Multimedia Tools and Applications (2020) 79:35275–35295
https://doi.org/10.1007/s11042-019-7702-5

Unsupervised deep learning system for local anomaly event detection in crowded scenes

Anitha Ramchandran [1] · Arun Kumar Sangaiah [1]

Received: 29 January 2019 / Revised: 2 April 2019 / Accepted: 26 April 2019 /
Published online: 12 May 2019

Abstract
Anomaly detection in video surveillance is a significant research subject because of its immense use in real-time applications. These days, open spots like hospitals, traffic areas, airports are monitored by video surveillance cameras. Strange occasions in these recordings have alluded to the anomaly. Unsupervised anomaly detection in the video be endowed with many challenges as there is no exact definition of abnormal events. It varies as for various situations. This paper aims to propose an effective unsupervised deep learning framework for video anomaly detection. Raw image sequences are combined with edge image sequences and given as input to the convolutional auto encoder-ConvLSTM model. Experimental evaluation of the proposed work is performed in three different benchmark datasets such as Avenue, UCSD ped1 and UCSD ped2. The proposed method Hybrid Deep Learning framework for Video Anomaly Detection (HDLVAD) reaches better accuracy compared to existing methods. Investigating video streaming in big data is our further research work.

Keywords Video surveillance · Abnormal event detection · Crowd analysis · Convolutional auto encoder · Convolut LSTM

1 Introduction

Video analysis is an active research area in computer vision domain. The massive amount of video data is available today because of surveillance cameras installed in almost every part of our society. Growth in hardware technologies and processing power, less cost of surveillance cameras and existing of large video data made video analysis as a trending and significant

✉ Arun Kumar Sangaiah
arunkumarsangaiah@gmail.com

Anitha Ramchandran
anitha66r@gmail.com

[1] School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

research area [12]. It has many real-time applications such as human behavior recognition, traffic monitoring, and violence detection. In most cases, video surveillance cameras are installed for security concerns. Analyzing gigantic amount of video data is a tedious task. So, intelligent surveillance is highly needed where human operators can be alerted automatically when there is any abnormality in the recorded video. In general, intelligent video analysis research area has two paths, event recognition and anomaly detection. The first one concentrates on interpreting video semantically (e.g. human activity recognition) whereas the second one concentrates on finding unusual or rare events. Video anomaly detection is a meticulous task for intelligent surveillance cameras. It is really important in real-world applications as it is able to capture rare or abnormal events. Angela A Sodemann et al., [32] provided a detailed review of anomaly detection in the surveillance video.

Video anomaly detection can be done in crowded scenes and uncrowded scenes. Finding an abnormal event in crowded scenes gives us an additional challenge because of occlusions. Video anomaly detection can further be classified into local anomaly detection and global anomaly detection. Global anomaly detection refers to the process of finding global anomaly behavior of people in surveillance videos [8]. For example, if any unusual event occurs such as bomb blast, accident or violence, most of the peoples run in different directions. Local anomaly detection refers to the process of finding local anomaly behaviour of an individual or a few people. For example, a person goes by bicycle in a pathway where all the people go by walk.

In previous works, the models of video anomaly detection are classified as object-centric or trajectory based and holistic methods or non-trajectory based. The models which detect anomaly using techniques like object detection, behavior recognition, and tracking trajectories are called as object centric. In this method, the crowd is considered as a group of individuals. Trajectory-based models [26] segment image frames and tracks each object. But, holistic method based models consider crowd as a whole entity. In this method, normal events or abnormal events are modeled. Spatiotemporal features are extracted for detecting unusual events in videos. The main drawback of trajectory-based video anomaly detection is the performance decrease when crowd density increases. It is difficult to track all individuals when the density of the crowd is high and also, the trajectories will look clumsy which makes anomaly detection difficult. Non-trajectory based method or holistic method is more suitable to detect video anomaly detection in crowded scenes. Non-trajectory based methods can further be classified as follows based on the approaches followed. They are probabilistic based, distance based, information theoretic based, domain based and reconstruction based (Fig. 1).

The probabilistic-based model assumes that abnormal events have a low probability than normal events. The model trains with the data and calculates its probability. The probability which is lower than the threshold is classified as anomalous [17]. The distance-based model assumes that the normal events occur in the dense region and the anomalous event will be far from normal events. The nearest neighbor or clustering algorithm is used to calculate distance which finds similar data points. Domain-based models train the data and learn its classes which forms a boundary. With respect to this boundary, the location of test data is calculated and the abnormal event is detected. Information theoretic based models assume that the anomalous event has a big impact on changing information in the dataset. So, the data is considered as anomalous if its removal makes a big information change in the entire dataset.

Reconstruction error based models assume that anomalous events have high reconstruction error than the normal events. Generally, the image sequences are encoded and decoded by any specific model. Then the reconstruction error is calculated. If the reconstruction error is higher than the given threshold, then it is considered to be abnormal events. Video abnormal event

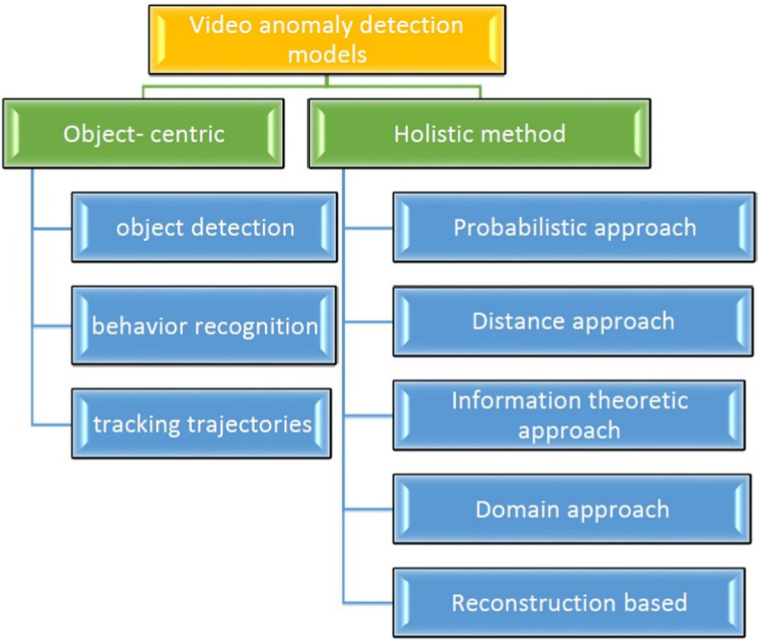

Fig. 1 Classification of video anomaly detection models

detection is having high importance because of its real-time applications. In most cases, the monitored environments are public areas. These areas are often crowded. So, designing video anomaly detection in crowded scenes is again an important task. It is familiar that abnormal labels are not available in all collected data. In spite of the lack of labels, the video anomaly detection system should be able to detect abnormal events. To solve these challenges, we proposed an unsupervised deep learning framework for local event detection in crowded scenes.

