2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)
Karnataka, India. Dec 23-25, 2022

2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) | 978-1-6654-5499-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMARTGENCON56628.2022.10084028

Anomaly Detection Methods in Surveillance Videos: A Survey

1st Lokesh Borawar
Dept. of Computer Science
Chandigarh University
Mohali, India
borawarlokesh26@gmail.com

2nd Dr. Ravinder Kaur
Dept. of Computer Science
Chandigarh University
Mohali, India
drravinder2920@gmail.com

***Abstract*—Surveillance system continuously generates massive amount of video data in the newest technological era, analysing these data is a tedious task for security specialists. With the popularization of surveillance monitoring system and the evolution of information technology, how to immediately and without human interaction detect unusual behaviours in surveillance footage is becoming more and more crucial for smart cities and public safety. Finding of abnormal footage physically in these massive video recordings is a laborious work, as they do not happen often in the real world. This clearly shows the necessity of automated anomaly detection, afterward that can detect crimes and aid investigations. The progress of anomaly detection has substantially benefited from deep learning, and much outstanding work has been published on this subject. This survey paper provides a comprehensive review of various anomaly detection and recognition methods. Researchers will get a better perspective of anomaly detection task with GAN approach, fine-tuned approach, and keyframes extraction plus shallow network approach and also their issues as well.**

***Index Terms*—anomaly detection, intelligent surveillance networks, keyframe extraction, spatial augmentation, 3DConvNets, U-Net**

I. Introduction

In the new century era, people lost lives due to increase in crime rate [1] and this is the prime reason for it. One possible time and cost-efficient solution for detecting unforeseen criminal events is a smart and automatic video surveillance system. For public safety, government has put large number of surveillance cameras in various areas. Due to the manual monitoring constraint, it makes difficult for law enforcement organizations to identify or avoid abnormal actions. An intelligent computer vision method is required to identify odd behaviour through frames, one that can distinguish between normal and abnormal situations with no human input. Benefit of such an automated system is not to do only monitoring, but its also decreases 24-h manual video observation done by humans.

In the literature [2]–[5], anomaly detection methods depended on sparse coding have shown auspicious outcomes. These methods are taken as standard for identifying anomalies. These methods are trained to create a dictionary of typical events from the first few frames of a footage. Although computationally efficient, this method performs poorly when it comes to reliably identifying anomalous occurrences. Anomaly detection methods based on weakly multi-instance learning (supervised) (MIL) are also investigated in [3], [5], [6]. In such methods, the footages are sundered into predetermined parts throughout the training phase. These parts create a collection of examples for both negative and positive samples.

Because of sparse coding methods dynamic nature, they are not ideal for surveillance contexts. For example, changing the dictionary to abnormal events from normal events steers in a significant amount of false-negative and false-positive results. Also, it is extremely challenging to spot abnormal events in noisy, low-resolution recordings. Machines must rely on visual features, whereas people can distinguish uncommon or regular events based on their intuition. The existing approaches mostly have a large percentage of false alarms, which lowers performance. Furthermore, these methods have a restricted performance when used in real-world scenarios despite performing well on tiny datasets.

Researchers have built many methods. In short, it can be understood using three categories. One is an autoencoder. In autoencoder [31]–[34], model tries to learn how to generate a normal frame and not an abnormal frame. Researchers proposed different models and approaches to build an autoencoder. The second category has pre-trained models [27], [35]. To train on anomaly videos classification tasks researchers have used different pre-trained model which already trained on video data and ready to use as temporal features extraction. The third category has unique techniques like weakly labels [5], [19], [20] and keyframes extraction [25] etc. Upcoming sections describe each category by one method in detail.

TABLE I
Area under the curve in percentage for different datasets.

