Materials Today: Proceedings 58 (2022) 162–167

Contents lists available at ScienceDirect

Materials Today: Proceedings

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

Survey on anomaly detection in surveillance videos

S. Anoopa *, A. Salim

Department of CSE, College of Engineering Trivandrum, India

ARTICLE INFO

Article history:
Available online 21 January 2022

Keywords:
Anomaly detection
Video surveillance
CCTV cameras
Deep learning
Representation learning

ABSTRACT

The demand for video surveillance has seen a staggering growth in the last few years due to the rapid growth of urbanization & industrialization. Video Surveillance is a system which observe scenes, activities, behaviour or other kind of information with the help of CCTV Cameras & associated software. The main objectives of video surveillance are detection and tracking of moving objects, loss prevention, traffic monitoring, and capturing variety of real world anomalies. The major challenges in Anomaly detection are extraction of appropriate features, addressing normal behaviour, environment divergence and occurrence of abnormal events in sparse manner. Many algorithms have been proposed in this area for object detection and tracking related applications, however, the problem is still challenging because of dynamic behaviour in anomalies. In this review paper a comparative study of different anomaly detection methods in deep learning and representation learning is performed and the limitations of each method are listed. Also, we outline open research challenges for proceeding further research. This paper will use as a quick reference to the researchers working in the same field.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Artificial Intelligence & Energy Systems.

1. Introduction

During the last decade, the requirement of video surveillance has received worldwide acceptance and unparalleled growth. Major factors like new government policies, positive economic growth, increase in employment rate, R & D, which has catalyst the rapid growth of urbanization & modern industries. Being a cost effective & efficient method, different agencies are deploying video surveillance system to maintain the law & order and ensure the safety of citizen & valuable assets with the help of CCTV Cameras and associated software [1]. However, it is impractical & unrealistic to observe and analyze every anomaly in the video with the help of human observer. Hence, an Intelligent video surveillance system has been deployed with different techniques for detection and localization of anomalies, which are very closely related. Detection is the process of identifying outlier within a frame whereas Localization focus on identifying the location of anomaly by using a bounding box. Considering the importance and advantages, a large number of works are associated with this field, however, the area is quite challenging.

In this paper we discuss about eight different methods, of which four methods falls under deep learning and others in representation learning. Deep Learning models are learned from experiences by simply training the artificial neural networks. During learning process features are extracted automatically from input videos which are efficient, robust and compact. The deep learning models discussed here are Incremental spatio-temporal learner approach [4], Unsupervised Spectral Mapping [21], Multilayer Perceptron RNN [24] and GE based Promotion method [25]. Whereas in Representation learning models, the learning is performed by extracting useful information from the input video by considering prior information of the problem domain. With this method, the computational complexity is reduced to a certain extent by applying dimensionality reduction. The important methods based on the representation learning discussed are Optical flow based Convolutional Autoencoder [8], detection using low dimension descriptor [18], local motion based joint video representation [20] and low rank dictionary learning [22].

1.1. Contribution

(a) Different approaches in video anomaly detection algorithms are analysed and a comparative study is performed.
(b) Open research challenges are listed.

* Corresponding author.
E-mail address: mailanoopas@gmail.com (S. Anoopa).

https://doi.org/10.1016/j.matpr.2022.01.171
2214-7853/Copyright © 2022 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the International Conference on Artificial Intelligence & Energy Systems.

2. Video anomaly types and approaches

Video anomaly detection means to find the abnormal behaviour in videos. Anomalies are of four types (a) local anomaly, (b) global anomaly, (c) interaction anomaly and (d) point anomaly. Local anomalies mean behaviour of an object which differ from neighbourhood objects [2]. Global anomalies are group of abnormal behaviour also termed as crowd anomalies [2,3]. A difference in single data point from other dataset is a point anomaly and the unusual interaction between the human individuals is termed as interaction anomalies [3].

2.1. Incremental spatio-temporal learner approach (ISTL)

ISTL is an unsupervised deep learning method using active learning with fuzzy aggregation [4]. The three phases of ISTL includes

(a) Spatio temporal learning
(b) Anomaly detection and localization
(c) Active Learning with Fuzzy aggregation.