1.1 Research challenges

1. Video anomaly detection is challenging due to distortion in camera viewpoint, variations in both intra and inter normal class and versatility of anomalies.
2. Anomaly events in surveillance videos rarely occur so, monitoring the events manually is a tedious and meticulous task.
3. Abnormal events vary in a different environment. Let us see with an example, people running at a beach would be normal but running at a shop or movie hall would be abnormal.
4. There is no universal definition for the anomaly.
5. Finding video anomalies is a highly difficult task because it occurs very rare and also lasts only for a short period of time.
6. Obtaining training datasets which cover all type of normal behavior is a challenging one.
7. Video anomaly detection in the crowded scenario has additional challenges because of the huge number of people on foot in nearness, the unpredictability of individual appearance, the successive incomplete impediments that they create, and the unpredictable crowd movements.

8. Crowd scenes frequently contain vulnerabilities, for example, size, density, and shape of subjects vary. The limits of these elements that can convey ambiguities to the definition and elucidation of the importance crowd anomalies.

1.2 Research motivation

Monitoring surveillance videos for finding abnormal events manually is a herculean job which raises the need for intelligent video surveillance. As there is no universal definition for video anomalies, either we should label anomalies manually or design the unsupervised model. The first choice is again tedious and also expensive. Further, the research challenges discussed in section 1.1 motivates us to design an unsupervised deep learning framework for local anomaly event detection in crowded scenes which is capable to detect abnormal event without class label and able to achieve better accuracy than the existing systems.

1.3 Research contribution

The main aim of the proposed framework is to improve effectiveness of video anomaly detection without using class labels.

1. We combine the benefits of both hand-crafted feature and hierarchical feature learning in the proposed framework for effective anomaly detection.
2. We proposed an unsupervised Hybrid Deep Learning framework for Video Anomaly Detection (HDLVAD) which detects video anomalies effectively.

This paper is structured as follows: section 2 describes existing works on unsupervised video anomaly detection. Section 3 defines video anomaly detection in the context of reconstruction based. In section 4, the proposed framework Hybrid Deep Learning Video Anomaly Detection (HDLVAD) is elaborated. Section 5 describes deep learning architecture, dataset used and experiment method performed. Results are discussed in section 6. Finally, section 7 concludes with summary and future work.

2 Related works

Anomaly detection in surveillance video has gained attention from both computer vision researchers and applications developers. Earlier research works of reconstruction based video anomaly detection can be divided into two groups. One is based on hand–crafted features and other one is based on feature learning. Hand-crafted feature-based models are less effective as the extracted features are done manually. Hand–crafted feature based models works by extracting spatial and temporal features from raw video manually. Algorithms such as optical flow, MHOF (multi-scale histogram of optical flow), HOG (histogram of oriented gradients), HOF (histogram of optical flow) [7] and saliency information are used for extracting features. Then these features are given as input to the classification model. Finally, it results whether the given data is an anomaly or not. Shifeng Li et al., [18] proposed model which extract motion features using optical flow and calculates the likelihood for finding an anomaly in input frames. Additionally, to reduce the computational cost, they performed background subtraction. This method may lead to failure in detecting anomalies

sometimes, due to its likelihood computation from the global maximum grid in training data. Zhijun Fang et al., [9] used saliency information and MHOF (Multi histogram optical flow) for extracting low-level features and PCANet for extracting high-level features. Then the output is given to the SVM model for anomaly classification. Tian Wang et al., [37] proposed a Hidden Markov model for anomaly classification and HOFO (Histogram of optical flow orientation) is used for feature extraction purpose. Dandan Ma et al., [23] proposed anomaly detection through online learning. Feature extraction is done by optical flow algorithm. It selects experts which are outstanding and updated dynamically. Medhini G Narasimhan & Sowmya Kamath [25] proposed Gaussian classifier as a model to classify anomaly. Local features are extracted using SSIM and global features are extracted using sparse denoising autoencoder. Roberto Leyva et al., [16] performs compact feature extraction. Both optical flow and HOF are used for feature extraction. GMM (Gaussian mixture model) performs anomaly classification. Because of less feature consumption, this work is suitable for online performance. Jing Wang & Zhijie Xu [35] performs feature extraction using statistical model and wavelet transform detects anomaly frame. It achieves good real-time performance. Ying Zhang et al., [44] extracts spatial features using SVDD and detects the spatial anomaly. Similarly, optical flow is used to extract motion features and then detects motion anomaly. As a final step, spatial anomaly and motion anomaly are combined to detect abnormal events. This method is effective even though it avoids too many parameters. Weixin Li et al., [17] proved that mixture dynamic texture is better than optical flow in extracting temporal features. Spatial features are extracted using saliency information. The discriminative model combines scores of both anomaly maps (spatial and temporal). Yugen Yi et al., [42] considers the manifold structure of video data and proposed a sparse representation model. This model learns reconstruction vector in the test dataset. Sparsity reconstruction cost is calculated. If SRC (sparsity reconstruction cost) is above the threshold, then the frames are labeled as an anomaly. Peng Liu et al., [21] detects anomaly by double sparse representation. MHOF is used for feature extraction. Two different sparse representation gives two results. Fuzzy classifier combines it and finally detects an anomaly. Dictionary gets updated whenever training samples are added. The major advantage of this work is updating dictionary dynamically. Yuan et al., [43] proposed a dictionary learning method for sparse representation. Limitation of this work is normal events can be characterized as sparse linear combinations in dictionary learning, but abnormal events cannot be done. Shifeng Li et al., [19] implemented a global grid motion template and optical flow for local grid template. Local histogram and global histogram are plotted. The assumption of this work is histogram bins differs for normal and abnormal events. Shifu Zhou et al., [47] extracts the moving pixels in video frames using optical flow and the convolutional neural network is trained to detect t abnormal frame. Rima Chaker et al., [4] extract features using social similarity measure. The frames are divided according to spatiotemporal features. These features are given as input to the local social network model. Combining all these, the global social network model is formed which decides whether the frame is anomalous or not. This work covers both local and global anomaly detection. Updating global network makes possible to extend this work for online detection. Siqi Wang et al., [36] segment frames into spatiotemporal cuboids, then foreground localization is done. Spatial features are extracted by SL-HOF algorithm and temporal features are extracted by ULGP-OF algorithm. These features are given as input to one class extreme learning machine model separately. As a result, the model detects anomalous frames. Ying Zhang et al., [45] proposed robust anomaly detection. It is effective and also suits for online performance. They used locality sensitive hashing filters to isolate anomaly pixels. Dan Xu et al., [41] fuse both stacked denoising auto encoder and optical flow for feature extraction and one class SVM detects abnormal events. Hanhe Lin et al., [20] performs online weighted clustering for anomaly detection. Adaptive MHOF multi histogram optical flow) extracts

temporal features. This work also suits online detection due to its adaptability. S Amraee et al., [2] connected component analysis to find out suitable cell size. Then, important regions alone extracted by eliminating repeated information. Histogram optical flow gradient method and the Gaussian model are used for extracting spatial features. By calculating average optical flow in cells, temporal features are calculated. The main advantage of this method is it reduces the size of training data in the pre-processing stage. S Amraee et al., [1] proposed method to detect abnormal events based on one class SVM. Proper cell size is computed with the help of normal size of objects in training data. HOG-LBP method is applied to extract appearance features. HOF is applied to extract motion features. Then, it is given as input to one class SVM model. The main benefit of this approach is that this framework is capable to differentiate between overlapping of normal objects and abnormal objects. KW Cheng et al., [5] proposed framework which aims to find a local abnormal, global abnormal event and both at the same time. Frequent geometric relations are found to solve the problem of detecting interaction in normal events. These interaction templates are modelled using Gaussian process regression. The advantage of this framework is that it can handle noisy data and imbalance data. X Hu et al., [13] proposed histogram of oriented contextual gradient descriptor method. This method applies gradient descriptor based on context rather than the pixel. Hence, this descriptor efficiently detects contextual information in the video data. This method outperforms traditional HOG based on pixel.

