World Wide Web (2019) 22:571–601
https://doi.org/10.1007/s11280-018-0582-1

Deep learning approaches for video-based anomalous activity detection

Karishma Pawar[1] · **Vahida Attar**[1]

Received: 16 August 2017 / Revised: 16 April 2018 / Accepted: 26 April 2018 /
Published online: 3 May 2018

Abstract The pervasive use of cameras at indoor and outdoor premises on account of recording the activities has resulted into deluge of long video data. Such surveillance videos are characterized by single or multiple entities (persons, objects) performing sequential/concurrent activities. It is often interesting to detect suspicious behavior of such entities in an automated manner without any intervention of human personnel, and to this end, anomalous activity detection from surveillance videos is an important research domain in Computer Vision. Detecting the anomalous activities from videos is very challenging due to equivocal nature of anomalies, context at which events took place, lack of ample size of anomalous ground truth training data and also other factors associated with variation in environment conditions, illumination conditions and working status of capturing cameras. Though automated visual surveillance is one of the highly sought-after research domains, use of deep learning techniques for anomalous activity detection is still in nascent stage. Deep learning models like convolution neural networks, auto-encoders, Long Short Term Memory network models have achieved remarkable performance on different domains like image classification, object detection, speech processing, and expediting towards achieving excellence in anomaly detection tasks. This paper aims at studying and analyzing deep learning techniques for video-based anomalous activity detection. As outcome of the study, the graphical taxonomy has been put forth based on kinds of anomalies,

This article belongs to the Topical Collection: *Special Issue on Deep vs. Shallow: Learning for Emerging Web-scale Data Computing and Applications*
Guest Editors: Jingkuan Song, Shuqiang Jiang, Elisa Ricci, and Zi Huang

✉ Karishma Pawar
kvppawar@gmail.com

Vahida Attar
vahida.comp@coep.ac.in

[1] Department of Computer Engineering & IT, College of Engineering Pune (COEP), Pune, India

level of anomaly detection, and anomaly measurement for anomalous activity detection. The focus has been given on various anomaly detection frameworks having deep learning techniques as their core methodology. Deep learning approaches from both the perspectives of accuracy oriented anomaly detection and real-time processing oriented anomaly detection are compared. This paper also sheds light upon research issues and challenges, application domains, benchmarked datasets and future directions in the domain of deep learning based anomaly detection.

Keywords anomalous activity detection · anomaly modeling · computer vision · deep learning · real time detection · video surveillance

1 Introduction

Imagine the smart scenario in which the robot is proactively detecting all suspicious activities performed by human, avoiding the crowd turbulences and violence before they get worse, acting as a surveillance agent at public and private places, avoiding the robberies/theft at sensitive areas by informing concerned authorities, and many more. Though we may not have reached that stage yet, current technology is expediting towards such amazing era with self-operated and autonomous robots working continuously without human intervention. This survey talks about progress made by newly emerged deep learning technology and other non-deep learning approaches in the area of video based anomalous activity detection.

Automated visual surveillance as an active area in computer vision has been one of the most sought-after research domains in academia and business firms due to its wide applicability for monitoring of public and private places, crowd management, elderly health care systems, defense systems, transportation systems. So, installing Closed Circuit Television (CCTV) cameras has been popular option for monitoring the ongoing activities and achieving global security. Due to less cost, ease of use and customized design of cameras, global surveillance camera market is anticipated to increase at compound annual growth rate of 16.6% from 2017 to 2025 [102].

This proliferation of cameras for effective monitoring has resulted into deluge of video data. In 2015, around 566 petabytes of data were produced by video surveillance cameras installed worldwide, and would generate 2500 petabytes of data daily by the end of 2019 [19]. As continuous monitoring of such videos is beyond the capacity of video operator personnel, there is a need of automated, online visual surveillance system to operate continuously and detect suspicious/anomalous behavior of objects and human from video in near-real time manner.

Due to its wide scope for providing global security, the past 2 decades witnessed great improvements in video based anomaly detection approaches. Many review papers have been put forth in the domain of human activity recognition, behavior understanding, and crowded scene analysis which are directly or indirectly relevant to video based anomaly detection [8, 40, 44, 46, 52, 72, 89, 90, 94, 104, 117, 124]. It can be observed from the existing literature that no review paper has assessed deep learning approaches for video based anomalous activity detection. The work by Chong and Tay discussed the use of deep architectures for anomaly detection [14]. But it covers a very short review of deep learning methods for anomaly detection. Therefore,

the aim of this survey paper is to thoroughly analyze the progress made by deep learning techniques in the field of video based anomaly detection.

The work is contributed as follows:

- The graphical taxonomy of video-based anomalous activity detection has been put forth.
- Thorough survey of state-of-the-art deep learning approaches for video based anomaly detection is done.
- The trade-offs in anomaly detection from video are mentioned from both the view-point of accuracy-oriented approaches and real-time processing oriented approaches using deep learning techniques.
- Newly introduced datasets in the past lustrum, need and issues of anomaly detection are explored.
- The current challenges, application domains and possible future directions in the domain of deep learning applicable to anomaly detection are thoroughly put forth.

The roadmap of the paper is depicted in Figure 1. Section 2 deals with taxonomy of anomalous activity detection from videos. Section 3 and 4 deal with traditional and deep learning approaches for anomaly detection respectively. Application domains and benchmarked datasets are briefly overviewed in section 5. Research challenges and future directions in anomaly detection are enunciated section 6. Section 7 concludes the paper.

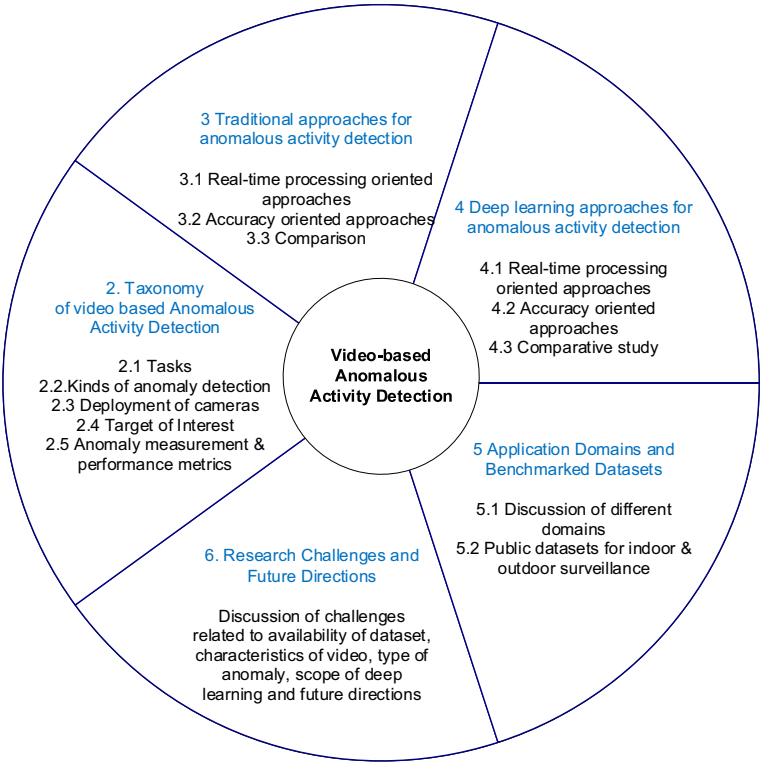

Figure 1 Roadmap of the paper (Figure to be read from left in clockwise manner)

2 Taxonomy of video based anomalous activity detection

"*Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior*" [10]. Anomalous activity is also known as irregular behavior, suspicious activity, surprising event [36], unusual activity and so on. Anomalous events are context and subject dependent, new, unknown, rare and therefore, challenging to detect from videos. Figure 2 depicts the taxonomy of video based anomaly detection. The taxonomy is depicted based on various factors to be considered for performing anomalous activity detection.

2.1 Tasks

Anomalous activity detection focuses on finding whether the given video frame exhibits an anomaly or not. It addresses the question, *"Does the given frame contain an anomaly or not?"* Anomalous activity localization performs the localization of anomalies by determining actual location of anomaly in the given video frame by bounding box. It addresses the question, "*Where is anomaly occurring in the given frame?"* Localizing the groups performing activities has been well handled by Lei et al. [92] using latent graph model. The tasks of detection and localization have been jointly performed in [12, 16, 66, 79, 81, 114, 115].

2.2 Kinds of anomaly detection

- Referring to the literature [12, 15, 16], anomalous events can be classified into 2 classes' viz. local anomaly and global anomaly. Local anomalous event differs from spatio-temporal neighboring events and deals with finding how the activity of an individual

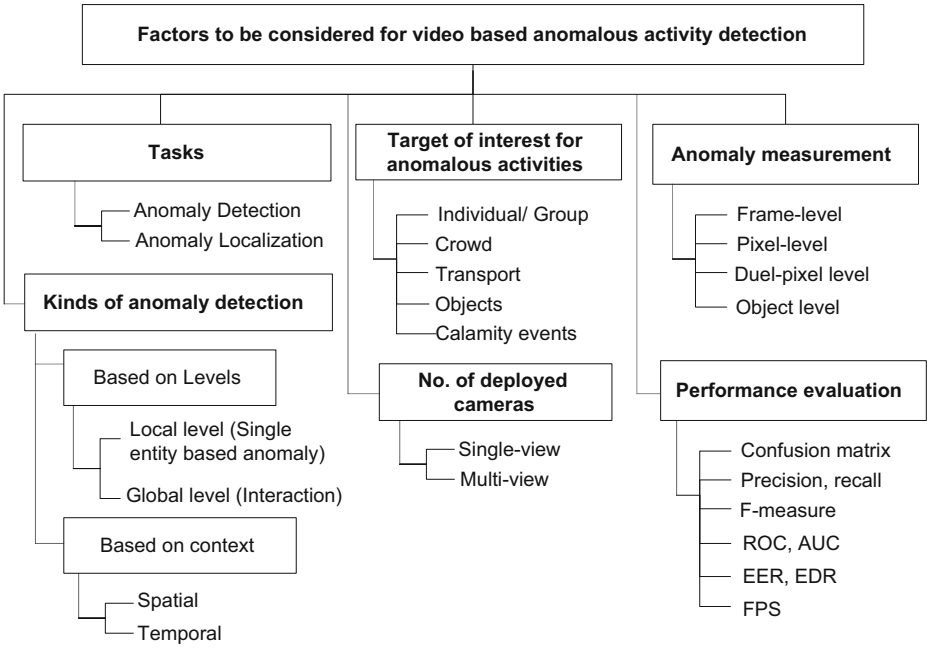

Figure 2 Taxonomy of video based anomaly detection

varies from its neighbors (For example, driving a vehicle in wrong direction). Local anomalous activity detection has been well investigated in [2, 45, 57, 84]. On the contrary, global anomalous events globally interact with each other in an unusual way, even if any local events are normal or anomalous individually i.e. multiple events though seem normal, interact with each other in a suspicious/unusual manner (For example, car accidents; crowd dispersion due to explosion). It also involves entities behaving in suspicious manner and their collective activity is harmful, for example, violence and robbery. Joint modeling of local and global anomalous events is done in [12, 78]. Both [78] and [12] used spatio-temporal video descriptors for joint modeling of local and global events. Cong et al. [15] used spatio-temporal features extracted using Histogram of Optical Flow for detecting anomaly at multiple locations and scales. Their approach is based on sparse coding technique.