The term spatio temporal learning refers to the continuous learning process with the help of both locations based and time-interval information. Spatio temporal autoencoder is an artificial neural network which is developed for learning both motion and appearance from the input image. The reason for using the autoencoder is to reduce the reconstruction cost by using backpropagation algorithm [5]. The architecture of autoencoder contains different layers. The pre-processing step is performed in input layer to increase the learning of autoencoder [4], Optical low based Convolutional Autoencoder. Dimension of each frame is reduced by converting frames in to grayscale. Convolutional layers are responsible for preserving spatial relationship between frames. The ISTL layers model contain two convolutional layer & two deconvolutional layers and each layer have set of filters for feature extraction. Long Short Term Memory (LSTM) networks are Recurrent Neural Network capable of learning long term dependences and reduce the time lag between the processing of frames. The LSTM system contains cell, input gate, output gate, forget gate which helps to identify the relevant information. Convolutional LSTM layers [6] are used for capturing temporal behaviour of input data. The architecture of the spatio-temporal encoder is illustrated in Fig. 1.

The trained autoencoder does not have the ability to reconstruct the abnormal scenes from the input video. Therefore, Reconstruction error [5] represents the score value of abnormal temporal cuboid. The reconstruction error is computed using the equation

$$E = \sum_{i=0}^{T}\sum_{j=0}^{W}\sum_{k=0}^{h}\sqrt[2]{\varphi(i,j,k)}$$

$$\varphi(i,j,k) = |X_{(i,j,k)} - \bar{X}_{(i,j,k)}|^{2} \qquad (2)$$

where X and $\bar{X}$ are temporal cuboid of input and reconstructed image. If reconstruction error is above the anomaly threshold, the resulting frame is set as an anomaly. One of the most important advantages of this method is active learning with fuzzy aggregation [7]. The significance of active learning is to improve the accuracy of anomaly detection. Active learning is improved by human observer and fuzzy aggregation [7]. Based on the feedback of human observer, learning is performed continuously and hence improve the accuracy.

2.2. Detection using optical low and convolutional autoencoder

Optical flow is used to obtain the apparent motion of an objects in video frames. It helps to obtain depth information relative to direction and speed of objects. This method is applied along with convolutional autoencoder [8] for improving accuracy and reducing computational complexity. The entire framework consists of three different stages; first stage is pre-processing which helps to extract dense optical flow of frames [9]. Second stage is called Convolutional Autoencoder helps to obtain spatial structure of a frames and third stage is Convolutional Long Short Term Memory to learn the temporal patterns of frames.

The steps included in the pre-processing are rescaling and finding optical flow of video frames. Farneback optical flow [10] method is used to obtain both velocity and direction of foreground objects. After the pre-processing video frames are given to the deep learning model f to maintain the spatial features of video frames using back propagation algorithm. The model contains two parts spatial encoder and spatial decoder having LSTM Layers. The reason for using LSTM layers are it has memory cell [11] which hold the information for a longer time. The network is trained by normal behaviour and square filters are used to maintain spatial relationship between video frames. Finally, Regularity Score is computed after training using multi scale structural similarity method [12]. The regularity score $r(x)$ is computed as follows.

$$r(x) = 1 - \frac{a(x) - min_t a(t)}{max_t a(t) - min_t a(t)} \qquad (3)$$

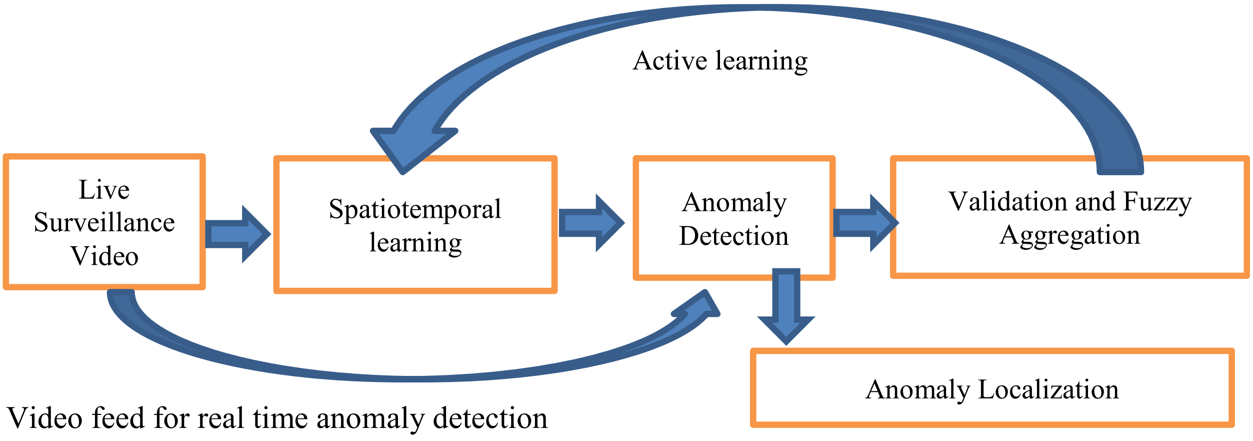

Fig.1. Architecture of Incremental Spatio Temporal Learning (ISTL) [4].