In contrast, feature learning based method models don't have a separate module for feature extraction and classification. It learns the hierarchical features automatically and then classify the given data as an anomaly or not. Feature learning based models are effective but its computational cost is high. Hanh TM Tran & DC Hogg [33] combines auto encoder and one class SVM, where the auto encoder learns features and one class SVM classifies normal and anomalous frames. The main disadvantage of this system is it incurs a high computational cost. Yong Shean Chong & Yong Haur Tay [6] proposed spatiotemporal auto encoder which learns features and detects abnormal events. Limit of this work is as crowd complexity increases, false alarm may also increase. Mohammad Sabokrou et al., [30] used the cascaded network. Feature extraction is done by cascaded stack auto encoder and classification is done by the cascaded convolutional neural network. Yachuang Feng et al., [11] also performs feature extraction using stacked denoising auto encoder but LSTM (long short temporal memory) network calculates scores for each frame. Graph based ranking model is used in this work. Achieves less false alarm rates. Hung Vu et al., [34] cluster input frames and then RBM (Restricted Boltzmann Machine) is trained for reconstructing images. If the reconstruction error reaches above the given threshold, then the frame is labelled as abnormal. Manassés Ribeiro et al., [28] proposed convolutional auto encoder to reconstruct image frames. Effect of input frames, spatiotemporal filters, and video complexity have been well studied. Yiru Zhao et al., [46] proposed 3D convolutional auto encoder to reconstruct image frames. Instead of finding only reconstruction error, weight decreasing prediction loss is also calculated. Finally, the regularity score is computed to find abnormal events. M Sabokrou et al., [29] combines both auto encoder and sparse representation for feature extraction. Cascade classifier performs anomaly detection. Y. Feng et al., [10] propose deep Gaussian mixture model for anomaly detection but the feature learning is done by PCANet. X Wang et al., [38] proposed deep auto encoder which is a combination of two networks namely 3DCNN and convolutional GRU. Features are learned by the 3DCNN network and bidirectional convolutional GRU. Both spatial and temporal features are learned. The network has two branches called reconstruction branch and prediction branch. Reconstruction branch is responsible to reconstruct the input frames whereas prediction branch is responsible for learning features during the training phase. The advantage of this model is over fit of the data is avoided and accuracy is improved. M Sabokrou et al., [31] combined pre-trained convolutional network called AlexNet and

another convolutional layer which is trained with the given input data. This fully convolutional neural network architecture gives good accuracy at 370 fps speed. Tudor Ionescu et al., [15] proposed a framework based on unmasking method which doesn't need training sequences. A binary classifier is trained to differentiate between two videos while evacuating at each progression the most discriminant features. If training accuracy is high, then the events are recognized as abnormal events. This method has been compared with both supervised and unsupervised approaches. It shows the best accuracy with 20fps. M Ravanbakhsh et al., [27] applied generative adversarial network to detect abnormal events in videos. The network is trained only with normal distribution data. The generator is used as a supervisor for discriminator and vice versa. During the testing period, discriminator detects abnormal events without using any label. The generator may learn trivial identity function. So, it has been forced to transform raw data. This method outperforms traditional methods in both frame level and pixel level evaluation. S Huang et al., [14] proposed a multi-modal framework based on deep learning. Mid-level features are extracted using CDBN and high-level features are learned by using multimodal RBM which is stacked above CDBN. This method shows an effective performance. T Wang et al., [39] proposed S2VAE algorithm which is a combination of two networks SF-VAE and SC-VAE. The first one is a shallow network whereas the latter one is a deep network. SF VAE is responsible to eliminate less important samples and learning features. The second network SC VAE is responsible to detect abnormal event detection. This approach is experimented in four datasets and proved the effectiveness of it.

It is well-known that most of the existing method works either using the handcrafted feature or learned features. Hand-crafted feature based methods are less effective but has less computational cost. In the other hand feature learning based methods are highly effective but has high computational cost. In proposed work, we address this gap by combining both handcrafted feature and feature learning method to achieve better accuracy with less computational cost. Canny Edge detection algorithm is used to extract spatial features manually and features are learned by deep learning model. Both ROC-AUC and EER performance metrics are used for evaluation.

3 Anomaly detection

We propose a reconstruction based anomaly detection framework. In this category, the main hypothesis is that abnormal frames in video differ from normal frames. Hence, reconstruction error of abnormal frames will be higher than the reconstruction error of normal frames. The Euclidean distance between the original image and the reconstructed image is calculated as a reconstruction error.

$$D(t) = \|x(t) - l_w(x(t))\|_2 \tag{1}$$

Where $D(t)$ refers to Euclidean distance of frame t.

l_w refers to learned weights by the model

To find abnormality value, $D(t)$ is scaled from 0 to 1,

$$x_a(t) = \frac{D(t) - D(t)_{min}}{D(t)_{max}} \tag{2}$$

Finally, regularity score is calculated by simply subtracting abnormality score from 1,

$$x_r(t) = 1 - x_a(t) \tag{3}$$

4 Proposed framework

In the proposed framework, raw video is divided into frames. All the frames are pre-processed according to the need of a model. Edge frames are produced using a canny edge detector. These edge frames serve as a spatial feature. Combination of original frames and edge frames are given as input to the model. Here, both original frames and edges frames are packed into a single sequence. The input is given as single channel not as separate two channels to the deep learning model. As we are giving spatial information manually to the model, the number of convolution layers in the model can be reduced. Hence, less number of parameters is utilized. Inspired by [6], the proposed model also consists of both convolutional auto encoder and convolutional LSTM such that spatial features are learned by convolutional layers and temporal features are learned by LSTM. But architecture and input given to the model differs from [6]. Both training and testing are unsupervised. Ground truths are used only for evaluating purpose. The output of this model is reconstructed frames. Finally, the reconstructed error is calculated to detect an anomaly (Figs. 2, 3, 4, 5 and 6).