- On the similar lines, Yu et al. [118] defined single point anomaly and interaction based anomaly. Point anomaly can be mapped to anomalous activity of individual entity, termed as single entity based anomaly. The interaction of group in unconventional manner maps to interaction based anomaly. The variants of interaction based anomalies can be person interacting with object i.e. human-object interaction (For example, person keeping the bag unattended at public place), human-human interaction (For example, fighting People) or object-object interaction (For example, vehicles colliding with each other). Complexity and time required for anomaly detection and localization increases from anomaly due to single entity to interaction based entities and finally crowd anomaly detection and localization. This is well depicted in Figure 3.
- The definition of anomaly varies as the context varies. For example, car running on highway is normal activity whereas running of the same car on pedestrian walkway is anomalous one. This is known as contextual anomaly in which activity in specific scenario is said to be anomalous according to some context, whereas the same activity is normal in

Figure 3 Kinds of Anomaly detection based on number of entities involved

other context. Therefore, activities related to each other by space and time forms the context [123], and it is necessary to model the appearance features obtained from spatial domain and motion features obtained from temporal domain in a joint manner. Contextual anomalies are divided into spatial and temporal anomalies [48, 64]. By and large, spatial anomalies are detected from single frame, whereas minimum two-frames (observations collected over consecutive time stamps) are required for temporal anomaly detection.

2.3 Cameras deployed for surveillance

One of the important factors for accurately detecting the anomaly is the number of cameras deployed for capturing the video and the view/angle at which the cameras are fixed. It's important to capture the videos from multiple views/cameras since there are chances that all activities (normal/anomalous) may not be captured by single camera and framework may miss detecting the anomalous activity if any. Sometimes, suspicious individual may also deliberately avoid camera and hide the activities. These issues can be addressed by capturing the multiple views of ongoing activities.

Multiple views can be captured from multiple cameras, sensors, or thermal cameras [20, 87]. Once all the videos are obtained, video summarization can be performed by removing redundant views and decision of anomaly detection is taken. Most research work has focused on detecting the anomaly from single view camera [51]. Though the use of multi-camera (multi-view) for anomaly detection is complex and challenging task compared to the anomaly detection using single view, multi-camera anomaly detection has potential to provide accurate detection capability since it helps to capture the spatio-temporal features (context) of the video efficiently.

2.4 Target of interest for anomalous activities

Anomaly detection is applied to both indoor and outdoor environment, and therefore the challenges associated with surveillance videos of such environments need to be handled carefully. Indoor surveillance videos are characterized by change in illumination levels of room, light perturbation, reflections in architectural components like windows or doors. This kind of surveillance mainly covers offices, shops, ATMs, home-based healthcare systems. On the other hand, outdoor surveillance videos involve change in illumination levels based on time of day, weather conditions based on rain, snow and fog. This surveillance covers both controlled environments and uncontrolled environments like sports arena, crowd scenario, pedestrian walkways, transportation systems and many more. In summary, target of anomaly detection can be individuals/groups, crowd scenarios, transport domain, naturally or artificially occurring calamities like flood detection, fire detection, etc.

2.5 Anomaly measurement and performance metrics

Different performance metrics are used for anomaly detection. This includes True positive rate, False-positive rate, precision, recall. Confusion matrix depicting performance measures used for anomaly detection is shown in Figure 4. For this, ground-truth is delineated as follows. Presence of anomalous activity is understood as "positive" whereas, its absence as

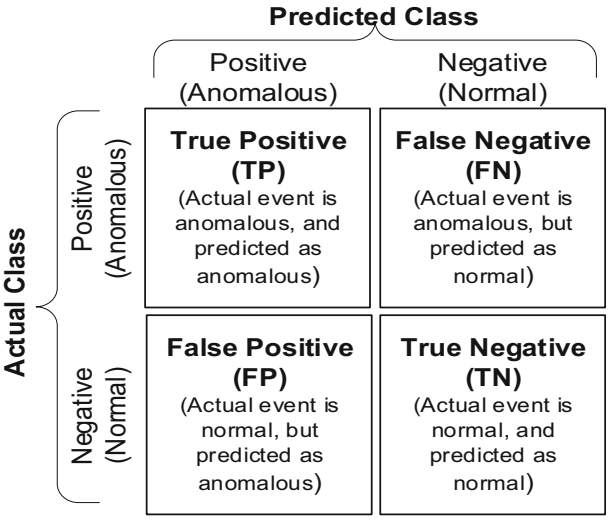

Figure 4 Confusion matrix for anomaly detection

"negative" in the confusion matrix. Receiver Operating Characteristic (ROC) curves are preferably used for visualizing and comparing the performance of classification methods used in anomaly detection. It is a two-dimensional graphical representation of true positive rate plotted on Y-axis and false positive rate plotted on X-axis. In case of anomaly detection, it is used for checking the performance trade-off between benefits of true positive (accurately predicted events) versus false positive (inaccurately predicted events). Three more performance metrics are evaluated based on ROC curves viz. Area under the ROC Curve (AUC), Equal Error rate (EER), and Equal Detection Rate (EDR). EER is determined by the ratio of uncategorized frames when FP rate = 1 – TP rate. The Detection Rate (DR) calculated at equal error rate is termed as EDR. In case of real time processing oriented anomaly detection approaches, frames per second (FPS), or time required for processing each frame in the video is also considered. For evaluating the performance of the anomaly detection models, various levels of detection are considered viz. frame-level, pixel level, duel pixel level and object level.

- Frame-level anomaly detection: The whole frame is said to be anomalous if at least one pixel for a test frame is predicted to be anomalous.
- Pixel-level anomaly localization: It measures the accuracy of spatially located anomalous region as predicted by system. As mentioned by Li, Mahadevan, & Vasconcelos [51], comparison of predicted anomalous pixels is done with the pixel-level ground truth; if there is 40% overlap of predicted anomalous pixels with that of ground-truth region, then given test frame is assigned true positive metric, on the other hand, frame is considered as a false positive.
- Duel Pixel level (DPL): If some region detected by algorithm possesses some anomaly and suppose this region (either obtained by frame-level or pixel-level) overlaps with ground truth anomalous region, then that region is termed as "lucky guess". Frame-

level and pixel-level measures do not take into account this false region ("lucky guess"). In order to detect such "lucky guess" region, duel pixel measure is introduced [79]. A frame is said to contain anomalous activity, if following conditions are satisfied. (1) Frame contains anomaly at frame-level. (2) There should be at least $\beta\%$ overlap of predicted anomalous pixels with that of ground-truth region (3) In addition to anomalous region, if unimportant regions are also considered as anomaly, then duel-pixel measure identify it as true positive.

– Object level: Since pixel-level anomaly check for 40% overlap of predicted anomalous pixels with that of ground-truth region, setting higher true positive rate may result into large false positive rate. Therefore, object level anomaly localization [26] define true positive rate by setting threshold Θ shown in Eq. 1.

$$\frac{Detected\ anomaly \cap True\ anomaly}{Detected\ anomaly \cup True\ anomaly} \geq \Theta \tag{1}$$

3 Traditional approaches for anomaly detection

Anomaly detection from video is the widely investigated research topic since decade. Various frameworks for tracking, surveillance, and anomaly detection in different domains have been put forth till date for commercial use. IBM Smart Surveillance System (S3) [95] is the world's first, event based, distributed middleware to be used in surveillance system for video based behavioral analysis, automatic scene monitoring, event based retrieval and real time event alert system. PFinder [111] and W^4 [33] are two systems applicable to be used for human behavior tracking. PFinder tracks and interprets human behavior activities and it is applicable to be used in video databases, wireless interfaces. W^4 [33] system is operated by monocular video imagery and works in an outdoor environment for detection and tracking of multiple interacting people with other objects. Mobileye [116] is commercially available system for vehicle tracking in the domain of transportation. It is used for detection of large objects lying over small distances using monocular camera, however, its detection is limited to certain classes of objects (vehicles and pedestrians). Fujitsu's Intelligent Transportation System [27] works at 30 FPS running on Intel Xeon 3.2 GHz with 4 GB memory for detecting and tracking vehicles and other entities at real time. Knight [86], automated surveillance system developed at University of Central Florida works with multiple cameras for detecting and tracking the objects. Apart from monitoring the sterile and dangerous zones, this system also summarizes the key frames in video and delineates the textual information of the trajectories for the ease of monitoring personnel.

Monitoring and tracking the human behavior and finding out the anomalies from surveillance video are well investigated topics since decade. But, this topic has been investigated more from the point of view of detecting anomalies without regard of how much length of videos are buffered before processing starts, support for online learning, and time it takes to detect and classify the anomalies. Very few papers serve as candidate for real-time processing oriented models [55, 78]. As video based anomalous activity detection is the promising area in Computer Vision, it has great applicability to be deployed in public places. Deploying such

systems at public places requires the practical constraints like how much amount of video the system buffers before processing, time required for detecting and classifying anomalies so that detection would be performed in a timely manner to avoid mishaps proactively. It also requires how much frames are processed per second, speed of streaming video and running time of anomaly detection algorithms. Considering this factors into account, this paper classifies the anomaly detection approaches into accuracy-oriented and real-time processing oriented approaches. This would facilitate how to modify the traditional approaches to real-time one or develop new models focused on real-time processing so that these models are readily deployable in real life scenarios for anomaly detection. So, state-of-the-art approaches of video based anomalous activity are mainly divided into 2 categories as accuracy-oriented approaches and real-time processing oriented approaches. The aim of accuracy-oriented approaches is to detect and localize anomalies with a focus on accurately detecting the anomalies, whereas, real-time processing oriented approaches focus on online processing of video frames in order to detect anomalies in real time manner. The classification of video based anomaly detection approaches is depicted in Figure 5. As majority of the traditional approaches have focused on accuracy-oriented methods for anomaly detection, this paper covers real-time processing oriented anomaly detection approaches both using traditional and deep learning methods. But for comparing trade-off among accuracy and real-time, significant accuracy-oriented approaches are also mentioned.