Method	UCSD ped1	UCSD ped2	CUHK Avenue	UCF Crime
EADN	93	97	97	98
3DConvNets	-	-	-	82
Bidirectional prediction	87.24	96.54	87.39	-

Only developing a good model is not an approach to detect anomaly in given frames. Extracting key frames from a video that can define an activity clearly also be an approach. So, section III (EADN) is about extracting key frame and a shallow network. A model which is already trained on video classifi-

cation task also can use to classify anomaly task, described in section IV (3DConvNets). A GAN approach is proposed in section V (Bidirectional prediction). First model learns to generate a normal frame, so that way it will understand what a normal video contains. And when at test time if a normal video came, model will generate the frame with less error but if an anomaly video came, model will not be able to generate the frame with less error. The authors have given a complete explanation of these methods in their paper [25], [27], [31] respectively and AUC of these methods showed in I.

II. Literature review

The literature on anomaly detection techniques is divided into two primary groups: traditional handmade feature-based techniques and deep learning feature-based methods for recognising anomalous events. Anomaly detection used to be very reliant on low-level, manually created feature-based approaches. These approaches are typically based on 3 stages. First, recognising patterns (low-level) from the training set known as feature extraction. Second, differentiating encoding normal occurrences distribution known as feature learning. Last one is, separation of clusters and outliers which are recognised as aberrant events. As, Zhang et al. [7] used spatiotemporal features to characterise frequent occurrences using the Markov random field. Likewise, in a social interaction model created by Mehran et al. [8], cooperative forces were estimated and optical flow was used to distinguish between normal and abnormal behaviour. In further, an explainable anomaly detection methodology was put forth by Nor et al. [9] to help with prognostic and health management PHM. A Bayesian deep learning model with predetermined prior and likelihood distributions serves as the foundation of their methodology. To develop PHM task's global and local explanations, it offers additive explanations. Likewise, an attention-based LSTM is developed by Ullah et al. [10] for the action recognition in sport events. To enhance the spatial features, they applied convolution block attention mechanism. The refined feature maps are similarly classified into various sports acts by a fully connected dense neural network with a softmax activation function. In contrast to visual data, finding genetic data's anomalies is prime focus of Selicato et al. [11]. For instance, ensemble based method has been provided by them which uses principal component analysis (PCA) and hierarchical clustering to identify aberrant and normal matrices for gene expressions. For the purpose of identifying abnormalities in complicated sceneries, Riaz et al. [12] presented an ensemble of deep models. A human position estimate model that recognises the human joints is comprised in the initial stage. The second phase involves treating the discovered joints and providing for anomaly detection to deeply connected CNN as features. More recently, Zhao et al. [13] recently, reported a method (unsupervised) which utilises sparse reconstruction skills which is a learning from all events vocabulary and online query signals for detecting abnormalities in footages. Sparse coding style with time variation is used by this method. It has remained difficult to acquire the capacity to spot anomalies in a timely manner, which has drawn the attention of various academics. Sparse Combination Learning (SCL), for instance, was used by Lu et al. [4] who examined their method on both local and cloud servers.

Conventional approaches have been surpassed in a number of high-dimensional and non-linear fields with help of deep feature based models, which include video summarization and activity recognition. A method given by Liu et al. [14] used a CNN to encode video frames and a following ConvLSTM to identify abnormal events. Their encoder method encodes motion variation to discover irregularity in surveillance video. Hasan et al. [15] gave an RNN and a convolution autoencoder method. Luo et al. [14] developed an autoencoder for anomaly detection which contains an LSTM (convolutional) model. They also extended on their work by employing an autoencoder-stacked RNN to find abnormalities. Liu et al. [16] combined spatial and temporal detector and put out a method for video anomaly detection. In this model regular data training events has been created from handling of saliency detector and dynamic texture feature collection. Cheng el al. [17] gave a deep autoencoder based on clustering to effectively extract information from normal events. To learn spatial temporal two modules are used, the spatial autoencoder and feature regularity, the last individual video frame is operated in second module. In difference, the temporal autoencoder in second module generates RGB variation between the frames. In addition, anomalies in videos are found using generative models. For example, for detecting abnormalities in videos Sabokroul et al. [18] proposed a generative adversarial network (GANs). In this method normal distribution learn using GANs with generator and discriminator. In recent, researchers proposed weakly supervised approaches to label the video and to detect anomalous act C3D and multiple instance learning (MIL) methods [19], [20] are used. For instance, Sultani et al. [5] came up with the MIL method and weak video labels for detecting anomaly. Normal and anomaly videos separated in two different bags so by training model can learn differences between normal and abnormal event train the model and then, to detect anomaly scores MIL technique is used. A tube extraction method proposed by Landi el al. [21] which uses a regression technique for abnormality detection. Shortly, before passing spatial features with temporal features of optical flow into the regression model, it passes through the inception block. Zhong et al. [22] introduced a supervised method for activity recognition with noisy labels and for detecting weakly supervised anomalies. due to uncertain behaviour of the abnormal events, only in abnormal event labels were noised. In addition, to remove these noisy labels a convolution neural network model is used, and after that an activity classifier was used to classify the activities.