4.1 Pre-processing

In this phase, raw data frames are processed such that it is suitable for the model. Each video is divided into frames. The frames are resized to 128 × 128. Pixel values in all the images are

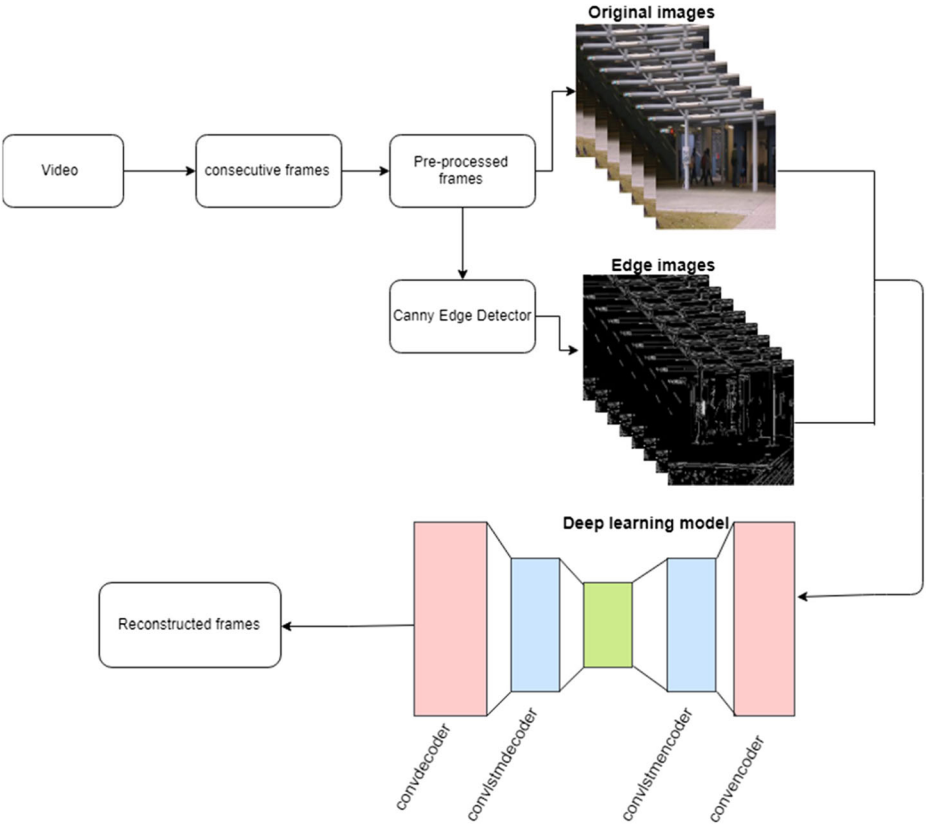

Fig. 2 Proposed deep learning framework (HDLVAD)

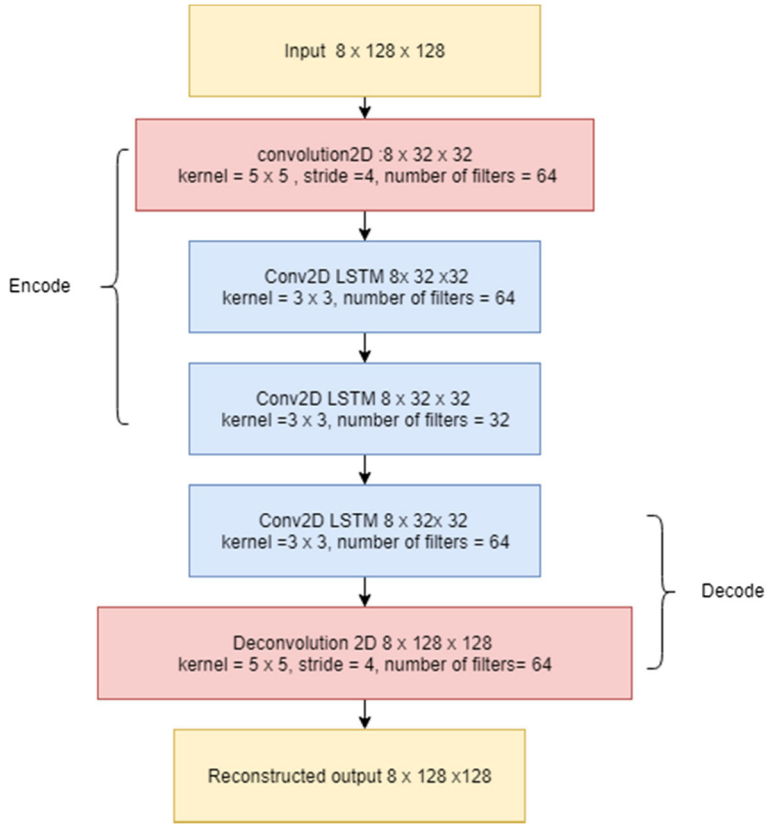

Fig. 3 Architecture of convolutional auto encoder - ConvLSTM model

scaled from 0 to 1, to ensure all the frames are in the same scale. As a next step, each and every frame is subtracted from the global mean image to achieve normalization. Calculating the average of pixel values at each location provides us with a global mean image. Video data are high dimensional data. To reduce dimension, we converted all images to grayscale.

Fig. 4 Normal(top) and abnormal(bottom) samples of avenue dataset

Fig. 5 Normal (top) and abnormal(bottom) samples of UCSD ped1 dataset

4.2 Appearance filter

Canny edge detector algorithm is used as the appearance filter. The motivation behind edge detection is to altogether decrease the amount of information in an image while protecting the structural properties to be utilized for image processing. Canny edge detector algorithm is a standard method created by John F. Canny in the year of 1986 [3]. This algorithm has five stages, smoothing, finding gradients, Non-maximum suppression, double threshold and edge tracking. It is known that all the raw images contain at least a few noises. So, to reduce noise the images are smoothed using Gaussian filter. Generally, the algorithm finds the edges on the basis of the assumption that intensity of grayscale images is high in edges. Gradients of images reveal where the intensity is higher compared to other regions. Sobel-operator calculates gradients of each pixel in the given image. As a next step, blurred edges are converted to sharp edges using non-maximum suppression method. Double threshold is applied to select only strong edges. Edge tracking is the final step in which strong edges are included in the final image whereas weak edges are included only if it is connected with strong edges.

Algorithm 1: Canny Edge detector
Input: Original Image
Output: Edge image

1. Noise reduction by Gaussian filter
2. Calculate intensity gradient I_x , I_y and magnitude M using Sobel operators.

   i. $M = \sqrt{I_x^2 + I_y^2}$

3. For each pixel calculate gradient vector orientation $\theta = a\,tan2(I_x \wedge I_y)$. Calculate new array A : if

   $$M(p_a) < M(p) < M(p_b)$$

   p is a current pixel, p_a and p_b are neighbour pixels, then $N(p) = M(p)$, else $N(p) = 0$.

4. Determine two thresholds of M, t_{min} and t_{max}. From $N(p) \geq t_{max}$, find paths of all pixels with $N(p) \geq t_{max}$ and include them in the image.