3.1 Real-time processing oriented approaches

The state-of-the-art traditional anomaly detection approaches are divided into 2 categories viz. Local feature modeling methods and Holistic feature modeling methods. Local feature modeling methods learn the model based on local visual features to represent the events and apply statistical, computer vision based techniques for detection of anomalies. This method assumes video as a collection of entities. Holistic feature modeling methods assumes the entities in the video as a whole and performs anomaly detection based on modeling the holistic features like motion and density.

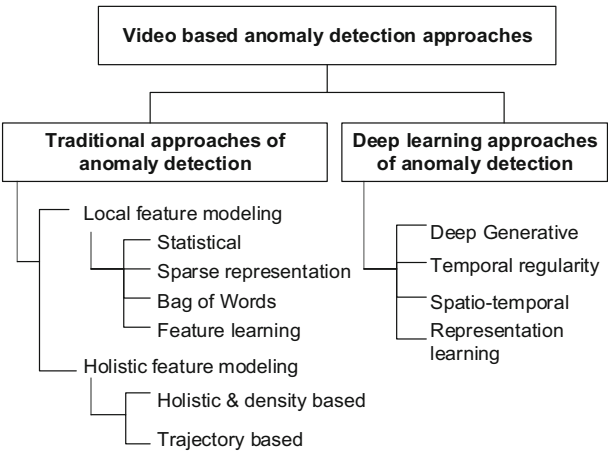

Figure 5 Classification of video based anomaly detection approaches

3.1.1 Local feature modeling methods

Statistical approach The statistical method for describing anomalous activity is given as follows [85]. In this, ℓ features are distributed with a probability density function (pdf),$g_0(.)$, if they come from a nominal distribution. Anomalous instances are distributed with pdf, $g_1(\ell)$. So, problem of anomalous activity detection amounts to predicting whether an instance ℓ is distributed according to nominal or anomalous pdf. This is given in an Eq. 2.

$$H_0 : \ell \sim g_0(.) \text{ versus the alternative (anomaly) } H_1 : \ell \sim g_1(.) \tag{2}$$

If both pdfs are either known or can be estimated from training data, this task reduces to the well-known Likelihood Ratio Test (LRT).

Non-parametric methods are not dependent on parametric model for learning motion and appearance based features from video, and directly learns scene normality based on descriptor instances. This approach uses normal data to train itself and follows unsupervised method of training. Bertini et al. [6] put forth spatio-temporal volumes (STV)-based non parametric method for detecting anomalies using unsupervised learning. To support multi-scale analysis, descriptors are computed independently at different scales even if they are overlapped. For detecting contextual anomalies, likelihood of descriptor based on neighboring cells is calculated. To detect whether video stream contains anomalous event, range query is applied on the training data to check the neighboring cells. This is done using fast approximate nearest-neighbor search built over k-means trees.

Apart from videos, anomaly detection is performed on hyperspectral imagery. Most approaches working on hyperspectral imagery follow the statistical approach in which the statistical model is built for background image and anomaly rate is calculated based on deviation of the mean from the model [77]. Though this approach is unsupervised, hyperspectral data merely satisfies the requirements of this approach. Therefore, Olson and Doster [68] came up with approach to model the background. They have combined the kernel Principal Component Analysis (PCA) with sub-sampled image and calculated reconstruction error as a measure of anomaly detection. This method can be improved by jointly modeling the spectral and spatial information of the hyperspectral imagery.

Sparse representation approach Lu et al. [55] put forth framework based on sparse combination learning for speedily detecting the anomalies from surveillance video. The use of small-scale least square optimization shortens running time for detecting the anomalies.

Bag of words (BOW) approach Roshtkhari et al. [78] handled the problem of detecting contextual anomalies from video using probabilistic framework by measuring the likelihood of STVs. New normal events are learned incrementally using online and unsupervised learning. For faster detection of anomalies redundancy of spatio-temporal volumes is curbed by grouping STVs using codebook construction and in turn, reducing the search time for comparison of newly observed data with previously stored STVs.

Contextual information obtained from spatio-temporal volumes of video cubes is used for detection of global and local anomalies in [53]. Activity pattern codebook is constructed to infer global information from video whereas composition pattern dictionary is used to infer salient patterns in STVs. Sparse reconstruction model built over learned dictionary is used for

anomaly detection. In this paper, multi-scale analysis method is used for accurate localization of anomalies in the video.

Cheng et al. [13] applied one class Support Vector Machine (SVM) with Bays probability to detect anomalies from video and maximum subsequence search method for anomaly localization. In this, events in video are represented using 'subsequence' - subsequence of time series based spatial windows present in the proximity of each other. Though the approach achieves comparable performance in terms of faster processing, anomaly of small-scale and having short duration of occurrence is not detected.

Feature learning approach For real time detection of anomalies from video, Wang et al. [105] used low-level statistical features instead of relying on complex machine learning and computer vision algorithms. This approach is not suitable for low density crowd scenes in which behavior of individual entity is anomalous.

Leyva et al. [49] used optical flow features and foreground occupancy features for extracting descriptive features from cell structure. After extracting compact set of features, models like Gaussian Mixture Model (GMM), Markov chains and BOW are used for anomalous video volumes. Finally, inference mechanism is used for detection of anomalous activity using neighborhood cells described by local spatio-temporal features.

3.1.2 Holistic feature modeling methods

Holistic & density based approach Marsden et al. [59] used holistic and density based approach for crowd anomaly detection. In this, they have put forth scene-level holistic features in terms of 4 dimensions as crowd conflict, collectiveness, motion speed and density. They have used 2 classifiers based on availability of anomalous data. GMM is used for anomaly detection only when normal training data are available. SVM – discriminative model is used when both normal and anomalous behavior data are available. In this paper, authors used cross scene training, i.e. for detecting anomalies in UMN datasets, training frames of other datasets are used to generate Gaussian Mixture Model (GMM).

Trajectory based approach Motion instability defined in terms of direction randomness and motion intensity has been used for discriminating anomalous behavior from normal one using unsupervised approach [113]. This framework is useful for understanding how previously occurred observed patterns are deviating, but fails to detect appearance-based anomalies. In order to perform faster processing, feature tracking scheme is employed in this approach.

The majority of algorithms for anomaly detection work on decompressed videos containing pixel-level information. In pixel domain, complex feature extraction process requires huge amount of data and lowers the speed of execution. This problem of decompressed videos worsens when thousands of long duration decompressed video data are generated. Biswas and Babu [7] came up with new approach of utilizing motion vector cues from H.264/AVC compressed videos. Hierarchical processing of video frames is carried out using pyramid structure, namely motion pyramids. In this, initial processing is done at coarse level and if anomaly is occurred, then processing

moves to finer level i.e. actual frame level. This method of hierarchical processing reduces computational overhead, and thus detects anomalies at real time. This approach works well only when motion is encountered in video and therefore can't detect appearance anomaly since it is purely based on motion vector.

3.2 Accuracy oriented approaches

The substantial amount of work has been put forth in accuracy oriented approaches for anomalous activity detection. Some of significant works are based on Mixture of Dynamic Textures (MDT) [51, 57], sparse representation technique [63], Gaussian process regression [12], cascaded Hidden Markov Models [106], context-dependent approaches [123].

Most of the mentioned papers [12, 51, 57] train the model with normal videos and build normalcy model. During testing phase, anomalous videos are introduced to check the effectiveness of the model. This is known as unsupervised learning. Out of these, [63] uses both supervised and unsupervised learning method for detection of anomaly. In case of supervised learning, anomalous videos are also used during training phase for improving the accuracy of detection. Similar to [63], weakly supervised learning strategy is followed in [35]. In this paper, anomalous videos are used in training phase. Both multi-instance learning model and dictionary learning approach are used for anomaly detection.

3.3 Comparison of real-time processing oriented approaches and accuracy-oriented approaches

Real-time processing oriented approaches rely on online learning i.e. time required for frame processing is shorter than the time for processing the next frame in the sequence. Such approaches continuously update themselves for identifying whether a newly observed event is anomalous or not. In this, model parameters are updated incrementally based on new training data.

On the other hand, accuracy oriented approaches use offline algorithms which assume that all the training data are available in the outset. These methods use fixed parameters, predefined anomaly thresholds or fine-tuned thresholds obtained from training of batch data. And therefore, can't be used for real time detection of anomalies.

The traditional approaches of anomaly detection rely on hand-crafted features for extracting features from video frames, and require expertise to design the methods for feature engineering**.** Such hand-crafted features being suboptimal are very specific to the given scenario. Manually hand-crafted features are incapable to be used in cross domain datasets and exhibit poor support for inferring semantic information. Such approaches can't be generalized to work with scenes having unknown anomalies, adverse lighting conditions and drastic variation in motion and appearance of entities in video. On the contrary, deep learning architectures like Convolutional Neural Networks (CNNs), automatically learns and selects the features. Deep learning approach has ability to be generalized across multiple datasets i.e. once learning is done on one dataset, the pre-trained deep neural network can be applied to other dataset, called as Knowledge Transfer. Figure 6 the shows the difference between traditional machine approach and deep learning approach.

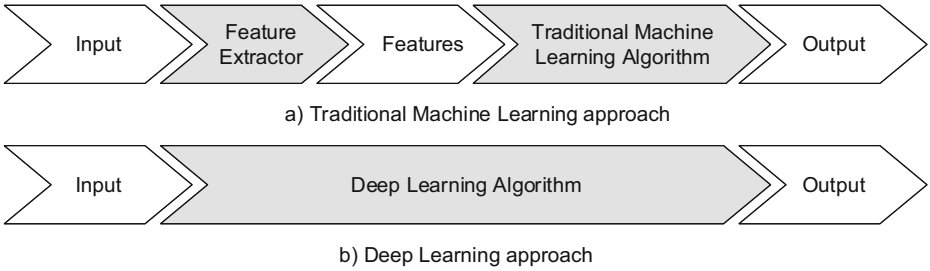

Figure 6 Traditional Machine learning versus deep learning methodology

4 Deep learning approaches for anomalous activity detection

The availability of large datasets and high availability of GPUs at lower costs has resulted into proliferation of deep learning techniques. State-of-the-art results have been achieved for image classification [47], object detection [69, 70, 76], activity recognition [88, 108] egocentric activity recognition [109], video hashing [91] and video captioning [29] using deep learning. The reason behind success of deep learning approaches is that non-linear transformations allow extracting useful and complex features from high dimensional data like video. This triggered use of deep learning techniques for anomaly detection from videos.

As the use of deep learning approaches for anomaly detection approaches is still in nascent stage, deep learning approaches from the viewpoint of both accuracy and real time are described in this paper. Various deep architectures have been used for anomaly detection viz. CNNs [47], Long Short Term Memory Networks [38], Auto-encoders (AE), and Generative Adversarial networks (GANs) [31].