III. EADN: Efficient Deep Learning Model for Anomaly

Key component and structure of EADN framework [25] will be discussed in this section. EADN framework is illustrated in Fig. 1 [25]. In short, classification of anomaly and

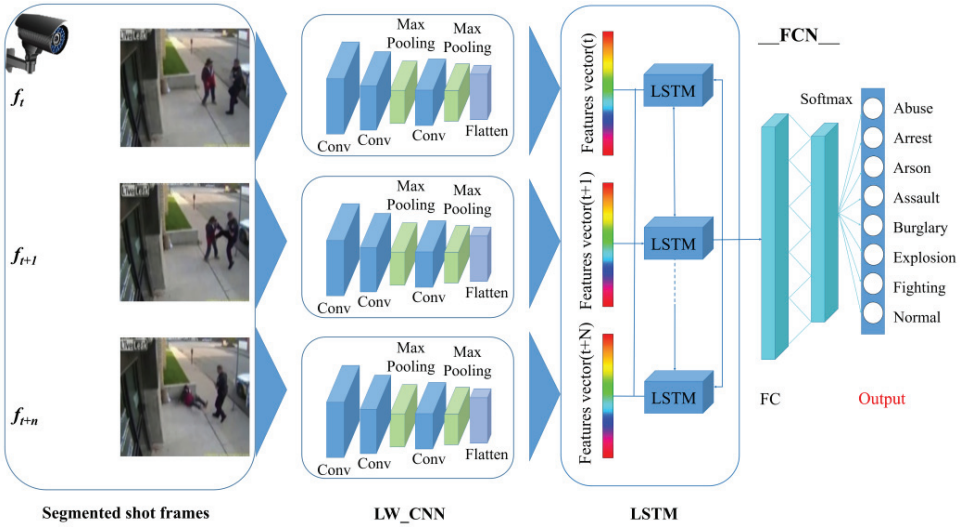

Fig. 1. Surveillance videos anomaly detection using EADN framework.

detection task contains majorly three important parts. First, keyframes extraction. Second, feature extraction. Last one is learning and temporal anomaly classification. In the first part, shot boundary detection algorithm [25] helps to segment key frames. In the second step, these key frames are passed through the Lightweight CNN (LW_{CNN}) architecture to learn spatiotemporal features. After that as the third step, to learn temporal features from an ordered key frames LSTM cells are used. LW_{CNN} took frames that are in sequential manner to detect action and movement. At last, trained LW_{CNN} with LSTM cells network is utilized to classify abnormal action in the video's segmented shot. Model parameters are updated with help of categorical cross entropy loss.