Fig. 6 Normal(top) and abnormal(bottom) samples of UCSD ped 2 dataset

4.3 Convolutional auto encoder

Auto encoders are one of the unsupervised deep learning algorithms. It encodes and then decodes the given image. Traditional auto encoders fail to preserve spatial properties and also produce repeated parameters. Hence, convolutional auto encoder is developed by Jonathan Masci et al., in the year of 2011 [24]. It has an alternate point of view in filter definition task. In other convolution algorithms, filters are defined manually. On the contrary, the convolutional auto encoder learns the optimal filters that lead to less reconstruction error. The learned filters can be used in any other image classification task. CAE is considered as a kind of convolution neural networks. The main difference is that CNN learns the features and performs classification whereas, CAE learn the features and reconstruct the images. Former is supervised and the latter one is unsupervised. It is also scalable for high dimensional images due to its convolutional nature.

4.3.1 Encode

It is well-known that a single convolutional filter is not able to learn all the features from the input image. So, let us say there are N convolutional filters for each convolutional layers with input depth D.

The convolution with the input volume $X = \{ X_1, X_2, X_3, \ldots, X_N \}$ and a set of N convolutional filters $\left\{ f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, \ldots, f_N^{(1)} \right\}$, each with input depth D, produces a set of N activation maps.

$$O_m(x,y) = \left(\sum_{d=1}^{D} \sum_{u=-2z-1}^{2z+1} \sum_{v=-2z-1}^{2z+1} f_{md}^{(1)}(u,v) X_d(x-u, y-v) \right) \tag{4}$$

Where, $m = 1, 2, \ldots, n$ and a= non-linear activation function.

Each convolution is wrapped by the activation function, such that in the training phase the model is trained to learn how to represent input by combining non-linear functions.

$$R_m = O_m = a\left(X * f_m^{(1)} + b_m^{(1)} \right) \tag{5}$$

Where, =1, 2, … . , m, $b_m^{(1)}$ = bias for the m^{th} feature map and R_m = latent variable. These created activation maps are termed as encoding for the input X in low dimensional space. Over all, the encoded convolution with width W and height h is,

$$\begin{aligned} O_w = O_h &= (X_w + 2(2z+1)-2)-(2z+1)+1 \\ &= X_w + 2(2z+1)-1 \end{aligned} \tag{6}$$

4.3.2 Decode

The latent representation $R_{m=1,2,\ldots,N}$ is used as input for the decoder to reconstruct image . Hyper parameters of the decoder are fixed by the encoding structure including filters volume f and number of filters D. Hence, the reconstructed image $\tilde{X}$ can be produced using convolution between the feature map volume $R = \{R_{x=1}\}^N$ and the convolution filter $f^{(2)}$ with the dimensions $(2z+1, 2z+1, N)$.

$$\tilde{X} = a\left(R * f_m^{(2)} + b^{(2)}\right) \tag{7}$$

As a result, the deconvolution with width w and height h is,

$$O_w = O_h = X_w + 2(2z+1)-1-(2z+1)+1 = X_w = X_h \tag{8}$$

4.4 Convolutional LSTM

Convolutional LSTM was proposed by Shi et al., in the year of 2015 [40]. Traditional LSTM method has matrix operations which are replaced as convolution in convolutional LSTM method. For both input to hidden connection and hidden to hidden connection, the convolution operation is performed. This yields better spatial features. Forget layer F_t is computed as,

$$F_t = \sigma\left(w_f \otimes [h_{t-1},\ \ i_t, c_{t-1}] + b_f\right) \tag{9}$$

New information is added using,

$$N_t = \sigma(w_N) \otimes [h_{t-1},\ \ i_t, c_{t-1}] + b_N \tag{10}$$

$$\tilde{C} = \tanh(w_c \otimes [h_{t-1}, i_t] + b_c) \tag{11}$$

Both old information and new information is combined by,

$$C_t = F_t * C_{t-1} + N_t * \widetilde{C}_t \tag{12}$$

Outputs of the convolutional LSTM are computed as,

$$O_t = \quad \sigma(w_o \otimes [h_{t-1}, i_t, c_{t-1}] + b_o) \tag{13}$$

$$h_t = O_t * \tanh(C_t) \tag{14}$$

Where i denotes input images, b denotes bias, C denotes cell state and w denotes weights. Note that ⊗ symbol denotes convolution operation and ∗ denotes Hadamard product. Weights refer

Table 1 Details of anomalous events in the datasets

S. No	Dataset	Anomalous events
1	Avenue	Strange action, Abnormal object, and wrong direction
2	UCSD [ped1 and ped2]	Skaters, bikers, small carts, people walking in the grass and across the pathway

to convolution filters. Hence, convolutional LSTM works better with images than the traditional LSTM method. Spatial features are propagated to each state temporarily.

5 Experimental evaluation

This experiment is performed in Ubuntu operating system with Nvidia GeForce GPU. ReLU activation function is used in the proposed model as it is good in regularization and Adam optimizer to adapt learning rate automatically. The proposed model is trained with 50 epochs and 16 mini batch sizes. Validation data has been partitioned to the rate of 0.15 from training data. Weights are initialized by using RandomNormal method. Both training and testing are performed without ground truth (unsupervised learning).

5.1 Architecture

As we combine the original image and edge image to the input of the model, the proposed architecture has less number of layers than the conventional spatiotemporal model [6]. Input shape is $8 \times 128 \times 128$, where 8 is number of frames given as input at a time. The first 2D convolutional auto encoder layer has a kernel size of 5 with stride 4. Then follows three 2D convolutional LSTM layers with kernel size 3. First two layers act as an encoder and the Last layer acts as a decoder. At the end of the model 2D convolutional auto encoder layer has kernel size of 5 with stride 4 which acts as a decoder.

5.2 Video surveillance datasets

5.2.1 Avenue dataset

The videos are recorded in CUHK campus avenue. It consists of 16 training video clips and 21 testing video clips. This dataset complements the research work [22]. Training videos contain normal samples and testing videos contain both normal and abnormal samples. Few challenges in this dataset are slight camera shake, outliers included in training data and few normal patterns occur rarely in training data.

Table 2 Comparison of AUC scores of three datasets

Dataset	HDLVAD	SpatioTemporal autoencoder [6]	Convolutional autoencoder [28]
Avenue	90.7	83.2	80.5
UCSD ped1	98.4	95.7	91.7
UCSD ped 2	98.5	81.6	82.1

Table 3 Comparison of EER values of three datasets

Dataset	HDLVAD	SpatioTemporal autoencoder [6]	Convolutional autoencoder [28]
Avenue	14.5	28.3	28.4
UCSD ped1	0.75	10.0	22.0
UCSD ped2	0.92	21.7	16.3

5.2.2 UCSD dataset (ped1 and ped2)

The videos are recorded using a stationary camera mounted at a height such that it overlooks pedestrian walkways. The crowd size varies meagre to exceptionally crowded. Normal training videos contain only pedestrians. Abnormal actions occur due to two reasons, either non-pedestrians appear in walkways or anomalous motion pattern of pedestrians. All anomalies are normally happening, i.e. they were not arranged for the motivations behind collecting the dataset. The video dataset was part into 2 subsets, each comparing to an alternate scene. The Peds1 dataset consists of 34 training videos and 36 testing videos. It includes clasps of gatherings of individuals strolling towards and far from the camera. The Peds2 dataset consists of 16 training videos and 12 testing videos. It includes clips with pedestrians walking parallel to the camera (Table 1).