$$a(x) = 1 - (MS\_SSIM(x, f_w(x)) \tag{4}$$

where $a(x)$ is the pixel wise error and MS_SSIM is the multi-scale structural similarity between frames. To avoid local minima problem, Persistence ID [13] algorithm is used.

2.3. Crowd anomaly detection using low dimensional descriptor.

Low dimensional descriptor is proposed for detecting anomalies in crowded scenes. The three different features extracted from the optical flow vector of video frames is used to develop a descriptor [14]. The three features are listed below

(a) Feature 1: It is used to eliminate perturbations of the background area in video frames. Feature 1 described here is sum of the threshold of optical flow magnitude of a video frame.

$$\overline{OF}(i,j,k) = \begin{cases} OF(i,j,k), ifOF(i,j,k) > T \\ 0, ifOF(i,j,t) \le T \end{cases} \tag{5}$$

where OF represents optical flow between objects. The motion of object in which OF value less than threshold (T) is neglected for further calculations.

(b) Feature 2: The Joint Entropy method is applied to find the dissimilarities between two consecutive video frames.
(c) Feature 3: This feature is used for computing variance in space and time. If a video frame has high variance the degree of crowd desperation is very high. Third feature is computed as

$$F3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{6}$$

After analysing three feature a post processing is also needed to obtain the smooth profile, reducing the effect of noise and outlier. For the classification unsupervised Support Vector Machine [15] is trained with only normal behaviour of crowd and it maps the normal behaviour in to hypersphere.

2.4. Anomaly detection and localization by local motion based joint video representation

This method helped to find anomalies in both crowded and uncrowded scenes. Here features are extracted by considering two motion based video descriptor [16]. Spatially Localized Histogram of Optical Flow (SL-HOF) and Uniform Local Gradient Pattern based Optical Flow (ULGP-OF) methods are applied to capture spatial and optical flow information respectively. PCA (Principal component analysis) is applied for localization of foreground region. First the training videos are used for extracting both spatial region and local texture with the help of SL-HOF and ULGP-OF descriptors. Both SL-HOF and ULGP-OF algorithms are modelled by One Class Extreme Learning Machine (OCELM). To obtain SL-HOF based video representation steps are

(a) Video frames split in to non-overlapping patches.
(b) Patches having similar spatial location are combined in to spatio- temporal cuboid.
(c) The spatio- temporal cuboid is divided in to different non overlapping $m \times n$ 3D local regions
(d) Sub histogram is extracted from $m \times n$ 3D local region.
(e) All the histogram is combined to obtain SL-HOF Feature.

The flow diagram of the work is described in Fig. 2.

2.4.1. Motion of local texture

With the help of the PCA algorithm & Sigmoid transformation, the foreground extractor & mapping operation are performed respectively. Foreground objects localization are carried out with Binarization method and enhancement of bounding boxes appearance through ULGP-OF descriptor. A Low Level Descriptor (LGP) code is used for image texture representation [16]. All the said features are modelled by OCELM (one class data description algorithm) by applying sparse density learning. Furthermore, Gaussian Kernel based OCELM [17] is implemented for increasing both accuracy and learning speed.

2.5. Anomaly detection based on low rank dictionary learning

Dictionary learning is the process of finding frames having sparsity. Low rank dictionary learning method [18] is applied for detecting abnormal events in crowded scenes. Features are extracted by using background subtraction and Binarization methods. The reason for using this method is to reduce the noise and variations of objects in background. During the training stage a low rank dictionary based similarity matrix is obtained and develop a new optimization method [18]. Low-Rank and Compact Coefficient Dictionary Learning (LRCCDL) is an iterative process to obtain low rank dictionary and mean vector. One of the important feature of this method is motion feature extraction. Motion information of two frames are obtained by optical flow field [19] by considering both amplitude and directions. The process of generation of histogram of maximal optical flow projection (HMOFP) [20] is shown in Fig. 3.