A. Network Architecture

Researchers of EADN have used (to build the network) two time-distributed 2D max-pool (MaxPool2D) layers following time-distribution 2D convolutional layers (Conv2D), to whom kernel sizes, strides and number of filters are mentioned in II [25]. Every time-distributed Conv2D layer with kernel size of 3 x 3 and stride size of 2 x 2, go along with the Rectified Linear Unit activation function. To reduce size of the network it has time distributed MaxPool2D layer which has 2 x 2 strides after the second and third convolution layer. Same-padding approach used to build time-distributed Conv2D layer to stop leaving knowledge at the edge of the input frame. Outcome of this is, same sized feature maps as input. Model architecture contains 64 featured maps in first convolution layer in the starting which followed by second layers which contain same number of feature maps. There are total of 128 feature maps in final convolutional layer. A segmented shot's pre-processed frame is fed into the proposed model as input. First, LW_{CNN} system takes out spatial features and then by giving sequence of spatial features in LSTM captures temporal features. Last time step output of LSTM passes into fully connected dense layer and at last for prediction it passes through softmax activation function [24], as seen in Fig. 1. For spatial features capturing, it passed into LW_{CNN} one by one frame wise. Moreover, the input frame at time t is f_t similarly f_{t+1} and f_{t+n} individually feed into LW_{CNN}, by this every frame converts into series of sequential spatial feature as seen in (1); where F_t represents spatial feature series. But, LSTM input is feed with representation series of spatial features to learn the temporal features H_{t+n}, as indicated in (2). At ongoing time step t, hidden state of LSTM represented by H_t, and similarly H_{t-1} represents previous time step $t-1$ hidden state [24]. Information of previous time step is passed into the current time step as an input. Hidden state H_{t+n} is output of last time step passed as input to the next fully connected dense layer [24]. At last, softmax layer is feeded with fully connected layer's output to predict each class probability, as presented in (3).

$$F_t = LW_{CNN}(f_t, f_{t+1}, f_{t+n}) \tag{1}$$

$$H_{t+n} = LSTM(F_t) \tag{2}$$

$$Prediction: y_j = softmax(H_{t+n}) \tag{3}$$

B. Findings of EADN paper

Authors used UCF-Crime [5], CUHK Avenue [4], and UCSD Pedestrian [23] datasets to analyse the EADN framework. UCSD is considered as a benchmark dataset because it is an extensively used dataset for abnormality detection. Likewise, The UCF-Crime and CUHK Avenue datasets are commonly used to assess anomaly detection algorithms and are also freely accessible. Receiver operating characteristic (ROC) [24] and area under curve (AUC) are used for performance analysis. EADN gained 93% [25] framed-based AUC on UCSDped1 and 97% [25] framed-based AUC on UCSDped2 dataset. In parallel, 97% [25] AUC achieved on Avenue dataset. Finally yet importantly, EADN obtained 98% [25] AUC on UCF-Crime dataset. Compared to recent work, EADN

TABLE II
DESCRIPTION OF LW_{CNN}.

Layer Type	Filters Number	Size	Padding Value	Stride	Activation	Output Shape
Time Distributed Conv2D_1	64	3 x 3	Same	2 x 2	Relu	5, 112, 112, 64
Time Distributed Conv2D_2	64	3 x 3	Same	2 × 2	Relu	5, 56, 56, 64
Time Distributed MaxPooling2D_1	1	2 x 2	-	2 × 2	-	5, 28, 28, 64
Time Distributed Conv2D_3	128	3 x 3	Same	2 × 2	Relu	5, 14, 14, 128
Time Distributed MaxPooling2D_2	1	2 x 2	-	2 × 2	-	5, 7, 7, 128
Time Distributed Flatten_1	-	-	-	-	-	56272

reduces false alarm rates significantly and also proposed fewer parameters (14.14M), smallest model size (53.9MB), faster processing time. However, EADN architecture has space for efficiency improvement and real time accuracy (because of shot boundary detection algo not going to work in real time).

IV. ANOMALY RECOGNITION FROM SURVEILLANCE VIDEOS USING 3D CONVOLUTION NEURAL NETWORK

Authors proposed this method in paper [27] and starts with dividing every video into certain frames. First, all frames were converted into 3D cubes to train the model after pre-processing. A concatenation of sequential frames is passed into the model as input. To not lose channel information, frames are passed in their original state and not in grayscale. The pre-trained model has convolution, pooling, and fully connected layers, but to use this model as a fine-tuned model, researcher also added batch normalization layers. It helps to overcome overfitting. The anomaly recognition technique is shown in Fig. 2 [27]. As per the flow diagram, 3D cubes (pre-processed frames) are passed to 3D ConvNets as an input to take out spatiotemporal features, and as input spatiotemporal is passed into a fully connected layer and then fully connected layer output passed to SoftMax function which calculates the probability of anomaly and helps to train the neural network. The output with the highest likelihood of the anticipated class is provided during the classification step. In this section, data preparation, spatial augmentation, architecture of the model, and findings of the model will be introduced.