5.3 Performance metrics

In general, three kinds of evaluations are performed in video anomaly detection. They are pixel-level, frame-level and event- level. We followed frame-level evaluation as it is widely used in the literature. In this type of evaluation, abnormalities are checked in each frame. Even one abnormal pixel is detected, the whole frame is labelled as an abnormal frame. Frame-level ground truths are available in all five datasets. The main limitation is, the anomalous frames are detected, but the location where the anomaly occurred is not considered. We use the AUC-ROC curve as a performance measurement. The problem is a binary classification in which the model has to predict whether the frame is an anomaly or not. AUC curve measures the model performance under the various threshold. The better model is the one which achieves high

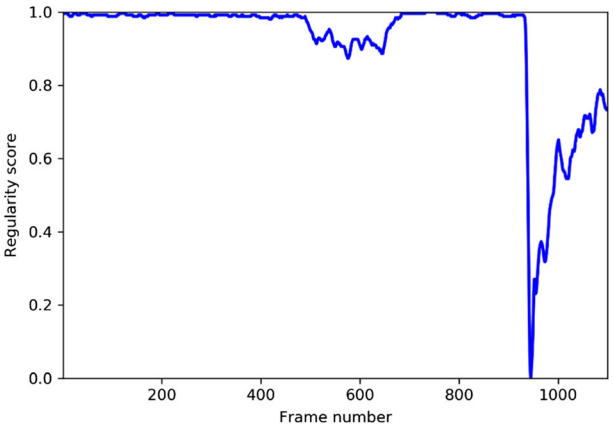

Fig. 7 Regularity score of frames in Avenue dataset

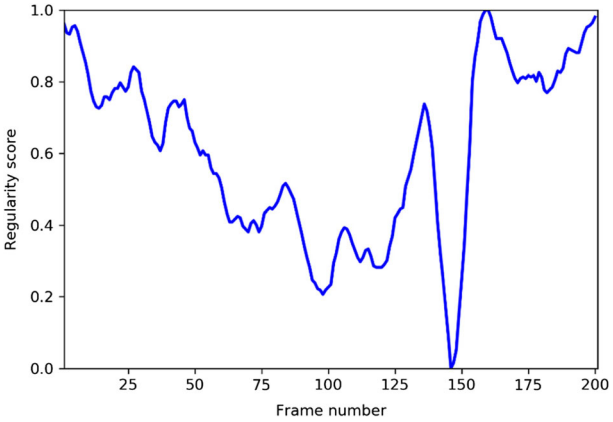

Fig. 8 Regularity score of frames in UCSD ped 1 dataset

AUC score. AUC-ROC curve is plotted with TPR (True Positive Rate) values on y-axis and FPR (False Positive Rate) values on the x-axis.

$$TruePositiveRate = \frac{TruePositive}{TruePositive + FalseNegative} \tag{15}$$

$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive} \tag{16}$$

EER (Equal Error Rate) as a performance metric is also calculated. EER is an optimized value where false positive intersects with a false negative. The better model has lower EER value.

6 Results and discussion

Table 2 shows the AUC scores of the proposed framework and other state of art methods in three different datasets. The proposed framework outperforms state art methods spatial

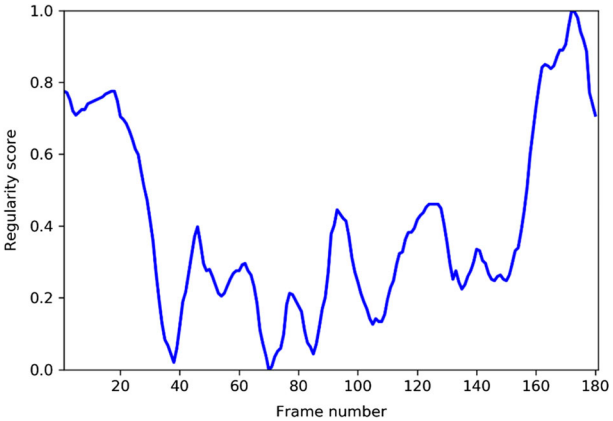

Fig. 9 Regularity score of frames in UCSD ped2 dataset

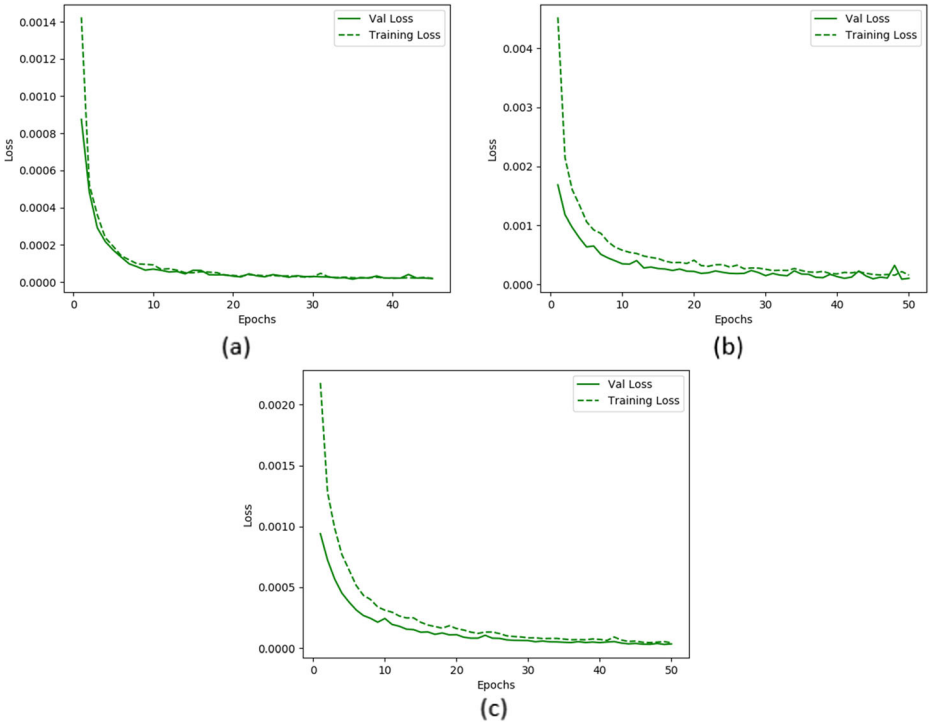

Fig. 10 Training loss vs Validation loss generated by the proposed method HDLVAD for (a) Avenue dataset (b) UCSD ped1 dataset (c) UCSD ped2 dataset

temporal autoencoder [6] and convolutional auto encoder [28]. Table 3 shows the EER values of the proposed framework and other state of art methods in three different datasets. As we mentioned before, the lower EER value, the better model. The proposed method HDLVAD achieves high AUC score and lower EER score in all the three datasets compared to state of art methods. As our method is based on reconstruction error, regularity score with frame number is plotted. Qualitative analysis is done by visualizing regularity score against frame.

