89 (2019) 197–210.

[53] P. Wu, J. Liu, F. Shen, A deep one-class neural network for anomalous event detection in complex scenes, IEEE Trans. Neural Network Learn. Syst. 31 (7) (2020) 2609–2622.

[54] M. Sabokrou, M. Fayyaz, M. Fathy, R. Klette, Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes, IEEE Trans. Image Process. 26 (4) (2017) 1992–2004.

[55] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, R. Klette, Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes, Comput. Vis. Image Underst. 172 (2018) 88–97.

[56] Y.S. Chong, Y.H. Tay, Abnormal event detection in videos using spatiotemporal autoencoder, International Symposium on Neural Networks, Springer 2017, pp. 189–196.

[57] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 6479–6488.

[58] R. Tudor Ionescu, S. Smeureanu, B. Alexe, M. Popescu, Unmasking the abnormal events in video, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 2895–2903.

[59] Y. Cong, J. Yuan, Y. Tang, Video anomaly search in crowded scenes via spatiotemporal motion context, IEEE Trans. Inform. Forens. Secur. 8 (10, 2013) 1590–1599.

[60] M. Bertini, A. Del Bimbo, L. Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization, Comput. Vis. Image Underst. 116 (3) (2012) 320–329.

[61] V. Reddy, C. Sanderson, B.C. Lovell, Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture, Computer Vision and Pattern Recognition (CVPR) 2011 WORKSHOPS, IEEE 2011, pp. 55–61.
[62] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, V. Murino, Analyzing tracklets for the detection of abnormal crowd behavior, Proceedings of the IEEE Winter Conference on Applications of Computer Vision, IEEE 2015, pp. 148–155.
[63] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, S. Gao, Video anomaly detection with sparse coding inspired deep neural networks, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–15.
[64] B.T. Morris, M.M. Trivedi, A survey of vision-based trajectory learning and analysis for surveillance, IEEE Trans. Circuit Syst. Video Technol. 18 (8) (2008) 1114–1127.
[65] K. Xu, X. Jiang, T. Sun, Anomaly detection based on stacked sparse coding with intraframe classification strategy, IEEE Trans. Multimedia 20 (5) (2018) 1062–1074.
[66] C. Piciarelli, C. Micheloni, G.L. Foresti, Trajectory-based anomalous event detection, IEEE Trans. Circuit Syst. Video Technol. 18 (11, 2008) 1544–1554.
[67] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L.O. Alvares, F. Brémond, Toward abnormal trajectory and event detection in video surveillance, IEEE Trans. Circuit Syst. Video Technol. 27 (3) (2016) 683–695.
[68] X. Song, X. Shao, Q. Zhang, R. Shibasaki, H. Zhao, J. Cui, H. Zha, A fully online and unsupervised system for large and high-density area surveillance: tracking, semantic scene learning and abnormality detection, ACM Trans. Intell. Syst.Technol. (TIST) 4 (2) (2013) 1–21.
[69] L. Brun, A. Saggese, M. Vento, Dynamic scene understanding for behavior analysis based on string kernels, IEEE Trans. Circuit Syst. Video Technol. 24 (10, 2014) 1669–1681.
[70] J. Snoek, J. Hoey, L. Stewart, R.S. Zemel, A. Mihailidis, Automated detection of unusual events on stairs, Image Vis. Comput. 27 (1–2) (2009) 153–166.
[71] A. Revathi, D. Kumar, An efficient system for anomaly detection using deep learning classifier, SIViP 11 (2) (2017) 291–299.
[72] D. Tran, J. Yuan, D. Forsyth, Video event detection: from subvolume localization to spatiotemporal path search, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2013) 404–416.
[73] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, Proceedings of the IEEE Confeference on Computer Vision and Pattern Recognition 2016, pp. 779–788.
[74] J. Redmon, A. Farhadi, Yolo 9000: Better, faster, stronger, Proceedings of the IEEE Confeference on Computer Vision and Pattern Recognition 2017, pp. 7263–7271.