A. Data preparation

In the pre-processing step, first sequential frames are extracted from each video. These sequential frames are changed to 170x170 dimensions. Then, all frames' pixel values are scaled from 0 to 1 to ensure that they are all on the same scale. Later on, each class's frames are transformed using spatial augmentation described in the upcoming section. After implementing augmentation method, certain length of 3D cubes prepared from each video to pass temporal and spatial features into fine-tuned deep model.

B. Spatial augmentation

Augmenting an image just a linear change in the spatial domain. For time-series based dataset (video dataset), data augmentation will be more challenging because implementing any augment without temporal knowledge may affect the data's information. Augmentation applied to UCF-Crime dataset in such a way that first n number of frames $V = \{f_1, f_2, f_3, ..., f_n\}$ are obtained from each video then augmented using vertical and horizontal flip approach as illustrated in Fig. 3 [27]. Applying other augmentation (like rotation) on video datasets occasionally results in loss in temoral information and adds noise to the dataset.

C. Fine-tuned 3DConvNet model

Anomaly recognition task fine-tuned on 3DConvNets [26]. The published architecture was trained on UCF-101 activity recognition dataset and showed remarkable results. When performed on the UCF-Crime dataset for anomalous activity recognition, pre-trained model faced difficulties that is overfitting. Therefore, updated weights in transfer learning used to recognise 13 anomaly and 1 normal class in the UCF-Crime dataset. Researchers find out, the direct use of 3DConvNets led to overfitting so they used three batch normalization layers to overcome it and to normalize output. Tab. III [27] shows the architecture of 3DConvNets after fine-tuning.

TABLE III
3D CONVNETS ARCHITECTURE AFTER FINE-TUNING.

Layer	Input	Kernel	Output
Input	16×170×170×3	N/A	16×170×170×3
conv1	16×170×170×64	3x3x3	16×170×170×64
batchNormalization_1	16×170×170×64	-	16×170×170×64
pool1	16×170×170×64	2x2x2	16×85×85×64
conv2	16×85×85×64	3x3x3	16×85×85×128
pool2	8×85×85×128	2x2x2	8×43×43×128
conv3a	8×43×43×128	3x3x3	8×43×43×256
conv3b	8×43×43×256	3x3x3	8×43×43×256
pool3	8×43×43×256	2x2x2	4×22×22×256
conv4a	4×22×22×256	3x3x3	4×22×22×512
conv4b	4×22×22×512	3x3x3	4×22×22×512
pool4	4×22×22×512	2x2x2	2×11×11×512
conv5a	2×11×11×512	3x3x3	2×11×11×512
conv5b	2×11×11×512	3x3x3	2×11×11×512
pool5	2×13×13×512	2x2x2	1×6×6×512
batchNormalization_2	1×6×6×512	-	1×6×6×512
fc6	(None,18,432)	-	(None,4096)
batchNormalization_3	(None,4096)	-	(None,4096)
fc7	(None,4096)	-	(None,4096)
fc9	(None,4096)	-	(None,14)

D. Findings of 3D Convolution Neural Network paper

Researchers' method successfully learned to extract 3D features by giving augmented frame-level information of the UCF-Crime dataset to fine-tuned pre-trained 3DConvNets. Additionally, the suggested technique beats state-of-the-art approaches substantially in terms of accuracy on the recognition of aberrant activity, and the learned features are sufficiently