In order to detect the anomalous frames, the reconstruction error [18] is computed and the compared with the threshold value. If the error is greater than the threshold value, the output frame is considered as abnormal.

2.6. Unsupervised spectral mapping and hyperspectral anomaly detection

Spectral mapping is the process by which map the data from high dimension to low dimensional by preserving noise and interference between the frames. While applying dimensionality reduction original vectors are represented in unsupervised manner, hence it is called unsupervised spectral mapping [21]. During spectral mapping the spectral consistency and Gaussianity are added to find out the nonlinear mapping relation between high-dimensional and low-dimensional feature space.

2.6.1. Spectral mapping

Let X represent the input HSI with $N \times N$ spectral vectors and contain C dimensions [21] which represents spectral maps in the deep latent space. Spectral mapping encoding (he) and decoding (hd) is defined as

$$he = f(w_i x_i + b_1) \tag{7}$$

$$hd = f(w_i m_i + b_1) \tag{8}$$

The Stochastic Gradient Decent SGD optimization algorithm [22] is used for optimizing the parameters of the spectral mapping autoencoder.

2.6.2. Feature selection

To represent the original vectors completely, the noise and interference in the deep feature space to be preserved. So that it is very easy to select a feature map with maximum local density.

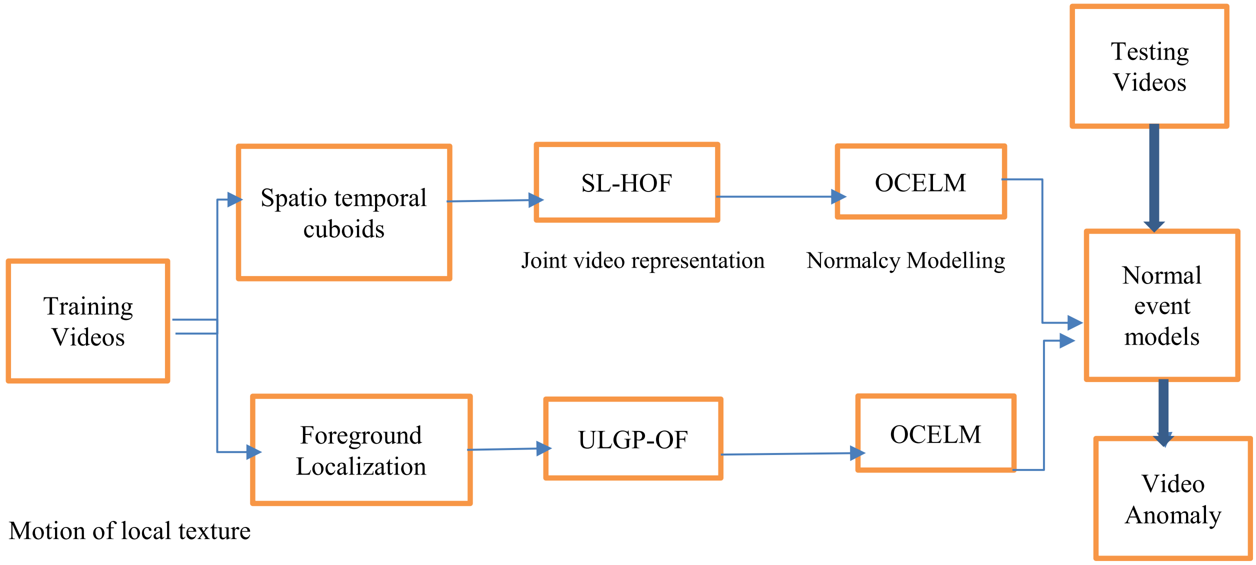

Fig.2. Anomaly Detection using SL-HOF and OCELM [16].