[75] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, Proceedings of the IEEE International Conference on Image Processing (ICIP) 2016, pp. 3464–3468.
[76] R. Nayak, M.M. Behera, U.C. Pati, S.K. Das, Video-based real-time intrusion detection system using deep-learning for smart city applications, Proceedings of the IEEE International Conference on Advanced Networks and Telecommunications Systems (IEEE ANTS), IEEE 2019, pp. 1–6.
[77] J. Redmon, A. Farhadi, Yolov3: An Incremental Improvement, arXiv preprint arXiv: 1804.02767 2020.
[78] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, Proceedings of the IEEE International Conference on Image Processing (ICIP) 2017, pp. 3645–3649.
[79] R. Nayak, M.M. Behera, V. Girish, U.C. Pati, S.K. Das, Deep learning based loitering detection system using multi-camera video surveillance network, Proceedings of the IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS), IEEE 2019, pp. 215–220.
[80] M. Elhamod, M.D. Levine, Automated real-time detection of potentially suspicious behavior in public transport areas, IEEE Trans. Intell. Transp. Syst. 14 (2) (2012) 688–699.
[81] Y. Zhang, L. Qin, R. Ji, H. Yao, Q. Huang, Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection, IEEE Trans. Circuit Syst. Video Technol. 25 (7) (2014) 1231–1245.
[82] Y. Yuan, Y. Feng, X. Lu, Statistical hypothesis detector for abnormal event detection in crowded scenes, IEEE Trans. Cybernet. 47 (11, 2016) 3597–3608.
[83] G. Xiong, J. Cheng, X. Wu, Y.-L. Chen, Y. Ou, Y. Xu, An energy model approach to people counting for abnormal crowd behavior detection, Neurocomputing 83 (2012) 121–135.
[84] B. Krausz, C. Bauckhage, Loveparade 2010: automatic video analysis of a crowd disaster, Comput. Vis. Image Underst. 116 (3) (2012) 307–319.
[85] S. Yi, X. Wang, C. Lu, J. Jia, L0 regularized stationary time estimation for crowd group analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 2211–2218.
[86] C.C. Loy, T. Xiang, S. Gong, Salient motion detection in crowded scenes, Proceedings of the 5th IEEE International Symposium on Communications, Control and Signal Processing, IEEE 2012, pp. 1–4.
[87] D.-G. Lee, H.-I. Suk, S.-K. Park, S.-W. Lee, Motion influence map for unusual human activity detection and localization in crowded scenes, IEEE Trans. Circuit Syst. Video Technol. 25 (10, 2015) 1612–1623.
[88] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 fps in matlab, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 2720–2727.
[89] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2013) 18–32.
[90] Y. Feng, Y. Yuan, X. Lu, Learning deep event models for crowd anomaly detection, Neurocomputing 219 (2017) 548–556.
[91] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, Z. Zhang, Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes, Signal Process. Image Commun. 47 (2016) 358–368.
[92] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 4489–4497.
[93] B. Zhao, L. Fei-Fei, E.P. Xing, Online detection of unusual events in videos via dynamic sparse coding, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 2011, pp. 3313–3320.
[94] W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked rnn framework, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 341–349.
[95] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE 2005, pp. 886–893.
[96] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, Proceedings of the European Conference on Computer Vision, Springer 2006, pp. 428–441.
[97] L. Kratz, K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 1446–1453.