Recently, vanilla deep models have been modified to solve the specific problem in hand. For example, 2D CNNs or 3D CNNs are used for automatically describing the videos. Moreover, to capture the temporal and spatial dynamics of long duration videos, different variants of LSTMs have also been put forth [32, 54]. Temporal and Spatial LSTM (TS-LSTM) put forth by Guo et al. [32] is a good candidate solution for detecting anomalies from video having long duration. It can be noted that CNNs (ability to extract features automatically) and generative models (ability to reconstruct the input pattern) have been widely used for anomaly detection.

The overview of variants of auto-encoders used in generative models is given here since most of anomaly detection approaches are based on AEs.

The beauty of generative models is that they learn the distribution of data and accordingly predict the future sequence of frame. Based on this principle, reconstruction error is generally used for calculating anomaly score. Each data instance x_i is reconstructed with help of learned network. The reconstructed output is given by o_{ij}. Then reconstruction error is calculated as follows.

$$\delta_i = \frac{1}{n} \sum_{j=1}^{n} \left(x_{ij} - o_{ij}\right)^2 \tag{3}$$

In the above equation, reconstruction error is denoted by δ_i, n denotes number of features for defining data. The reconstruction error δ_i, gives an anomaly score. The learned auto-encoder reconstructs the motion signatures from normal videos with low error but can't accurately reconstruct motions from anomalous videos. In other words, the auto-encoder is used for modeling the distribution of the regular dynamics of appearance changes. Generative

models generally assume that the features come from a predetermined type of distribution and therefore are likely to fail if the feature distribution changes.

- Sparse AE (SAE): Sparse auto-encoder is used for handling the transfer learning problem.
- Denoising auto-encoders (DAE): The Denoising auto-encoder is the extended stochastic version of auto encoder. Auto-encoder can be converted to denoising auto-encoder by adding stochastic corruption layer at the input of auto-encoder. For discovering robust features and refrain hidden layer from learning the identity of the input, this auto-encoder is trained to reconstruct the input while preserving the information of input and undoing the effect of corrupted version of input. Denoising auto-encoder predicts the missing (corrupted) values from the non-missing values (uncorrupted) for the given pattern. This model requires manually devised noise model for training. In case of training with unsupervised setting, it is very difficult to choose effective noise model for training.
- Stacked denoising auto-encoders (SDAE): Denoising Auto-encoders stacked together forms the initialization of deep architectures. They have been extensively used for new representation from videos. In order to denoise the corrupted (missing) values of the inputs, denoising auto-encoders are trained locally.
- Marginalized Denoising auto-encoder (MDA): SDAE require heavy computational processing and due to this, they are not efficient for large scale video analytics. Marginalized SDAE (mSDAE) handles this issue of computational processing and scalability to support high dimensional data. mSDAEs does feature learning in a faster way by using single layer structure of auto-encoder and achieves balanced trade-off among performance and speed.
- Cascaded stacked auto-encoder: Stacked auto-encoder with more than one layer is known as cascaded stacked auto-encoder. Stacked auto-encoders are useful for unsupervised feature learning.
- Generative Adversarial Networks (GANs): GANs are used for generating data and follow unsupervised learning. GAN can be considered as zero-sum two-player game. It consists of 2 different networks namely generator G and discriminator D. During training phase, generator's task is to generate data (images) whereas discriminator's task is to discriminate the generated data i.e. identify whether data are real or generated from G. In this way, discriminator is trained to output correct results.
- Conditional GANs: GANs can be modified to get conditional GAN by adding condition c as input to both generator G and discriminator D.
- Convolutional Winner-take-all encoder (CONV-WTA): The CONV-WTA [58] uses unsupervised approach for learning the sparse representations in a hierarchical manner. It is non-symmetric in nature. Encoder part is built by stacking multiple CNN based Rectified Linear Unit (ReLU) layers whereas decoder is built using linear deconvolutional layer.

4.1 Deep learning based Real-time processing oriented approaches

Initially, being the base of deep learning architectures, neural network models have been employed for real time anomalous activity detection [79, 80]. Sabokrou et al. [79] used independent feature learning method to model video using local and global descriptors using sparse auto-encoder model in unsupervised way. They have used Gaussian distribution to model normal video patches and Mahalanobis distance to denote anomaly measurement. On the similar lines of previous work, Sabokrou et al. [80] put forth two anomaly detectors based on auto-

encoder and sparse representation of video. As AE gives more reconstruction error for anomalous patch and sparse representation of video implies chances of anomalies, the cascaded effect of both detectors is used to detect anomaly at real time and achieves 120 FPS performance on UCSD ped 2 dataset using MATLAB 2015 running on 3.5 GHz CPU and 16 GB of memory.

4.1.1 Deep generative model

Fully convolutional neural networks have been used for anomaly detection for the first time by Sabokrou et al. [83]. This method uses transfer learning approach for extracting features using patch-operations from video frames with the help of CNN pre-trained on AlexNet [47]. The extracted features are represented using sparse auto-encoders whereas Gaussian model is used to evaluate anomalies. For automatically representing video frames and inferring appearance and motion cues, 3D gradients obtained from PCANet are used and normal event is modeled using deep GMM [26]. Deep GMM being scalable and generative in nature is constructed by stacking multiple layers of GMM. For time efficient and accurate anomaly localization, deep cascade approach based on competitive cascade of deep neural networks has been put forth by Sabokrou et al. [81]. This approach works for real time anomaly detection in surveillance systems. It combines two stages of deep stack auto-encoder and CNN. Intermediate layers of CNN or stack auto-encoders act as sub-stages of a cascaded classifier. For achieving time-efficient anomaly detection, shallow layers of cascaded DNN are used to detect background normal patches whereas complex patches in the neighborhood of simple patches are detected by deep layers. CNN is merely fine-tuned in this method and trained from scratch. Following the cubic patch based approach based on cascaded classifies, Sabokrou et al. [82] used local and global video descriptors for representation of video. Structural Similarity Metric (SSIM) is employed to check the similarity among the patches. Two one-class classifiers for each descriptor are used for anomaly detection based on weakly anomalous patches and strongly anomalous patches.

Wu et al. [112] used two stream network and Variational Autoencoder/Generative Adversarial network for detection and localization of anomalies. The beauty of this system is that it is based on client-server architecture and provides users with input channel for uploading the local videos and also accepts input stream of video using online mode.

4.1.2 Spatio-temporal model

The lack of large set of anomaly representations for training, most anomaly detection approaches follow unsupervised learning [113, 114]. The frameworks put forth by Giorno et al. [21] and Ionescu et al. [99] work on unsupervised learning method when no training data are available. Online anomaly detection based on unmasking technique is done in [99]. It works on the principle of change detection. The unmasking technique involves binary classifier which iteratively distinguishes between consecutive video frames while removing the discriminating features from frames iteratively. The higher degree of training accuracy shows the presence of anomaly.

4.2 Deep learning based accuracy oriented approaches

4.2.1 Temporal regularity model

This model focuses on evaluation of CNN features across time to capture local anomalous events from videos [74]. Ravanbakhsh et al. used CNN model pre-trained on object

recognition to detect anomalies and employed two-channel approach for representing video in terms of appearance and motion (optical flow) similar to [43, 57]. They put forth TCP (Temporal CNN pattern) network in which Binary Quantization layer is placed as the last layer of CNN to represent temporal motion patterns for anomaly segmentation. But, TCP network is not end-to-end trainable and suffers from heavy post-processing and require previously computed codebook of convolution feature maps.

Anomaly detection based on sparse coding approach [55, 120] involves building a dictionary over normal events associated with small reconstruction error, whereas anomalous events would result into large reconstruction error. Optimizing the sparse coefficients is time consuming and leads to bottleneck for dictionary learning. In addition, the neighboring frames (temporally related) are assigned different sparse coefficients which leads to loss of temporal coherence between those frames [55, 57].

In order to retain the locality information between the neighboring frames, temporally coherent sparse coding based method (TSC) has been put forth in [56] in which similar neighboring frames are encoded with same sparse coefficients. This TSC is mapped to its equivalent representation using stacked RNN (sRNN). Optimization in the parameters of Stacked RNN alleviated the need to select the hyperparameters in TSC and expedites the anomaly prediction due to use of shallow architecture.

4.2.2 Spatio-temporal model

Zhou et al. [122] pioneered the use of spatio-temporal CNN for anomaly detection and localization for the first time. Fang et al.'s spatio-temporal anomaly detection model is inspired from saliency information obtained from videos [24]. This model represents spatial information (SI) obtained from salient regions of frame. Temporal motion aspect is represented by multi-scale histogram optical flow (MHOF). Deep learning network – PCANet is used to obtain features from SI and MHOF for anomaly detection.

Once the anomalous events are detected, then it is also necessary to explain why the event is judged as anomalous. This is called as event recounting of anomalous activities. The approach put forth by Hinami et al. [37] performs joint detection and recounting of anomalous events by amalgamating the multi-task Fast-RCNN (MT-FRCN) and environment-specific anomalous event detector. Currently, semantic knowledge used for explaining the anomalies is restricted to actions. This deep knowledge of visual concepts can be extended to explain more complex object interactions occurring in the anomalous events. Sun et al. [93] fused one-class SVM (OC-SVM) with CNN for designing end-to-end trainable model for anomaly detection. For modeling the velocity and direction of entities in video, optical flow features are fed as input to CNN. Their Deep One Class (DOC) model equipped with Radial Basis Function (RBF) yields robust anomaly detection.

4.2.3 Representation learning model

Hu et al. [41] used deep incremental slow feature analysis network (D-IncSFA) to learn high level abstraction from video and detect anomalies in one step. Global anomaly detection is done using temporal modeling and local anomaly detection using multi-scale analysis based on summed squared derivative (SSD) value. This approach does not reply on any classifier model and does not use hand-crafted feature representation.

Though deep learning models are good at extracting high level abstraction from the given video, it is difficult to model regression tasks since labels do not hold enough capability to fine-tune learning parameters. This problem has been addressed by deep metric learning (DML) based on regression applicable to density based approaches [107]. DML not only extracts density based features but also learns better distance measurement. Currently, this method is shown to be applicable in congestion detection and crowd counting. But, it is still difficult to train deep networks even if guided DLM is used.