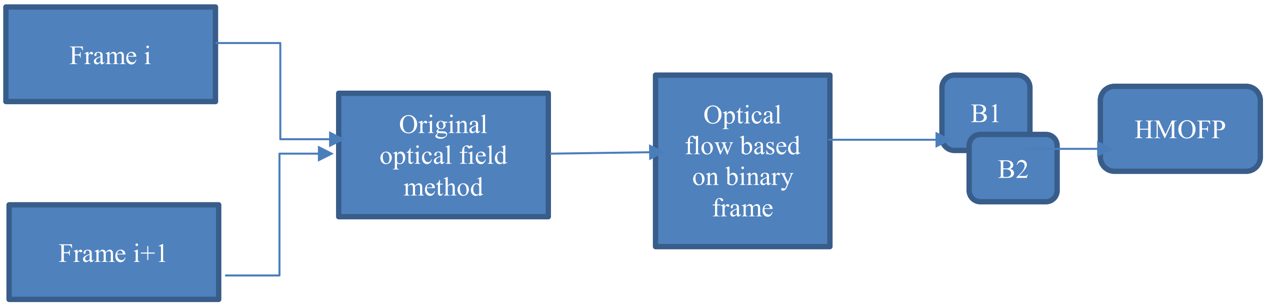
Frame i
Original optical field method
Optical flow based on binary frame
B1
B2
HMOFP
Frame i+1

Fig. 3. Generation of HMOFP [18].

To minimize the loss of information, similar points are considered [23]. Selected maps are calculated using

$$B = argmax(\varepsilon_i \times S_i) \tag{9}$$

where ε_i is the density and S_i is the distance between the maps.

2.6.3. Anomaly detection

Comprehensive model is selected for anomaly detection to reduce the rate of miss classification.

$$Q = \alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_3 \tag{10}$$

where P_1, P_2, P_3 are called attributes maps of anomalies and $\alpha_1, \alpha_2, \alpha_3$ are the parameters of selected features. Finally, Guided filter helps to eliminate cluster anomalies and also reduces false alarm rate.

2.7. Anomaly detection based on multilayer perceptron RNN

Multilayer perceptron is a feed forward neural network which is composed of large number of fully connected layer. It helps to overcome the utilization of high computation power by performing adaptive learning. An optimized multilayer neural network is used for detecting anomalies in real world [24]. The input video is given to frame extraction stage for extracting the frames contain anomalous information. Then the Gaussian mixture model is used for background estimation. After object detection, ROI (Region of Interest Selection) method is implemented for extracting foreground information from back ground pixels. Maximally stable external regions feature extraction strategies used for extracting features from the ROI of a frame. Finally, object detection and tracking step is implemented with the help of RNN (Recurrent Neural Network). The reason for using RNN is faster training as compared to back propagation networks. The architecture [24] of the network is described in Fig. 4.

The basis configuration of RNN has 12 input layers, 24 hidden layers and one output layer. The output of this method provides high quality, effective segmentation and less complexity.

2.8. Video anomaly detection using GE based promotion method

Generation error is a measure of how accurately an algorithm predicts the output from the input data. A new method called promotion method based on generation error [25] is implemented for detecting anomalies in videos. The entire process of GE based deep learning methods consist of two steps: training and testing. The promotion method is implemented during the testing stage of detection. One of the main reason for implementing GE- based method is to set a single threshold value to all the video frames subjected to detect anomalies and also reduce the high false alarm rate. This is done by considering maximum value of GE for detecting anomalies in frames using max pooling and mean filter operation. During training, a generator is used to produce normal patterns. To learn information of the whole scene, frame level GE is considered. After the training process a bidirectional LSTM [26] called GNN is used for generating GE Maps. Then the block level process is performed with help of max pooling and mean filtering.

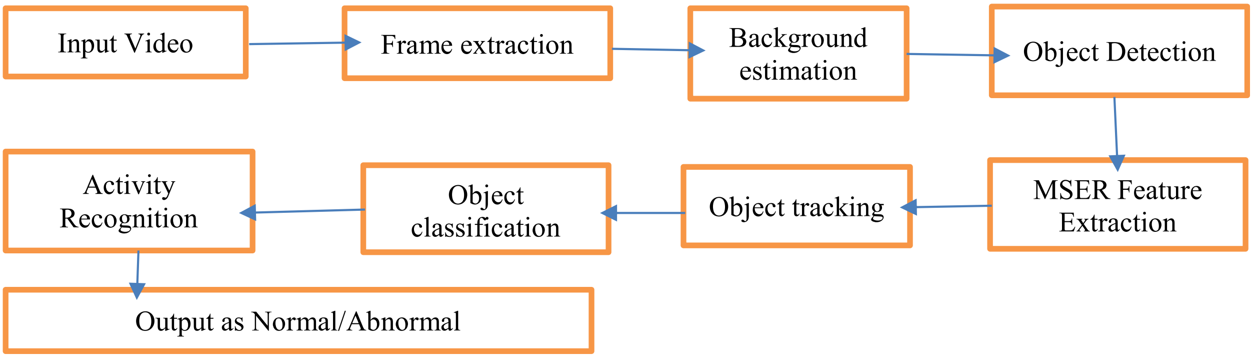

Fig.4. Anomaly detection using RNN [24].