[98] X. Cui, Y. Tian, L. Weng, Y. Yang, Anomaly detection in hyperspectral imagery based on low-rank and sparse decomposition, Fifth International Conference on Graphic and Image Processing (ICGIP 2013), vol. 9069, International Society for Optics and Photonics 2014, p. 90690R.
[99] S. Wu, B.E. Moore, M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE 2010, pp. 2054–2060.
[100] F. Jiang, J. Yuan, S.A. Tsaftaris, A.K. Katsaggelos, Anomalous video event detection using spatiotemporal context, Comput. Vis. Image Underst. 115 (3) (2011) 323–333.
[101] J. Kim, K. Grauman, Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 2921–2928.
[102] X. Mo, V. Monga, R. Bala, Z. Fan, Adaptive sparse representations for video anomaly detection, IEEE Trans. Circuit Syst. Video Technol. 24 (4) (2014) 631–645.
[103] K. Yu, Y. Lin, J. Lafferty, Learning image representations from the pixel level via hierarchical sparse coding, Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE 2011, pp. 1713–1720.
[104] C. Lu, J. Shi, J. Jia, Scale adaptive dictionary learning, IEEE Trans. Image Process. 23 (2) (2013) 837–847.
[105] S. Han, R. Fu, S. Wang, X. Wu, Online adaptive dictionary learning and weighted sparse coding for abnormality detection, Proceedings of the IEEE International Conference on Image Processing, IEEE 2013, pp. 151–155.
[106] W. Chu, H. Xue, C. Yao, D. Cai, Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos, IEEE Trans. Multimedia 21 (1) (2018) 246–255.
[107] P. Roy, B. Abdulhai, Gaid: genetic adaptive incident detection for freeways, Transp. Res. Rec. 1856 (1) (2003) 96–105.
[108] D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe, Learning Deep Representations of Appearance and Motion for Anomalous Event Detection, arXiv preprint arXiv: 1510.01553 2020.
[109] T. Bao, S. Karmoshi, C. Ding, M. Zhu, Abnormal event detection and localization in crowded scenes based on pcanet, Multimed. Tools Appl. 76 (22, 2017) 23213–23224.
[110] K. Xu, T. Sun, X. Jiang, Video anomaly detection and localization based on an adaptive intra-frame classification network, IEEE Trans. Multimedia 22 (2) (2020) 394–406.
[111] R. Ali, M.U.K. Khan, C.M. Kyung, Self-Supervised Representation Learning for Visual anomaly Detection, arXiv preprint arXiv:2006.09654 2020.
[112] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417.
[113] R. Chalapathy, A.K. Menon, S. Chawla, Robust, deep and inductive anomaly detection, Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer 2017, pp. 36–51.
[114] C. Zhou, R.C. Paffenroth, Anomaly detection with robust deep autoencoders, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2017, pp. 665–674.
[115] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, Proceedings of the International Conference on Artificial Neural Networks, Springer 2011, pp. 52–59.
[116] M. Ribeiro, A.E. Lazzaretti, H.S. Lopes, A study of deep convolutional auto-encoders for anomaly detection in videos, Pattern Recogn. Lett. 105 (2018) 13–22.
[117] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, Advances in Neural Information Processing Systems 2015, pp. 802–810.
[118] V. Patraucean, A. Handa, R. Cipolla, Spatio-temporal video autoencoder with differentiable memory, Workshop track of International Conference On Learning Representations (ICLR) 2015, pp. 01–13.
[119] L. Wang, F. Zhou, Z. Li, W. Zuo, H. Tan, Abnormal event detection in videos using hybrid spatio-temporal autoencoder, 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE 2018, pp. 2276–2280.
[120] M.G. Narasimhan, S. Kamath, Dynamic video anomaly detection and localization using sparse denoising autoencoders, Multimed. Tools Appl. 77 (11, 2018) 13173–13195.
[121] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J.T. Zhou, S. Yang, Structured autoencoders for subspace clustering, IEEE Trans. Image Process. 27 (10, 2018) 5076–5086.