4.2.4 Deep generative model

To learn the temporal dynamics in long-hours of video, end-to-end-trainable framework is developed using convolutional auto-encoder having ability to learn local features and classifiers [34]. Its working is based on following principle. Auto-encoder learns complex distribution of normal patterns in video and reconstructs motion in normal patterns with low error and does not reconstruct motion patterns in anomalous frame of video. The reconstruction error between real frame and reconstructed frame gives the anomaly score. Similar to this approach, Medel and Savakis [60] replaced the weights in fully connected LSTM with convolutional filters to yield Conv-LSTM architectures and used it for predicting near future terms by encoding and reconstructing video sequence. The predictive capability of network used in regulatory evaluation algorithm detects the anomalies. Xu et al. [114] used three stacked denoising auto-encoders (SDAE) to learn the joint representation of appearance and motion and three one-class SVMs for calculating the anomaly scores. Use of optical flow maps makes this method to capture only the short term motion and can't handle long-term temporal motion to infer useful regular pattern from video. Apart from this, contextual information required for finding relation between consecutive video shots is missing in their approach. Therefore, Feng et al. [25] came up with another approach based on SDAE to handle the problem of short term clues and contextual anomalies. They used LSTM for capturing long-term motion cues from video and Graph-based manifold ranking scheme to reduce false alarms from spatial contextual information.

In order to automatically learn the feature representations from video in unsupervised manner, Xu et al. extended their previous work [114] and put forth Appearance and Motion DeepNet (AMDN) approach based on Stacked Denoising Auto-Encoders (SDAE) [115]. The crux of the approach is double fusion scheme which performs joint representation of appearance and motion characteristics of video and this scheme does not rely on object level analysis. AMDN is not suitable for real time applications due to its heavy computational processing. In addition, this scheme uses shallow networks and small image patches as network input, so there are chances of overfitting on small scale data. As multiple one-class SVMs built over learned features are not jointly optimized with anomalous activity discrimination task, the learned features may be suboptimal.

Dearth of anomalous ground truth data and ambiguous nature of anomalies hinders the development of end-to-end trainable deep learning model. This issue has been addressed using conditional Generative Adversarial Networks [75]. As claimed by authors, it is a pioneer work using GAN for anomaly detection for the first time. The main feature of this end-to-end trainable deep learning model is that it uses cross-channel approach to refrain generator from learning identity function and uses multi-channel representation for fusing appearance and motion information.

Tran and Hogg [96] used Convolutional Winner-Take-All (WTA) [58] and one-class SVM for anomaly detection. In this, convolutional auto-encoder is used for extracting the motion

features and OC-SVM is used for building the normalcy model. This framework is based on motion feature representation. It can be extended by introducing mechanism for appearance feature modeling and also methods for modeling motion patterns of longer length.

4.2.5 Hybrid model

Inspired by the success of 3 dimensional CNNs [97], Zhao et al. [121] put forth hybrid approach for anomaly detection. They used 3D for modeling the spatio-temporal features for surveillance video and jointly utilized the reconstruction loss (for reconstructing the input frame) and weight decreasing prediction loss (for predicting the future frame) of auto-encoder for detection of anomalies. As the approach works on predicting frames, sudden appearance of objects in the field of view may hinder the performance of this model.

4.3 Comparative study

Tables 1 and 2 shows the comparison of selected traditional and deep learning approaches. The comparison of real time anomaly detection approaches using both traditional and deep learning techniques is done in Table 3. Please note that values mentioned in the Table 3 are taken from results mentioned in the corresponding research papers. It can be observed that deep learning technique has achieved highest performance for frame per second on the datasets. Though it is not verified from single research article, there is a great scope for deep learning to excel in anomaly detection tasks.

5 Application domains and benchmarked datasets

5.1 Discussion of application domains

Though there are enormous domains where anomaly detection can be applied, Figure 7 shows some of widely investigated scenarios in which research related to anomaly detection can be carried out. These domains include traffic, transportation, sports, crowd scenarios, health care domains, naturally occurring/manmade calamity detection, industrial domains, wildlife scenarios, etc. Some of the working use cases of anomaly detection are explained here. Periodical railway inspection in order to avoid railway mishaps is a *part and parcel* of safer railway transportation. Detection of obstacles, missing fastening bolts (by which rail is fixed to the sleepers), status of switches and other railway defects can be detected in real time manner [22]. This eliminates the need of human expertise to walk along the track to identify visual anomalies. Autonomous driving on urban highways or mountain regions is very challenging. Timely detection of anomalous objects guarantees to curb the chances of accidents and ensures the safety of people on highways [17]. Detecting unattended objects in timely manner is essential for maintaining enhanced security at public places to curb the chances of terrorism [67]. Anomaly detection is indirectly related to crowded scene analysis. This includes congestion detection, crowd counting [107]. Crowd counting and timely detection of congestion due to traffic, processions, or at pilgrims helps to avoid mishaps by applying proactive measures to control the crowd. This would ultimately help to avoid crowd disasters [117]. One of the applications of anomaly detection for immediately taking an action in elderly fall incidents [23].

Table 1 Deep learning approahces for anomalous activity detection

Deep learning approach	Ref., Year	Deep architectures	Technique	Anomaly formulation	Anomaly measurement			Perf. metric	Dataset
					Frame	Pixel	DPL		
Real-time processing oriented approaches									
Generative model	Deep-cascade [81], 2017	Stacked AE + CNN	Cubic patch based, cascaded classification	Mahalanobis distance between test patch & Gaussian model	√	√	√	EER, ROC, FPS	UCSD ped 1 & ped 2, UMN
Generative model	Deep-Anomaly [83], 2016	Fully CNN + Sparse AE, pre-trained AlexNet	Cascaded outlier detection, transfer learning	Mahalanobis distance between test patch & Gaussian model	√	√	–	ROC, EER, AUC, FPS	UCSD ped 2, subway
Generative model	Sabokrou et al. [82], 2017	Sparse AE	Cubic patch based, unsupervised feature l earning, cascaded classification, SSIM	Mahalanobis distance between test patch & Gaussian model	√	√	√	ROC, AUC, FPS, EER, DR	UCSD ped1 & ped2, UMN
Spatio-temporal model	Unmask [99], 2017	CNN, pre-trained VGG-net [11]	Change detection based Unmasking technique	Measure of Training accuracy obtained over classifying the consecutive frames	√	√	–	ROC, AUC, FPS	UCSD ped1 & ped2, Subway, Avenue, UMN
Accuracy oriented approaches									
Generative model	Hasan et al. [34], 2016	Convolutional AE, pre-trained ZFNet [119]	Temporal modeling and sliding window	Regularity score obtained by reconstruction error	√	√	–	EER, AUC	UCSD ped1 & ped2, Subway, Avenue
Generative model	AMDN [115], 2017	Stacked denoising AE	sliding window, fusion strategies	Patch-based binary categorization problem	√	–	–	EER, AUC, ROC	UCSD ped1 & ped2, Subway (exit)
Generative model	cGAN [75], 2017	Conditional GANS, pre-trained ZFNet [119]	Multi-channel & cross channel data representation, Adversarial Discriminator	Reconstruction error	√	√	–	EER, AUC	UCSD ped1 & ped2, UMN
Temporal regularity model	TSC & sRNN [56], 2017	Spatial CNN + Stacked RNN	Temporally coherent sparse coding, Iterative Soft-threshold for coefficient optimization	Regularity score obtained by reconstruction error	√	√	– –	ROC, AUC	UCSD ped 2, CUHK Avenue, ShanghaiTech Campus
Spatio-temporal model	MT-FRCN [37], 2017	Fast R-CNN [30]	Region proposal generation, and Generic knowledge learning	Threshold criterion for anomaly score of object proposal	√	√	– –	ROC, EER, AUC	UCSD Ped2, Avenue
	DOC [93], 2017				√	√	– –		UCSD ped1 & ped2

Table 1 (continued)

Deep learning approach	Ref., Year	Deep architectures	Technique	Anomaly formulation	Anomaly measurement			Perf. metric	Dataset
					Frame	Pixel	DPL		
Spatio-temporal model		Convolutional Neural Network	Dense optical flow, One-class SVM with RBF Kernel	Deviation of pixels from Hyperplane				ROC, AUC, EDR	
Generative model	Hasan et al. [34], 2016	Convolutional AE, pre-trained ZFNet [119]	Temporal modeling and sliding window	Regularity score obtained by reconstruction error	√	√	– –	EER, AUC	UCSD ped1 & ped2, Subway, Avenue
Generative model	Tran et al. [96], 2017	Convolutional Winner-Take-All (WTA) Auto-encoder [58]	Local normality modeling and unsupervised SVM (One-class SVM)	SVM based threshold criterion applied to optical flow patches	√	√	– –	ROC, EER, AUC	UCSD ped1 & ped2, Avenue
Hybrid model (Spatio-- temporal + Generative)	STAE [121], 2017	3D CNN + Auto-encoder	Weight decreasing prediction loss	Threshold criterion based on Euclidean distance	√	√	– –	ROC, EER, AUC	UCSD ped1 & ped2, Avenue

Table 2 Traditional approaches for anomalous activity detection

Approach	Ref., Year	Video representation	Technique	Anomaly formulation	Anomaly Measurement			Perf. metric	Corpus
					Frame	Pixel	DPL		
Real time processing oriented approaches									
Statistical	Bertini et al. [6], 2012	Spatio-temporal volume (STV)	Multi-scale non-parametric approach, nearest-neighbor search over k-means tree	Cumulative distribution based nearest-neighbor distances	√	–		ROC, EER, FPS	UCSD ped1 & ped2
Statistical	Biswas & Babu [7], 2013	Motion pyramids	behavior modeling by histogram of motion magnitudes, dense statistical modeling	Probability of occurrence of behavior	√	√		EER, AUC, FPS	UCSD ped1 & ped2, UMN
Sparse representation	Lu & Jia [55], 2013	Spatial-temporal cubes	Small-scale least square optimization, sparsity constraint for combination size	Scale-wise summation of cubes gives anomaly measure	√	√		EER, EDR, AUC, FPS	UCSD Ped1, Avenue, Subway
Bag of Words	STC [78], 2013	Spatio-temporal volumes	Statically learning behavior method based on dense local spatio-temporal features	Reconstruction process (Events not similar to observed events are anomalies)	√	√		ROC, FPS	UCSD ped1 & ped2, Subway, Weizmann
Holistic & density based	Marsden et al. [59], 2016	4 dimensional features a	Scene-level holistic features, tracklet and feature extraction, cross-scene anomaly detection method	Distribution of log probabilities signifies anomaly detection based on Otsu's threshold selection method	√	√	–	FPS, ROC	UMN, Violent flow
Feature learning	Leyva et al. [49], 2017	Coarse-to-fine (variable sized) cell structure	Feature learning based on structure similarity, sparse auto-encoder, descriptor based similarity metric	Threshold criterion based on Mahalanobis distance	√	√	√	FPS, ROC, AUC, EER	UCSD ped1 & ped2, UMN, LV
Neural Network	SAE [79], 2015	Non-overlapping cubic patches with local and global descriptors	Feature learning based on structure similarity, sparse auto-encoder, descriptor based similarity metric	Threshold criterion based on Mahalanobis distance	√	√	√	FPS, ROC, EER, AUC	UCSD ped2, UMN