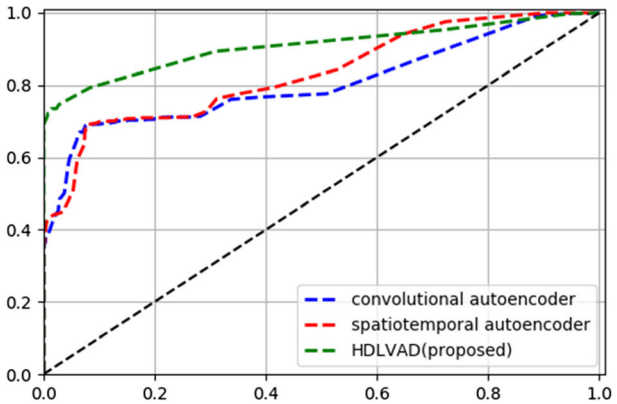

Fig. 11 ROC- AUC curve compare of proposed framework (HDLVAD) with other methods for avenue dataset

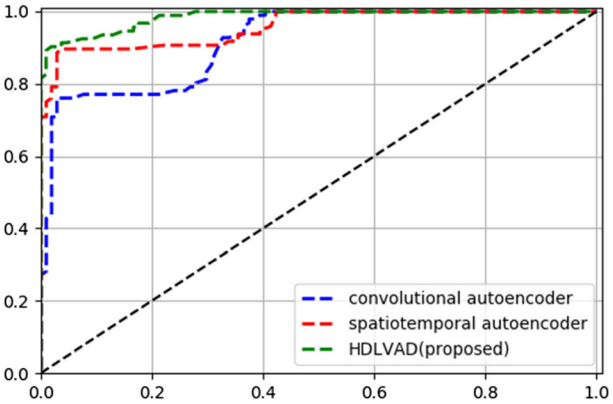

Fig. 12 ROC- AUC curve compare of proposed framework (HDLVAD) with other methods for UCSD ped 1 dataset

To visualize frame abnormality, regularity scores are plotted against each frame number. Normal frames will have high regularity score whereas abnormal frames will have low regularity scores. Figure 7 shows regularity scores of avenue dataset. We can clearly see the sudden drop in regularity score between the frame number 800–1000. It reveals that abnormal event occurs in those frames. Figure 8 shows regularity scores of UCSD ped1 dataset. In this dataset, there are few drops in regularity score in many frames. The weak abnormal events occurred in frames near 100 and strong abnormal events occurred in frames near 150. Similarly, Fig. 9 shows regularity scores of UCSD ped2 dataset. Strong abnormal events occurred in many frames such as near 40,70 and 80. Weak abnormal events occurred between 100 and 120. A good learning model should have good fit on the data. In other words, if the model learns the training data perfectly, then the new unknown data cannot be classified accurately. Similarly, if the model doesn't learn the data well, then also the new unknown data cannot be classified properly. Hence, over fitting and under fitting plays significant role. The proposed method HDLVAD neither over fit the data nor under fit the data.

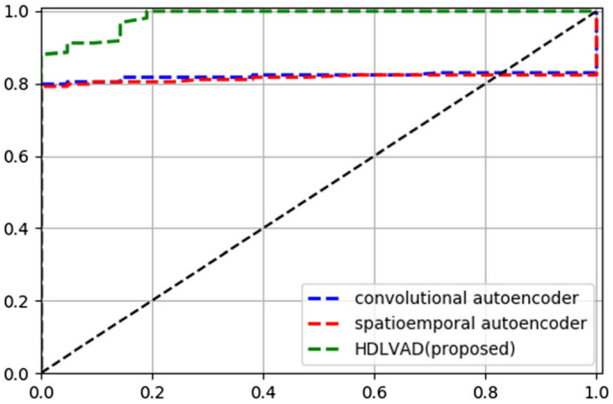

Fig. 13 ROC- AUC curve compare of proposed framework (HDLVAD) with other methods for UCSD ped 2 dataset

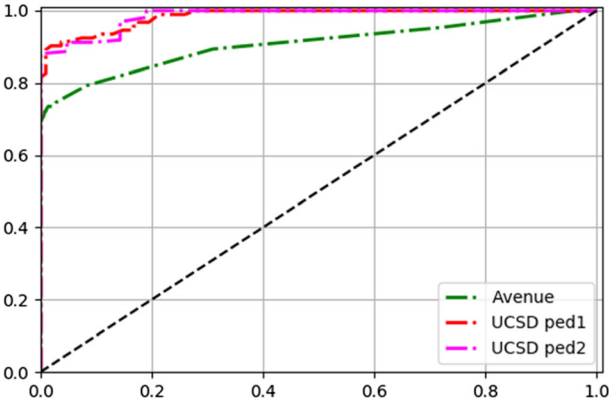

Fig. 14 ROC – AUC curve of proposed framework (HDLVAD) for three datasets

The fig. 10 shows the training loss and validation loss of the proposed framework HDLVAD of all the three datasets. Figure 10(a) reveals validation loss and training loss of avenue dataset. The model reaches minimum loss before 50 epochs. Figure 10(b) reveals validation loss and training loss of UCSD ped1 dataset. Similarly, Fig. 10(c) reveals validation loss and training loss of UCSD ped2 dataset. As there is not much difference between training loss and validation loss in all the three datasets, the HDLVAD method has good fit on the data. Comparison of the proposed work with state of art methods is given in Fig. 11 for avenue dataset, Fig. 12 for UCSD ped 1 dataset and Fig. 13 for UCSD ped 2 dataset. The overall performance of the proposed framework in three datasets is shown in the Fig. 14. ROC- AUC curve interprets that our model performs better in all three datasets.

7 Conclusion and future work

Abnormal event detection in video data has many real-time applications. Hence it is a significant research area in computer vision domain. We proposed unsupervised deep learning framework(HDVLAD) for anomaly detection in videos. Input frames are resized to 128 × 128 and normalized in pre-processing stage. Edge frames are extracted using canny edge detection algorithm. Both the input frames and edge frames are given to convolutional autoencoder and convolutional LSTM model. Reconstructed images are given as output. Finally, reconstruction error is calculated and abnormal events are detected. AUC and EER performance metric is used for evaluation. The proposed framework achieves 90.7% accuracy in avenue dataset, 98.4% accuracy in UCSD ped1 dataset and 98.5% accuracy in UCSD ped 2 datasets. Combining hand crafted features and learned features made model effective. Limitation of the proposed work is that it detects only the frame in which abnormal event occurs. It doesn't localize the abnormal object in those frames. As a future work, video streaming in big data can be investigated.