[122] H. Larochelle, I. Murray, The neural autoregressive distribution estimator, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics 2011, pp. 29–37.

[123] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, D. Wierstra, Deep Autoregressive Networks, arXiv preprint arXiv:1310.8499 2020.

[124] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress 2011, pp. 833–840.

[125] M. Germain, K. Gregor, I. Murray, H. Larochelle, Made: Masked autoencoder for distribution estimation, Proceedings of the International Conference on Machine Learning 2015, pp. 881–889.

[126] N. Ballas, L. Yao, C. Pal, A. Courville, Delving Deeper into Convolutional Networks for Learning Video Representations, arXiv preprint arXiv:1511.06432 2020.

[127] H. Vu, D. Phung, T.D. Nguyen, A. Trevors, S. Venkatesh, Energy-Based Models for Video Anomaly Detection, arXiv preprint arXiv:1708.05211 2020.

[128] L. Wiskott, T.J. Sejnowski, Slow feature analysis: unsupervised learning of invariances, Neural Comput. 14 (4) (2002) 715–770.

[129] Z. Zhang, D. Tao, Slow feature analysis for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 436–450.

[130] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, S. Yan, Dl-sfa: Deeply-learned slow feature analysis for action recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 2625–2632.

[131] V.R. Kompella, M. Luciw, J. Schmidhuber, Incremental slow feature analysis: adaptive low-complexity slow feature updating from high-dimensional input streams, Neural Comput. 24 (11, 2012) 2994–3024.

[132] A. Del Giorno, J.A. Bagnell, M. Hebert, A discriminative framework for anomaly detection in large videos, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 334–349.

[133] P.M. Joshi, Generative vs Discriminative Models, https://medium.com/@mlengineer/generative-and-discriminative-models-af5637a66a3, medium Blog, Accessed July 07, 2020 2018.

[134] M.J. Roshtkhari, M.D. Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions, Comput. Vis. Image Underst. 117 (10) (2013) 1436–1452.

[135] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 555–560.

[136] A. Zaharescu, R. Wildes, Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing, Proceedings of the European Conference on Computer Vision, Springer 2010, pp. 563–576.

[137] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, N. Papanikolopoulos, Real time, online detection of abandoned objects in public areas, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE 2006, pp. 3775–3780.

[138] E. Ribnick, S. Atev, O. Masoud, N. Papanikolopoulos, R. Voyles, Real-time detection of camera tampering, Proceedings of the IEEE International Conference on Video and Signal Based Surveillance, IEEE 2006, pp. 10–15, https://doi.org/10.1109/AVSS.2006.94.

[139] Y. Benezeth, P.-M. Jodoin, V. Saligrama, C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 2458–2465.

[140] P. Christiansen, L. Nielsen, K. Steen, R. Jørgensen, H. Karstoft, Deepanomaly: combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field, Sensors 16 (11, 2016) 1904.

[141] R. Hinami, T. Mei, S. Satoh, Joint detection and recounting of abnormal events by learning deep generic knowledge, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 3619–3627.

[142] L. Sun, H. Ai, S. Lao, Localizing activity groups in videos, Comput. Vis. Image Underst. 144 (2016) 144–154.

[143] P. Singh, V. Pankajakshan, A deep learning based technique for anomaly detection in surveillance videos, Proceedings of the Twenty Fourth National Conference on Communications (NCC), IEEE 2018, pp. 1–6.

[144] K.-E. Ko, K.-B. Sim, Deep convolutional framework for abnormal behavior detection in a smart surveillance system, Eng. Appl. Artif. Intell. 67 (2018) 226–234.

[145] M.S. Ryoo, J. Aggarwal, Ut-interaction dataset, icpr contest on semantic description of human activities (sdha), Proceedings of the IEEE International Conference on Pattern Recognition Workshops, vol. 2, 2010, p. 4.

[146] Boss, , URL http://www.multitel.be/image/researchdevelopment/research-projects/boss.php.

[147] S. Majhi, R. Dash, P.K. Sa, Two-stream cnn architecture for anomalous event detection in real world scenarios, Proceedings of the International Conference on Computer Vision and Image Processing, Springer 2019, pp. 343–353.

[148] Y. Li, Y. Cai, J. Liu, S. Lang, X. Zhang, Spatio-temporal unity networking for video anomaly detection, IEEE Access 7 (2019) 172425–172432.