Table 3 Comparative study of real time performance of anomaly detection approaches based on traditional and Deep Learning (DL) methods (Values mentioned in the table are directly taken from corresponding references)

Paper	DL?	Approach	Platform	Performance in Frames per Second on datasets				
				UCSD Ped 1	UCSD Ped 2	Avenue	Subway	UMN
SAE [79]	No	Neural network	MATLAB 2012a, 3.5 GHz CPU, 8GB RAM	–	25 FPS	–	–	–
[80]	No	Neural network	MATLAB 2015, 3.5 GHz CPU 16GB RAM	–	120 FPS	–	–	–
[55]	No	Sparse	MATLAB 2012, 3.4 GHz CPU, 8GB RAM	143.57 FPS	–	141.34 FPS	155.97 FPS	–
[7]	No	Statistical	MATLAB, 3.4 GHz CPU	70 FPS	70 FPS	–	–	150 FPS
[105]	No	Feature	2 × 3.07 GHZ CPU, 4 GB	55.55 FPS	55.55 FPS	–	–	62.5 FPS
[83]	Yes	Deep Generative	Caffe library, MATLAB 2014a, NVIDIA TITAN GPU	–	370 FPS	–	–	–
[81]	Yes	Deep Generative	MATLAB, Intel core i5 CPU 2.4 GHz, NVIDIA GT 620, 8 GB RAM	15 FPS (Using CPU)	130 FPS (Using GPU)	–	–	–
Unmask [99]	Yes	Spatio-temporal	Intel Core i7 2.3 GHz CPU, 8 GB RAM	–	20 FPS	–	–	–
STC [78]	No	BoW	Intel Q9550 CPU, 4GB RAM	5.26 FPS	–	–	4.16 FPS	–
LV [49]	No	Feature learning	MATLAB, 2.7 GHz CPU, 8 GB	32.25 FPS	–	–	–	32.25 FPS
[113]	No	Trajectory	C++ OpenCV, 3.3 GHz CPU, 4 GB RAM	23.80 FPS	–	–	–	–

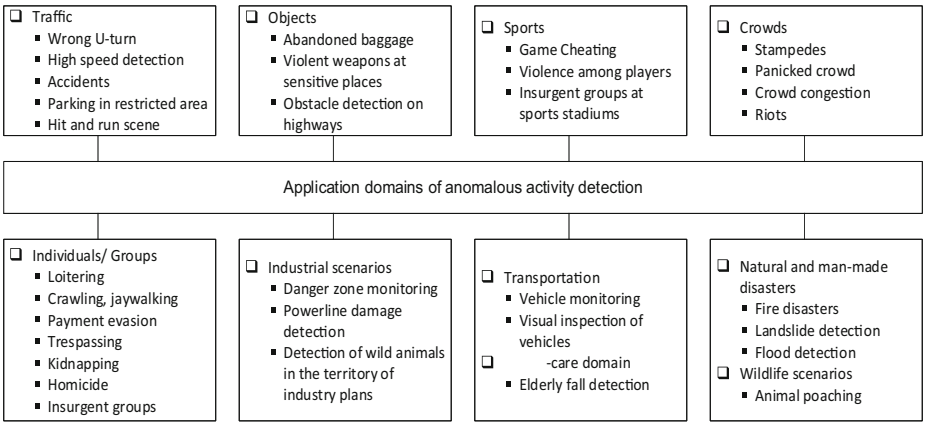

Figure 7 Application domains of anomaly detection

5.2 Public datasets for indoor and outdoor surveillance

Considering the requirements of real-life scenarios, various datasets for anomaly detection have been put forth till date obtained from indoor and outdoor surveillance. Table 4 shows the widely used datasets for anomaly detection. The datasets are compared based on the features, scenarios covered for anomaly detection, availability of ground truth (GT) and the resolution of videos.

6 Research challenges and future directions

Automated video surveillance for detecting anomalous activity has been a topic of great interest in computer vision and cognitive science for enhancing the security of indoor and outdoor places. Following identified research issues are still open in this domain and needs to be addressed for efficient detection of anomalous activities from videos. The list of issues and challenges is by no means exhaustive and continues.

- *Challenges related to indoor and outdoor environment:* Handling noise in video data due to the sensor, camera jitter and various video decoding artifacts, occlusion of independently moving objects, illumination changes, intra-class and inter-class variation of objects. Camouflage detection is also a major challenge.
- *Challenges related to scale at which normalcy model is defined*: It relates to multi-scale (resolution) of normalcy model, handling variation of normalcy model according to the anomaly to be detected
- *Challenges related to dearth of labeled anomalous behavior training dataset:* Due to dearth of labeled anomalous behavior training data, use of unsupervised learning is trivial option. There is a need of detection of anomalous activity based on lesser contextual information, and miniaturized size of training dataset.
- *Challenges related to trade-offs among performance metric:* Achieving balanced trade-off among real time processing and desired level of accuracy is very critical.
- *Challenges related to multi-view anomaly detection:* Target of interest seems to be normal from one view but exhibits abnormality if checked from another view. Substantial amount of work has been done in single-view anomaly detection. However, lesser work has been

Table 4 Benchmarked datasets of anomaly detection

Dataset	Features	Scenarios	GT	Resolution
UCSD Pedestrian 1 (Ped 1) and Pedestrian 2 (Ped 2) [100]	Use of stationary camera to capture the videos of pedestrian walkways where crowd density varies as sparse to crowded	Traversing of entities other than pedestrians across the pathway like skaters and vehicles on pedestrian walkway	Yes	158 × 238 240 × 360
UMN [101]	Sequence of normal behavior followed by anomalous one, Covers indoor and outdoor premises addressing occlusion	Unattended objects, Unusual behavior of crowd at indoor and outdoor premises, camera sabotage, movement in restricted area, loitering	No	320 × 240
Live Video (LV) [50]	Covers many domains from indoor/ outdoor premises, traffic, roadways	Robberies, wrong U-turns, crowd panic, loitering, fighting, trespassing, kidnapping, fire, driving in wrong direction, falling of people	Yes	176 × 144 1280 × 720
Avenue [18]	Covers traffic dataset and pedestrian dataset	Unusual behavior, walking in wrong direction, unattended object	Yes	640 × 360
Anomalous Behavior Dataset [3]	Scenarios with illumination, clutter, jitter, varied motion and appearance	anomalous scenarios involving boarding on and off the train, wrong direction	Yes	320 × 240
PETS' 09 [71]	Multi-sensor sequence of various crowd activities possessing calibration data	Crowd with variable density: walking, running, multiple flows of crowd, sudden dispersion, splitting	Yes	768 × 576 720 × 576
VIOLENT-FLOWS [103]	Covers datasets of both violent and non-violent activities of crowd collected from YouTube	Violence caused by people at public places like stadium	No	320 × 240
Weizmann [110]	Covers anomalous patterns in images and videos	Suspicious walking pattern, person walking with a gun, salient behavior	No	–
ShanghaiTech Campus [56]	Covers diverse anomalous scenes (approx. 13) captured from multiple cameras with different view angles under varying illumination conditions	Suspicious activities characterized by violent motions like brawling, chasing, skaters, bikers and trolley on the pedestrian walkways	Yes	846 × 480
CAVIAR [9]	Indoor premise covering entrance lobby and shopping center addressing appearance detection, occlusions	Single entities involving walking, falling, resting, roaming, abandoning the luggage and interaction anomalies involving fighting, people walking together and splitting	Yes	384 × 288
BEHAVE [5]	Video with understanding the behavior and interaction of the people	Interaction of people: approach each other, walk together, fight, run together, meet, split	Yes	640 × 480
QMUL Junction [73]	Covers the traffic Intersection at the junction	Wrong direction of vehicles	No	360 × 288

Table 4 (continued)

Dataset	Features	Scenarios	GT	Resolution
MIT Traffic [62]	Covers traffic video captured by stationary camera	Detection of pedestrian as anomaly on the public road	Yes	720 × 480
Subway Entrance & Exit [2]	Cover the view of underground train station both at entrance and exit gate	Avoiding turnstiles, wrong direction, loitering, suspicious interactions	Yes	512 × 384
i-Lids bag and vehicle detection challenge [42]	Subset of i-Lids dataset covering 2 scenarios of abandoned object and parking	Unattended objects, parking of vehicle in forbidden area	Yes	720 × 576

done in multi-view anomaly detection. It is very challenging to incorporate different levels of anomaly detection using multiple views in a single framework

- *Challenges related to camera anomaly detection:* Cameras used for surveillance are the basic sensors used for capturing the surveillance videos. Detecting the tampering, malfunctioning of surveillance camera in a real time i.e. camera anomaly detection has become the topic of research in recent years. The techniques for sabotage detection of cameras and self-improvement of camera's status needs further research investigation.
- *Anomaly detection from videos obtained from 360-Degree camera*: Till date, most approaches for anomaly detection used the video data obtained from statically positioned cameras [51], multiple cameras [56], and moving cameras [65]. With the advent of technology, the market of 360-Degree camera is anticipated to grow with CAGR of 34.4% from 2017 to 2024 [1]. This leads to need of detecting anomalies in the videos obtained from 360-Degree cameras. There is great scope for detecting anomalies from such videos since existing research is focused on videos from stationary and moving cameras only.
- *Convergence of frameworks aimed to detect anomalies from multiple domains:* Though biometric spoof detection (Detection of spoofed face, iris, and finger) is widely researched topic on its whole [4, 61], it can be integrated with video-based anomaly detection frameworks to improve the accuracy of detection especially in indoor environments. Apart from this, frameworks for camera anomaly detection used for detecting damages caused to the cameras can be incorporated with anomaly detection frameworks. This can be achieved by implementing interoperability measures among multiple detection frameworks to converge them into generalized anomaly detection platform.
- *Event recounting of anomalous activities:* The objective of Multimedia Event Recounting (MER) is to generate the summary of the events occurring in the given video clip [98]. Motivating from the success of MER on TRECVID datasets [28, 39], event recounting has been applied for anomaly detection in [37]. Justifying why the event is anomalous apart from detecting the anomalous event from video is still untouched research area except the work by [37]. Developing anomaly detection framework along with the visual concept representation for description of anomalous events is truly challenging due to ambiguous nature of anomalies and availability of semantic domain knowledge of anomalies, and deserves high scope for further research.
- *Use of deep learning technique to develop real time systems for anomaly detection*: Current literature still lags behind when it comes to process the videos at real time. The possible solution can be to follow the online learning methods to understand the anomalies at real time and incorporate online learning with deep models.