Table 1
Comparative analysis of algorithms.

METHOD	Objectives	Limitations
Incremental spatio-temporal learner approach	System to detect anomaly from real time videos using deep learning methods	1. Spatial and temporal streams are processed together causes false negative detection.2. Human feedback is needed for continuous learning.
Optical low and convolutional autoencoder	System to detect anomaly in videos using autoencoder and optical flow model.	1. Local Minima problem.2. Fails in pixel level anomaly detection.3. Doesnot get robust result.
Low Dimensional descriptors	System to detect crowd anomalies.	1. Accuracy depends on the quality of data.2. Comparatively low recall and high Localization error.
Local motion based joint video representation	System to detect anomalies in both crowded and uncrowded region	1. Cannot be model with high sparsity.2. Computation overhead is high.
low rank dictionary learning	Develop a novel approach for abnormal event detection in crowded scenes.	1. Running time is very high.2. Cannot be used for high resolution videos.
Unsupervised spectral mapping	An efficient method to detect point and contextual anomalies.Based on spectral mapping and feature selection	1. System cannot produce better result if the resolution of hyperspectral images is high.
Multilayer perceptron RNN	Optimized multilayer perceptual regression neural network based on real world anomalies.	1. Feature extraction is not accurate.2. Result depends on ROI Selection.
GE based promotion method	Introduced a GE based promotion method for detecting anomalies in videos.	1. Model cannot be used for all types of data set.2. U-Net is not the best choice for extracting temporal features

In block level process GE saliency of normal and abnormal frames are compared. The saliency is computed using [25]

$$saliency = \frac{L_{abnormal} - L_{normal}}{L_{normal}} \tag{11}$$

The maximum value of block level saliency is used for detecting anomalies in frames.

$$L_B(t) = max\{L_{B1}(t), L_{B2}(t) \cdots\cdots..L_{Bk}(t)\} \tag{12}$$

where k is the number of blocks in a frames and $L_{Bk}(t)$ is the k^{th} block level GE.

$$L_{Bk}(t) = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} B_{i,j,k} \tag{13}$$

w and h are the height and width of the block. Finally, an anomaly score is calculated to find the frames that contain anomalous behaviour.

3. Comparative study

While reviewing the above eight methods, it has been analyzed that selection of a best method for a specific anomaly detection problem is a very crucial aspect. A suitable method is selected based on the different factors like dataset, performance, and application domain. In this review we have discussed about two learning methods, in which deep learning models produce better results than representation learning models. In the case of representation learning, accuracy depends on the selection of a good representation of data and prior information about the problem domain, however, deep learning networks have the ability to learn efficient and compact representations of data by combing simple and complex structures. Hence, deep learning methods require a large amount of data to give better performance and are very expensive for training complex data. The said limitations of deep learning can be overcome by adding active learning to make more accurate decisions. This approach helps to develop a high-performance system from a training dataset by actively selecting valuable data points. As a result, this method suits well to solve real-time problems with high accuracy. Table 1 shows the main objectives and limitations of algorithms proposed in anomaly detection in video surveillance applications.

4. Research challenges

Based on the review of papers discussed above, few important challenges are listed as follows

1. *Unavailability of better dataset*: - Publically available benchmarked dataset for anomaly detection are very less due to the imbalanced distribution of normal and abnormal behaviour of data.
2. *Environmental challenges*: - Complex environmental challenges like background clutter, occlusions and illuminations which affect the accurate detection of anomaly.
3. *Time and Space complexity:* - Most of the existing algorithms have high computational space and time complexity. Hence developing a simple, efficient and accurate system is still a challenging problem.

4. *Dynamic behaviour in anomalies:* -Anomalous events are rarely occurred and its behaviour is always in sparse manner, so a single technique is not capable of detecting all the types of anomalies.
5. *Atmospheric Turbulences*: - Common atmospheric challenges are variations due to refraction and reflection of light, smoke and fog which blur the images in the video.