[149] K. Deepak, S. Chandrakala, C.K. Mohan, Residual spatiotemporal autoencoder for unsupervised video anomaly detection, Signal, Image and Video Processing 2020, pp. 1–8.

[150] R. Leyva, V. Sanchez, C.-T. Li, The lv dataset: A realistic surveillance video dataset for abnormal event detection, Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF), IEEE 2017, pp. 1–6.

[151] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Special Lecture on IE, 2, 2015, pp. 1–18.

[152] Y. LeCun, C. Cortes, C. Burges, Mnist Handwritten Digit Database, at&t labs, http://yann. lecun. com/exdb/mnist 2010.

[153] S. Hettich, S. Bay, The uci kdd Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed July 07, 2020 1999.

[154] M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, Training Adversarial Discriminators for Cross-Channel Abnormal Event Detection in Crowds, arXiv preprint arXiv:1706.07680 2020.

[155] A. Dimokranitou, Adversarial Autoencoders for Anomalous Event Detection in Images, Ph.D. thesis Purdue University, Indianapolis, Indiana, 2017.

[156] CVPR, Pets 2009 Benchmark Data, http://www.cvg.reading.ac.uk/PETS2009/a.html 2020 (Jun. 2009).

[157] F. Dong, Y. Zhang, X. Nie, Dual discriminator generative adversarial network for video anomaly detection, IEEE Access 8 (2020) 88170–88176.

[158] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 6536–6545.

[159] D. Chen, P. Wang, L. Yue, Y. Zhang, T. Jia, Anomaly detection in surveillance video based on bidirectional prediction, Image Vis. Comput. 103915 (2020).

[160] R. Nawaratne, D. Alahakoon, D. De Silva, X. Yu, Spatiotemporal anomaly detection using deep learning for real-time video surveillance, IEEE Trans. Ind. Inform. 16 (1) (2019) 393–402.

[161] C.C. Loy, T. Xiang, S. Gong, Modelling multi-object activity by gaussian processes, British Machine Vision Conference (BMVC), Citeseer 2009, pp. 1–11.

[162] C.C. Loy, T. Xiang, S. Gong, Stream-based active unusual event detection, Proceedings of the Asian Conference on Computer Vision, Springer 2010, pp. 161–175.

[163] C.C. Loy, T. Xiang, S. Gong, Detecting and discriminating behavioural anomalies, Pattern Recogn. 44 (1) (2011) 117–132.

[164] C.C. Loy, T.M. Hospedales, T. Xiang, S. Gong, Stream-based joint exploration-exploitation active learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 1560–1567.

[165] M. Wang, X. Wang, Automatic adaptation of a generic pedestrian detector to a specific traffic scene, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 2011, pp. 3401–3408.

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

[167] Weizmann dataset, URL http://www.wisdom.weizmann.ac.il/vision/Irregularities.html 2020.

[168] R.B. Fisher, The pets04 surveillance ground-truth data sets, Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance 2004, pp. 1–5.

[169] I. I. C. on Advanced Video, S. based Surveillance, i-lids bag and Vehicle Detection Challenge, URL http://www.eecs.qmul.ac.uk/andrea/avss2007_d.html 2020.

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

[171] S. Ali, M. Shah, Crowd Flow Segmentation & Stability Analysis, URL http://www.cs.ucf.edu/sali/Projects/CrowdSegmentation/index.html 2020.

[172] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2007, pp. 1–6.

[173] M. Rodriguez, J. Sivic, I. Laptev, J.-Y. Audibert, Data-driven crowd analysis in videos, Proceedings of the International Conference on Computer Vision, IEEE 2011, pp. 1235–1242.

[174] Behave, URL http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/ 2020.

[175] O. Zendel, M. Murschitz, M. Humenberger, W. Herzner, How good is my test data? Introducing safety analysis for computer vision, Int. J. Comput. Vis. 125 (1–3) (2017) 95–109.