7 Conclusion

A proliferation of deep learning is changing the way of solving the real-world problems and anomaly detection is no exception. It is just the recent lustrum in which deep learning has been employed for anomalous activity detection, and this paper is an attempt to analyze and summarize deep learning techniques for video based anomaly detection in a nutshell and would act as profound research contribution for further investigation of deep learning for automated visual surveillance domain.

It can be understood that use of deep learning for anomaly detection has achieved remarkable results on both the accuracy oriented and real time processing oriented objectives

of anomaly detection. This research domain is very promising area since it will act as foundation stone in many future computer vision based projects like elderly fall detection (health care) systems, self-driving cars, robotics, and many more domains alleviating the need of human personnel for continuously monitoring the sensitive places.

Considering the deep learning aspect, there is much scope improved anomaly detection approaches by implementing parallel and distributed architectures models of deep learning. Motivated by the recent success of AlphaGo using deep reinforcement learning, use of deep reinforcement learning for online learning of anomalous activities and real time processing of anomaly detection would still be untouched research to be explored.

Acknowledgements Authors would like to thank anonymous reviewers for their valuable comments and guidance. This work has been supported by the Center of Excellence for Signal and Image Processing (CoE - SIP), College of Engineering Pune, India.

References

1. 360-Degree camera Market: https://www.researchnester.com/reports/360-degree-camera-market-global-demand-analysis-opportunity-outlook-2024/385
2. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 555–560 (2008)
3. Anomalous behaviour dataset: http://vision.eecs.yorku.ca/research/anomalous-behaviour-data/
4. Arashloo, S.R., Kittler, J., Christmas, W.: An anomaly detection approach to face spoofing detection: a new formulation and evaluation protocol. IEEE Access. **5**, 13868–13882 (2017)
5. BEHAVE: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/
6. Bertini, M., Del Bimbo, A., Seidenari, L.: Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Comput. Vis. Image Underst.* **116**, 320–329 (2012)
7. Biswas, S. & Babu, R. V.: Real time anomaly detection in H. 264 compressed videos. in *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on* 1–4 (2013)
8. Candamo, J., Shreve, M., Goldgof, D.B., Sapper, D.B., Kasturi, R.: Understanding transit scenes: A survey on human behavior-recognition algorithms. *IEEE Trans. Intell. Transp. Syst.* **11**, 206–224 (2010)
9. CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (2002)
10. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. *ACM Comput. Surv.* **41**, 15 (2009)
11. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference, {BMVC} 2014, Nottingham, UK, September 1–5, 2014 (2014)
12. Cheng, K.-W., Chen, Y.-T., Fang, W.-H.: Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. *IEEE Trans. Image Process.* **24**, 5288–5301 (2015)
13. Cheng, K.-W., Chen, Y.-T., Fang, W.-H.: An efficient subsequence search for video anomaly detection and localization. *Multimed. Tools Appl.* **75**, 15101–15122 (2016)
14. Chong, Y. S. & Tay, Y. H.: Modeling video-based anomaly detection using deep architectures: Challenges and possibilities. in *Control Conference (ASCC), 2015 10th Asian* 1–8 (2015)
15. Cong, Y., Yuan, J. & Liu, J.: Sparse reconstruction cost for abnormal event detection. in *CVPR 2011* 3449–3456 (2011). https://doi.org/10.1109/CVPR.2011.5995434
16. Cong, Y., Yuan, J., Liu, J.: Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognit.* **46**, 1851–1864 (2013)
17. Creusot, C. & Munawar, A.: Real-time small obstacle detection on highways using compressive RBM road reconstruction. in *Intelligent Vehicles Symposium (IV), 2015 IEEE* 162–167 (2015)
18. CUHK Avenue, http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
19. Data generated by surveillance cameras: http://www.securityinfowatch.com/news/12160483/data-generated-by-new-surveillance-cameras-to-increase-exponentially-in-the-coming-years
20. de Leo, C., Manjunath, B.S.: Multicamera video summarization and anomaly detection from activity motifs. ACM Trans. Sen. Netw. **10**(27), 1–27:30 (2014)

21. Del Giorno, A., Bagnell, J. A. & Hebert, M. A Discriminative Framework for Anomaly Detection in Large Videos. in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V* (eds. Leibe, B., Matas, J., Sebe, N. & Welling, M.) 334–349 (Springer International Publishing, 2016). https://doi.org/10.1007/978-3-319-46454-1_21
22. Distante, A., Marino, F., Mazzeo, P. L.: Nitti, M. & Stella, E. Automatic Method and System for Visual Inspection of Railway Infrastructure. (2009)
23. Fan, Y., Levine, M.D., Wen, G., Qiu, S.: A deep neural network for real-time detection of falling humans in naturally occurring scenes. *Neurocomputing*. **260**, 43–58 (2017)
24. Fang, Z., et al.: Abnormal event detection in crowded scenes based on deep learning. *Multimed. Tools Appl.* **75**, 14617–14639 (2016)
25. Feng, Y., Yuan, Y. & Lu, X.: Deep representation for abnormal event detection in crowded scenes. in *Proceedings of the 2016 ACM on Multimedia Conference* 591–595 (ACM, 2016).
26. Feng, Y., Yuan, Y., Lu, X.: Learning deep event models for crowd anomaly detection. *Neurocomputing*. **219**, 548–556 (2017)
27. Fujitsu's Intelligent transportation system: http://www.fujitsu.com/cn/en/about/resources/news/press-releases/2016/frdc-0401.html
28. Gan, C., Wang, N., Yang, Y., Yeung, D.-Y., Hauptmann, A.: G. DevNet: A Deep Event Network for multimedia event detection and evidence recounting. in *2015 I.E. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2568–2577 (2015). https://doi.org/10.1109/CVPR.2015.7298872
29. Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T.: Video captioning with attention-based LSTM and semantic consistency. IEEE Trans. Multimed. **19**, 2045–2055 (2017)
30. Girshick, R: Fast r-cnn. IEEE international conference on computer vision, 1440–1448 (2015)
31. Goodfellow, I. et al. Generative Adversarial Nets. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 2672–2680 (Curran Associates, Inc., 2014)
32. Guo, Y., Zhang, J., Gao, L.: Exploiting long-term temporal dynamics for video captioning. World Wide Web. (2018)
33. Haritaoglu, I., Harwood, D., Davis, L.S.: W^4: real-time surveillance of people and their activities. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 809–830 (2000)
34. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K. & Davis, L. S.: Learning temporal regularity in video sequences. in *Computer Vision and Pattern Recognition (CVPR), 2016 I.E. Conference on* 733–742 (2016)
35. He, C., Shao, J., Sun, J.: An anomaly-introduced learning method for abnormal event detection. Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-5255-z
36. Hendel, A., Weinshall, D., Peleg, S.: Identifying Surprising Events in Videos Using Bayesian Topic Models. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) *Computer Vision – ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8–12, 2010, Revised Selected Papers, Part III*, pp. 448–459. Springer, Berlin Heidelberg (2011). https://doi.org/10.1007/978-3-642-19318-7_35
37. Hinami, R., Mei, T., Satoh, S.: Joint detection and recounting of abnormal events by learning deep generic knowledge. The IEEE International Conference on Computer Vision (ICCV). **2017**, (2017)
38. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997)
39. Hou, J., Wu, X., Yu, F. & Jia, Y.: Multimedia event detection via deep spatial-temporal neural networks. in *2016 I.E. International Conference on Multimedia and Expo (ICME)* 1–6 (2016). https://doi.org/10.1109/ICME.2016.7552981
40. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man, Cybern. Part C Applications Rev.* **34**, 334–352 (2004)
41. Hu, X., Hu, S., Huang, Y., Zhang, H., Wu, H.: Video anomaly detection using deep incremental slow feature analysis network. *IET Comput. Vis.* **10**, 258–267 (2016)
42. i-Lids bag and vehicle detection challenge: http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html
43. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A.: Image-to-image translation with conditional adversarial networks. *arXiv Prepr.* (2017)
44. Junior, J.C.S.J., Musse, S.R., Jung, C.R.: Crowd Analysis Using Computer Vision Techniques. *IEEE Signal Process. Mag.* **27**, 66–77 (2010)
45. Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L.J., Strintzis, M.G.: Swarm intelligence for detecting interesting events in crowded environments. *IEEE Trans. image Process.* **24**, 2153–2166 (2015)
46. Kok, V.J., Lim, M.K., Chan, C.S.: Crowd behavior analysis: A review where physics meets biology. *Neurocomputing*. **177**, 342–362 (2016)
47. Krizhevsky, A., Sutskever, I. & Hinton, G. E.: Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems* 1097–1105 (2012)