References

1. Amraee S, Vafaei A, Jamshidi K, Adibi P (2018) Abnormal event detection in crowded scenes using one-class SVM, vol. 12

2. Amraee S, Vafaei A, Jamshidi K, Adibi P (2018) Anomaly detection and localization in crowded scenes using connected component analysis. Multimed Tools Appl 77(12):14767–14782
3. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell (6):679–698
4. Chaker R, Aghbari ZA, Junejo IN Social network model for crowd anomaly detection and localization. Pattern Recogn 61:266–281
5. Cheng KW, Chen YT, Fang WH (2015) Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. pp. 2909–2917
6. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: International Symposium on Neural Networks, pp. 189–196
7. Colque RVHM, Caetano C, Andrade MTL, Schwartz WR (2017) Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Transactions on Circuits and Systems for Video Technology 27(3):673–682
8. Cong Y, Yuan J, Liu J (2013) Abnormal event detection in crowded scenes using sparse representation. Pattern Recogn 46(7):1851–1864
9. Fang Z, Fei F, Fang Y, Lee C, Xiong N, Shu L, Chen S. (2016) Abnormal event detection in crowded scenes based on deep learning, vol. 75
10. Feng Y, Yuan Y, Lu X:X (2016) Deep representation for abnormal event detection in crowded scenes. Neurocomputing 219:591–595
11. Feng Y, Yuan Y, Lu X (2017) Learning deep event models for crowd anomaly detection. Neurocomputing 219:548–556
12. Haering N, Venetianer PL, Lipton A (2008) The evolution of video surveillance: an overview, vol. 19
13. Hu X, Huang Y, Duan Q, Ci W, Dai J, Yang H. (2018) Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor, vol. 2018
14. Huang S, Huang D, Zhou X (2018) Learning multimodal deep representations for crowd anomaly event detection
15. Ionescu RT, Smeureanu S, Alexe B, Popescu M (2017) Unmasking the abnormal events in video. pp. 2895–2903
16. Leyva R, Sanchez V, Li CT (2017) Video anomaly detection with compact feature sets for online performance. IEEE Trans Image Process 26(7):3463–3478
17. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32
18. Li S, Liu C, Yang Y (2018) Anomaly detection based on maximum a posteriori. Pattern Recogn Lett 107: 91–97
19. Li S, Yang Y, Liu C (2018) Anomaly detection based on two global grid motion templates, vol. 60
20. Lin H, Deng JD, Woodford BJ, Shahi A (2016) Online weighted clustering for real-time abnormal event detection in video surveillance. pp. 536–540
21. Liu P, Tao Y, Zhao W, Tang X (2017) Abnormal crowd motion detection using double sparse representation. Neurocomputing 269:3–12
22. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. pp. 2720–2727
23. Ma D, Wang Q, Yuan Y (2014) Anomaly detection in crowd scene via online learning. p. 158. ACM
24. Masci J, Meier U, Ciresan D, Schmidhuber J. (2011) Stacked convolutional auto-encoders for hierarchical feature extraction
25. Narasimhan MG, Kamath S (2017) Dynamic video anomaly detection and localization using sparse denoising autoencoders
26. Piciarelli C, Micheloni C, Foresti GL (2008) Trajectory-based anomalous event detection. IEEE Transactions on 18(11):1544–1554
27. Ravanbakhsh M, Sangineto E, Nabi M, Sebe N (2019) Training adversarial discriminators for cross-channel abnormal event detection in crowds. IEEE
28. Ribeiro M, Lazzaretti AE, Lopes HS (2018) A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recogn Lett 105:13–22
29. Sabokrou M, Fathy M, Hoseini M (2016) Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electron Lett 52(13):1122–1124
30. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process 26(4):1992–2004
31. Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deepanomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput Vis Image Underst 172:88–97
32. Sodemann AA, Ross MP, Borghetti BJ:BJ (2012) A review of anomaly detection in automated surveillance. IEEE Trans Syst Man Cybern Part C Appl Rev 42(6):1257–1272
33. Tran HT, Hogg DC. (2017) Anomaly detection using a convolutional winner

34. Vu, H., Phung, D., Nguyen, T.D., Trevors, A., Venkatesh, S. (2017) Energy-based Models for Video Anomaly Detection. arXiv preprint arXiv:1708.05211
35. Wang, J., Xu, Z. (2015) Crowd anomaly detection for automated video surveillance
36. Wang S, Zhu E, Yin J, Porikli F (2018) Video anomaly detection and localization by local motion based joint video representation and OCELM. Neurocomputing 277:161–175
37. Wang T, Qiao M, Deng Y, Zhou Y, Wang H, Lyu Q, Snoussi H (2018) Abnormal event detection based on analysis of movement information of video sequence. Optik-International Journal for Light and Electron Optics 152:50–60
38. Wang, X., Xie, W., Song, J. (2018) Learning Spatiotemporal Features With 3DCNN and ConvGRU for Video Anomaly Detection
39. Wang T, Qiao M, Lin Z, Li C, Snoussi H, Liu Z, Choi C (2019) Generative neural networks for anomaly detection in crowded scenes. IEEE Transactions on Information Forensics and Security 14(5):1390–1399
40. Xingjian SHI, Chen Z, Wang H, Yeung DY, Wong WK, Woo W (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting
41. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst 156:117–127
42. Yi Y, Li X, Zhao R, Bi C, Wang J, Sun H. (2016) A constrained sparse representation approach for video anomaly detection
43. Yuan Y, Feng Y, Lu X (2018) Structured dictionary learning for abnormal event detection in crowded scenes. Pattern Recogn 73:99–110
44. Zhang Y, Lu H, Zhang L, Ruan X (2016) Combining motion and appearance cues for anomaly detection. Pattern Recogn 51:443–452
45. Zhang Y, Lu H, Zhang L, Ruan X, Sakai S (2016) Video anomaly detection based on locality sensitive hashing filters. Pattern Recogn 59:302–311
46. Zhao Y., Deng B, Shen C, Liu Y, Lu H, Hua XS (2017) Spatiotemporal autoencoder for video anomaly detection. pp. 1933–1941
47. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatialtemporal convolutional neural networks for anomaly detection and localization in crowded scenes, vol. 47

Anitha Ramchandran has received her Masters of Computer Science from Vellore Institute of Technology, in 2015. She is currently pursuing her Mtech by Research from Vellore Institute of Technology, India. Her area of research interest includes machine learning and video surveillance.

Arun Kumar Sangaiah received his Master of Engineering from Anna University and Ph.D. from VIT University, in 2007 and 2014, respectively. He is currently working as a Professor at the School of Computing Science and Engineering, VIT University, Vellore, India. His areas of research interest include machine learning, software engineering, computational intelligence, IoT, wireless networks, bio-informatics, and embedded systems. Dr. Sangaiah's outstanding scientific production spans over 250+ contributions published in high standard ISI journals, such as IEEE-Communication Magazine, IEEE Systems and IEEE IoT. His publications are distributed as follows: 250 papers indexed in ISI-JCR (Q1: 90, Q2: 80, Q3: 40, Q4: 50) and 21 papers indexed in Scopus.