[176] N. Patil, P.K. Biswas, A survey of video datasets for anomaly detection in automated surveillance, Proceedings of the Sixth International Symposium on Embedded Computing and System Design (ISED), IEEE 2016, pp. 43–48.

[177] S. Han, B. Dally, Efficient Methods and Hardware for Deep Learning, https://www.youtube.com/watch?v=eZdOkDtYMoo&list=PLC1qU-LWwrF64f4QKQT-Vg5Wr4qEE1Zxk&index=15 2020.

[178] S. Han, H. Mao, W.J. Dally, Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, arXiv preprint arXiv:1510.00149 2020.

[179] S.B. Shaw, A.K. Singh, A survey on cloud computing, Proceedings of the International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE) 2014, pp. 1–6.

[180] M.U. Bokhari, Q. Makki, Y.K. Tamandani, A survey on cloud computing, in: V.B. Aggarwal, V. Bhatnagar, D.K. Mishra (Eds.), Big Data Analytics, Springer Singapore, Singapore 2018, pp. 149–164.

[181] W. Yu, F. Liang, X. He, W.G. Hatcher, C. Lu, J. Lin, X. Yang, A survey on the edge computing for the internet of things, IEEE Access 6 (2018) 6900–6919.

[182] R. Roman, J. Lopez, M. Mambo, Mobile edge computing, fog, et al., A survey and analysis of security threats and challenges, Futur. Gener. Comput. Syst. 78 (2018) 680–698.

[183] R. Huang, Y. Sun, C. Huang, G. Zhao, Y. Ma, A survey on fog computing, in: G. Wang, J. Feng, M.Z.A. Bhuiyan, R. Lu (Eds.), Security, Privacy, and Anonymity in Computation, Communication, and Storage, Springer International Publishing, Cham 2019, pp. 160–169.

[184] M. Mukherjee, L. Shu, D. Wang, Survey of fog computing: fundamental, network applications, and research challenges, IEEE Commun. Surv. Tutorials 20 (3) (2018) 1826–1857.

[185] B. Ramachandra, M. Jones, Street scene: A new dataset and evaluation protocol for video anomaly detection, Proceedings of the IEEE Winter Conference on Applications of Computer Vision 2020, pp. 2569–2578.

[186] M. Sabokrou, M. Fathy, M. Hoseini, R. Klette, Real-time anomaly detection and localization in crowded scenes, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2015, pp. 56–62.

[187] S. Parameswaran, J. Harguess, C. Barngrover, S. Shafer, M. Reese, Evaluation schemes for video and image anomaly detection algorithms, Automatic Target Recognition XXVI, vol. 9844, International Society for Optics and Photonics, 2016 , 98440D.

[188] Wikipedia, Confusion matrix, URL https://en.wikipedia.org/wiki/Confusion_matrix 2020.

[189] R.A. Maxion, R.R. Roberts, Proper Use of ROC Curves in Intrusion/Anomaly Detection, University of Newcastle upon Tyne, Computing Science Tyne, UK, 2004.

[190] L.V.D. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[191] L. Hiley, A. Preece, Y. Hicks, Explainable deep learning for video recognition tasks: A framework & recommendations, arXiv preprint arXiv:1909.05667 2020.

[192] N. Xie, G. Ras, M. van Gerven, D. Doran, Explainable Deep Learning: A Field Guide for the Uninitiated, arXiv preprint arXiv:2004.14545 2020.

[193] K. Amarasinghe, K. Kenney, M. Manic, Toward explainable deep neural network based anomaly detection, Proceedings of the 11th International Conference on Human System Interaction (HSI), IEEE 2018, pp. 311–317.

[194] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 618–626.

[195] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE 2018, pp. 839–847.

[196] S. Bhakat, G. Ramakrishnan, Anomaly detection in surveillance videos, Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, CoDS-COMAD '19, Association for Computing Machinery, New York, NY, USA 2019, pp. 252–255.

[197] C.C. Aggarwal, Outlier analysis, Data Mining, Springer 2015, pp. 237–263.