48. Leach, M.J.V., Sparks, E.P., Robertson, N.M.: Contextual anomaly detection in crowded surveillance scenes. Pattern Recogn. Lett. **44**, 71–79 (2014)
49. Leyva, R., Sanchez, V., Li, C.-T.: Video anomaly detection with compact feature sets for online performance. *IEEE Trans. Image Process.* **26**, 3463–3478 (2017)
50. Leyva, R., Sanchez, V., Li, C.-T.: The LV dataset: a realistic surveillance video dataset for abnormal event detection. In: *Biometrics and Forensics (IWBF), 2017 5th International Workshop on* 1–6 (2017)
51. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 18–32 (2014)
52. Li, T., et al.: Crowded scene analysis: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **25**, 367–386 (2015)
53. Li, N., Wu, X., Xu, D., Guo, H., Feng, W.: Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. *Neurocomputing*. **155**, 309–319 (2015)
54. Li, X., Zhou, Z., Chen, L., Gao, L.: Residual attention-based LSTM for video captioning. World Wide Web. (2018). https://doi.org/10.1007/s11280-018-0531-z
55. Lu, C., Shi, J. & Jia, J.: Abnormal event detection at 150 fps in matlab. in *Computer Vision (ICCV), 2013 I.E. International Conference on* 2720–2727 (2013)
56. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. in *The IEEE International Conference on Computer Vision (ICCV)*. (2017)
57. Mahadevan, V., Li, W., Bhalodia, V. & Vasconcelos, N.: Anomaly detection in crowded scenes. in *Computer Vision and Pattern Recognition (CVPR), 2010 I.E. Conference on* 1975–1981 (2010)
58. Makhzani, A., Frey, B.J.: A winner-take-all method for training sparse convolutional autoencoders. Adv. Neural Inf. Proces. Syst. 2791–2799 (2015)
59. Marsden, M., McGuinness, K., Little, S. & O'Connor, N. E.: Holistic features for real-time crowd behaviour anomaly detection. in *Image Processing (ICIP), 2016 I.E. International Conference on* 918–922 (2016)
60. Medel, J. R. & Savakis, A. E.: Anomaly detection in video using predictive convolutional long short-term memory networks. *CoRR* abs/1612.0, (2016)
61. Menotti, D., et al.: Deep representations for Iris, face, and fingerprint spoofing detection. IEEE Trans. Inf. Forensics Secur. **10**, 864–879 (2015)
62. MIT Traffic dataset, http://www.ee.cuhk.edu.hk/~xgwang/MITtraffic.html (2018)
63. Mo, X., Monga, V., Bala, R., Fan, Z.: Adaptive sparse representations for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.* **24**, 631–645 (2014)
64. Munawar, A., Vinayavekhin, P. & De Magistris, G.: Spatio-temporal anomaly detection for industrial robots through prediction in unsupervised feature space. in *Applications of Computer Vision (WACV), 2017 I.E. Winter Conference on* 1017–1025 (2017)
65. Nakahata, M.T., Thomaz, L.A., da Silva, A.F., da Silva, E.A.B., Netto, S.L.: Anomaly detection with a moving camera using spatio-temporal codebooks. Multidimens. Syst. Signal Process. **29**, 1025–1054 (2018)
66. Narasimhan, M. G. & Sowmya Kamath S.: Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimed. Tools Appl.* (2017). https://doi.org/10.1007/s11042-017-4940-2
67. Ogawa, T., Hiraoka, D., Ito, S., Ito, M. & Fukumi, M.: Improvement in detection of abandoned object by pan-tilt camera. in *Knowledge and Smart Technology (KST), 2016 8th International Conference on* 152–157 (2016)
68. Olson, C. C. & Doster, T.: A Novel Detection Paradigm and Its Comparison to Statistical and Kernel-Based Anomaly Detection Algorithms for Hyperspectral Imagery. in *2017 I.E. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 302–308 (2017). https://doi.org/10.1109/CVPRW.2017.43
69. Pathak, A. R., Pandey, M., Rautaray, S. & Pawar, K.: Assessment of object detection using deep convolutional neural networks. in *Intelligent Computing and Information and Communication* (eds. Bhalla, S., Bhateja, V., Chandavale, A. A., Hiwale, A. S. & Satapathy, S. C.) 457–466 (Springer Singapore, 2018)
70. Pathak, A. R., Pandey, M. & Rautaray, S.: Deep learning approaches for detecting objects from images: a review. in Progress in Computing, Analytics and Networking (eds. Pattnaik, P. K., Rautaray, S. S., Das, H. & Nayak, J.) 491–499 (Springer Singapore, 2018)
71. PETS dataset: http://www.cvg.reading.ac.uk/PETS2009/a.html
72. Popoola, O.P., Wang, K.: Video-based abnormal human behavior recognition—A review. *IEEE Trans. Syst. Man, Cybern. Part C Applications Rev.* **42**, 865–878 (2012)
73. QMUL: QMUL junction dataset: http://www.eecs.qmul.ac.uk/~sgg/QMUL_Junction_Datasets/Junction/Junction.html

74. Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E. & Sebe, N.: Plug-and-Play CNN for Crowd Motion Analysis: An Application in Abnormal Event Detection. *CoRR* abs/1610.0**,** (2016)
75. Ravanbakhsh, M., Sangineto, E., Nabi, M. & Sebe, N.: training adversarial discriminators for cross-channel abnormal event detection in crowds. *CoRR* abs/1706.0, (2017)
76. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 779–788 (2016)
77. Reed, I.S., Yu, X.: Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. IEEE Trans. Acoust. **38**, 1760–1770 (1990)
78. Roshtkhari, M.J., Levine, M.D.: An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Comput. Vis. Image Underst.* **117**, 1436–1452 (2013)
79. Sabokrou, M., Fathy, M., Hoseini, M. & Klette, R.: Real-time anomaly detection and localization in crowded scenes. in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* 56–62 (2015)
80. Sabokrou, M., Fathy, M., Hoseini, M.: Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electron. Lett.* **52**, 1122–1124 (2016)
81. Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R.: Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. Image Process.* **26**, 1992–2004 (2017)
82. Sabokrou, M., Fathy, M., Moayed, Z., Klette, R.: Fast and accurate detection and localization of abnormal behavior in crowded scenes. Mach. Vis. Appl. **28**, 965–985 (2017)
83. Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R.: Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Underst.* (2018). https://doi.org/10.1016/j.cviu.2018.02.006
84. Saligrama, V. & Chen, Z.: Video anomaly detection based on local statistical aggregates. in *Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on* 2112–2119 (2012)
85. Saligrama, V., Konrad, J., Jodoin, P.-M.: Video anomaly identification. *IEEE Signal Process. Mag.* **27**, 18–33 (2010)
86. Shah, M., Javed, O., Shafique, K.: Automated visual surveillance in realistic scenarios. IEEE Multimed. **14**, 30–39 (2007)
87. Shao, M., Fu, Y.: Deeply Self-Taught Multi-View Video Analytics Machine for Situation Awareness. in *AFA Cyber Workshop. White Paper*. (2015)
88. Simonyan, K. & Zisserman, A. : Two-stream convolutional networks for action recognition in videos. in *Advances in neural information processing systems* 568–576 (2014)
89. Sjarif, N.N.A., Shamsuddin, S.M., Hashim, S.Z.: Detection of abnormal behaviors in crowd scene: a review. *Int. J. Adv. Soft Comput. Appl.* **4**, 1–33 (2012)
90. Sodemann, A.A., Ross, M.P., Borghetti, B.J.: A review of anomaly detection in automated surveillance. *IEEE Trans. Syst. Man, Cybern. Part C Applications Rev.* **42**, 1257–1272 (2012)
91. Song, J., et al.: Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Trans. Image Process. **27**, 3210–3221 (2018)
92. Sun, L., Ai, H., Lao, S.: Localizing activity groups in videos. *Comput. Vis. Image Underst.* **144**, 144–154 (2016)
93. Sun, J., Shao, J., He, C.: Abnormal event detection for video surveillance using deep one-class learning. Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-5244-2
94. Thida, M., Yong, Y.L., Climent-Pérez, P., Eng, H., Remagnino, P.: A Literature Review on Video Analytics of Crowded Scenes. In: Atrey, P.K., Kankanhalli, M.S., Cavallaro, A. (eds.) *Intelligent Multimedia Surveillance: Current Trends and Research*, pp. 17–36. Springer, Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-41512-8_2
95. Tian, Y., et al.: IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework. *Mach. Vis. Appl.* **19**, 315–327 (2008)
96. Tran, H. T. M. & Hogg, D. C.: Anomaly detection using a convolutional winner-take-all Autoencoder. in Proceedings of the British Machine Vision Conference 2017 (2017)
97. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: *Computer Vision (ICCV), 2015 I.E. International Conference on* 4489–4497 (2015)
98. TRECVID Multimedia Event Recounting (MER) Evaluation Plan: https://www.nist.gov/sites/default/files/documents/itl/iad/mig/MER_TRECVID_evaluation_spec_v23.pdf
99. Tudor Ionescu, R., Smeureanu, S., Alexe, B., Popescu, M.: Unmasking the abnormal events in video. in *The IEEE International Conference on Computer Vision (ICCV)*. (2017)
100. UCSD dataset: http://www.svcl.ucsd.edu/projects/anomaly/dataset.html
101. UMN dataset: http://mha.cs.umn.edu/proj_events.shtml#crowd

102. Video Surveillance Market: http://www.transparencymarketresearch.com/video-surveillance-vsaas-market.html
103. VIOLENT-FLOWS dataset: http://www.openu.ac.il/home/hassner/data/violentflows/
104. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **29**, 983–1009 (2013)
105. Wang, J., Xu, Z.: Spatio-temporal texture modelling for real-time crowd anomaly detection. *Comput. Vis. Image Underst.* **144**, 177–187 (2016)
106. Wang, B., Ye, M., Li, X., Zhao, F., Ding, J.: Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Mach. Vis. Appl.* **23**, 501–511 (2012)
107. Wang, Q., Wan, J. & Yuan, Y.: Deep metric learning for crowdedness regression. *IEEE Trans. Circuits Syst. Video Technol.* PP, 1 (2017)
108. Wang, X., Gao, L., Wang, P., Sun, X., Liu, X.: Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length. IEEE Trans. Multimed. **20**, 634–644 (2018)
109. Wang, X., et al.: Deep appearance and motion learning for egocentric activity recognition. Neurocomputing. **275**, 438–447 (2018)
110. Weizmann dataset: http://www.wisdom.weizmann.ac.il/~vision/Irregularities.html
111. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 780–785 (1997)
112. Wu, H., Shao, J., Xu, X., Shen, F. & Shen, H. T. A system for spatiotemporal anomaly localization in surveillance videos. in *Proceedings of the 2017 ACM on Multimedia Conference* 1225–1226 (ACM, 2017). https://doi.org/10.1145/3123266.3127912
113. Xie, S., Guan, Y.: Motion instability based unsupervised online abnormal behaviors detection. *Multimed. Tools Appl.* **75**, 7423–7444 (2016)
114. Xu, D., Ricci, E., Yan, Y., Song, J. & Sebe, N. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. *CoRR* abs/1510.0, (2015)
115. Xu, D., Yan, Y., Ricci, E., Sebe, N.: Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **156**, 117–127 (2017)
116. Yoffie, David B.: "Mobileye: The Future of Driverless Cars." Harvard Business School Case 715–421, October 2014. (Revised October 2015)
117. Yogameena, B., Nagananthini, C.: Computer vision based crowd disaster avoidance system: A survey. *Int. J. Disaster Risk Reduct.* **22**, 95–129 (2017)
118. Yu, R., Qiu, H., Wen, Z., Lin, C., Liu, Y.: A survey on social media anomaly detection. SIGKDD Explor. Newsl. **18**, 1–14 (2016)
119. Zeiler, M. D. & Fergus, R.: Visualizing and understanding convolutional networks. in *European conference on computer vision* 818–833 (2014)
120. Zhao, B., Fei-Fei, L. & Xing, E. P.: Online detection of unusual events in videos via dynamic sparse coding. in *CVPR* 2011 3313–3320 (2011). https://doi.org/10.1109/CVPR.2011.5995524
121. Zhao, Y. et al. Spatio-temporal AutoEncoder for video anomaly detection. in *Proceedings of the 2017 ACM on Multimedia Conference* 1933–1941 (ACM, 2017). https://doi.org/10.1145/3123266.3123451
122. Zhou, S., et al.: Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process. Image Commun. **47**, 358–368 (2016)
123. Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K.: Context-aware activity recognition and anomaly detection in video. *IEEE J. Sel. Top. Signal Process.* **7**, 91–101 (2013)
124. Zitouni, M.S., Bhaskar, H., Dias, J., Al-Mualla, M.E.: Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. *Neurocomputing*. **186**, 139–159 (2016)