5. Conclusion

Anomaly detection in video surveillance has received global acceptance and used in different application domain such as traffic monitoring, military surveillance, agricultural field, private and public places. In this survey, it is found that the deep learning techniques produces excellent result than the representation learning by adding active leaning concept. Also, open research challenges are briefly listed out here, which helps further research activities.

CRediT authorship contribution statement

S. Anoopa: Conceptualization, Writing – original draft. **A. Salim:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Liu, S. Chen, N. Kubota, Intelligent video systems and analytics: a survey, IEEE Trans. Ind. Inform. 9 (3) (2013) 1222–1233.
[2] X. Hu, S. Hu, Y. Huang, H. Zhang, H. Wu, Video anomaly detection using deep incremental slow feature analysis network, IET Comput. Vis. 10 (4) (2016) 258–267.
[3] R. Chalapathy, S. Chawla, Deep Learning for Anomaly Detection: A Survey, arXiv preprint arXiv:1901.03407.
[4] R. Nawaratne et al., Spatiotemporal anomaly detection using deep learning for real-time video surveillance, IEEE Trans. Ind. Inform. 16 (1) (2019) 393–402.
[5] P. Baldi, Autoencoders, unsupervised learning and deep architectures, in: Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop, Washington, USA, vol. 27, 2011, pp. 37–50.
[6] S.H.I. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810.
[7] D. De Silva, D. Alahakoon, Incremental knowledge acquisition and self-learning from text, in: The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8.
[8] E. Duman, O.A. Erdem, Anomaly detection in videos using optical flow and convolutional autoencoder, IEEE Access 7 (2019) 183914–183923.
[9] C. Dhiman, D.K. Vishwakarma, A review of state-of-the-art techniques for abnormal human activity recognition, Eng. Appl. Artif. Intell. 77 (2018) 21–45.
[10] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Proc. Scand. Conf. Image Anal. Springer, Berlin, Germany, 2003, pp. 363–370.
[11] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 802–810.
[12] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, Nov. 2003, pp. 1398–1402.
[13] Y. Kozlov, T. Weinkauf, Persistence1D: Extracting and Filtering Minima and Maxima of 1D Functions, pp. 1–11. <http://people.mpi-inf.mpg.de/weinkauf/notes/persistence1d.html> (Accessed: 2015).
[14] T. Qasim, N. Bhatti, A low dimensional descriptor for detection of anomalies in crowd videos, Math. Comput. Simul. 166 (2019) 245–252.
[15] B. Schölkopf et al., Estimating the support of a high-dimensional distribution, Neural Comput. 13 (7) (2001) 1443–1471.
[16] S. Wang et al., Video anomaly detection and localization by local motion based joint video representation and OCELM, Neurocomputing 277 (2018) 161–175.
[17] Q. Leng et al., One-class classification with extreme learning machine, Mathematical Problems in Engineering 2015, 2015.
[18] A. Li et al., Abnormal event detection in surveillance videos based on low-rank and compact coefficient dictionary learning, Patt. Recogn. 108 (2020) 107355.
[19] A. Li, Z. Miao, Y. Cen, Global abnormal event detection based on compact coefficient low-rank dictionary learning, 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2017.
[20] A. Li et al., Histogram of maximal optical flow projection for abnormal events detection in crowded scenes, Int. J. Distrib. Sens. Networks 11 (11) (2015) 406941.
[21] W. Xie et al., Unsupervised spectral mapping and feature selection for hyperspectral anomaly detection, Neural Netw. 132 (2020) 144–154.
[22] L. Bottou, Stochastic gradient learning in neural networks, Proc. Neuro-Nımes 91 (8) (1991) 12.
[23] S. Jia et al., A novel ranking-based clustering approach for hyperspectral band selection, IEEE Trans. Geosci. Rem. Sens. 54 (1) (2015) 88–102.
[24] M. Murugesan, S. Thilagamani, Efficient anomaly detection in surveillance videos based on multi-layer perception recurrent neural network, Microprocess. Microsyst. 79 (2020) 103303.
[25] Z. Wang, Z. Yang, Y.-J. Zhang, A promotion method for generation error-based video anomaly detection, Pattern Recogn. Lett. 140 (2020) 88–94.
[26] S. Lee, H.G. Kim, Y.M. Ro, STAN: Spatio-temporal adversarial networks for abnormal event detection, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018.