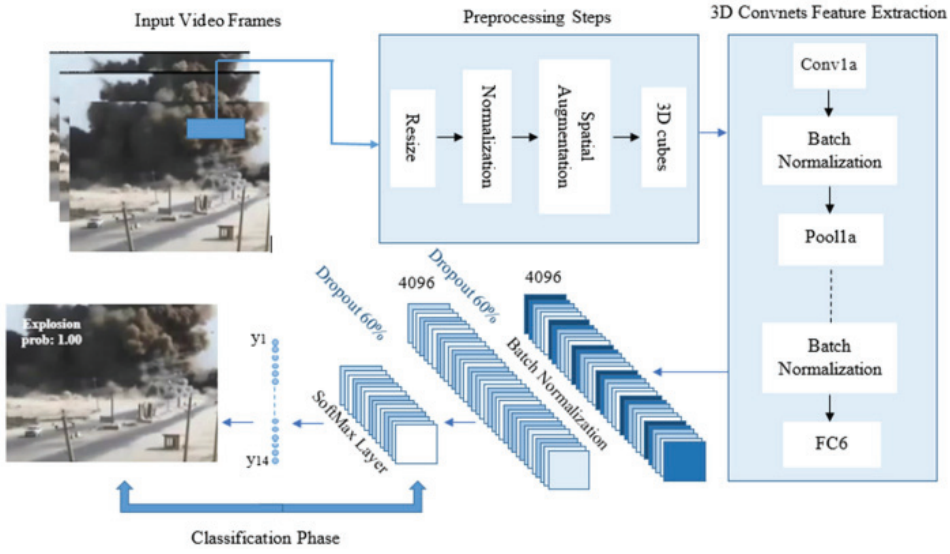

Fig. 2. Flow diagram of 3DConvNets.

Fig. 3. Vertical augmentation (Bottom) of UCF Crime dataset frames (Top).

compact to having an AUC of 82% [27] and achieved 0.45 [27] F-measure. This method did not achieve its full potential because of intra-class similarity between assault and fighting classes and arrest and shooting classes etc.

V. Anomaly Detection in Surveillance Video Based on Bidirectional Prediction

Before this approach, models were trained on past k frames to predict only one succeeding frame I_t at current time t. These techniques became common to people and have also given good results. But, the relation between the upcoming target frames and sequential frames cannot be determined by these techniques due to temporal information loss. So, a bidirectional prediction network is proposed in paper [31] to effectively utilise the temporal information in videos.

A. Network structure

Firstly, U-Net [30] opted as backbone network for both forward predictive subnetwork G_F and backward predictive subnetwork G_B as shown in Fig. 4 [31]. The setting of hyper-parameters is detailed in P-GAN [5] and both subnetworks G_F and G_B are identical in structure. The current target frames are denoted as I_t, while the preceding k frames are denoted as $I_{t-k}, ..., I_{t-1}$. Then (Batch, Frames, Height, Width, Channel) dimensional spatiotemporal tensors of each video of k frames are prepared as inputs and feed these tensors to the forward prediction subnetwork G_F. The produced frame that corresponds to G_F is labeled as $\hat{I}_t^F$. Similarly, the next followed k frames are $I_{t+1}, ..., I_{t+k}$ and their spatiotemporal tensors are prepared as inputs to the backward prediction subnetwork G_B. The corresponding generated frame is denoted as $\hat{I}_t^B$.

B. Loss function

Two symmetrical prediction subnetworks G_F and G_B make up the bidirectional prediction network that appears to have been trained independently but they are connected to one another because they share a same target frame. Hence, it is important to consider the similarity between two prediction frames and target frame to build the loss function.

First, as shown in Fig. 4, the forward mean square error MSE_F for a single prediction network is defined using the target frame I (ignore the time index t for simplicity) and the forward prediction frame $\hat{I}^F$, as given in (4). Similarly, definition of backward mean square error MSE_B has same meaning as shown in (5):

$$MSE_F = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I^F(i,j) - I(i,j)]^2 \tag{4}$$

$$MSE_B = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I^B(i,j) - I(i,j)]^2 \tag{5}$$

Predicting abnormal video frames $\hat{I}^F$ and $\hat{I}^B$ from G_F and G_B have worse consistency than those normal video. Mean square error between backward and forward prediction

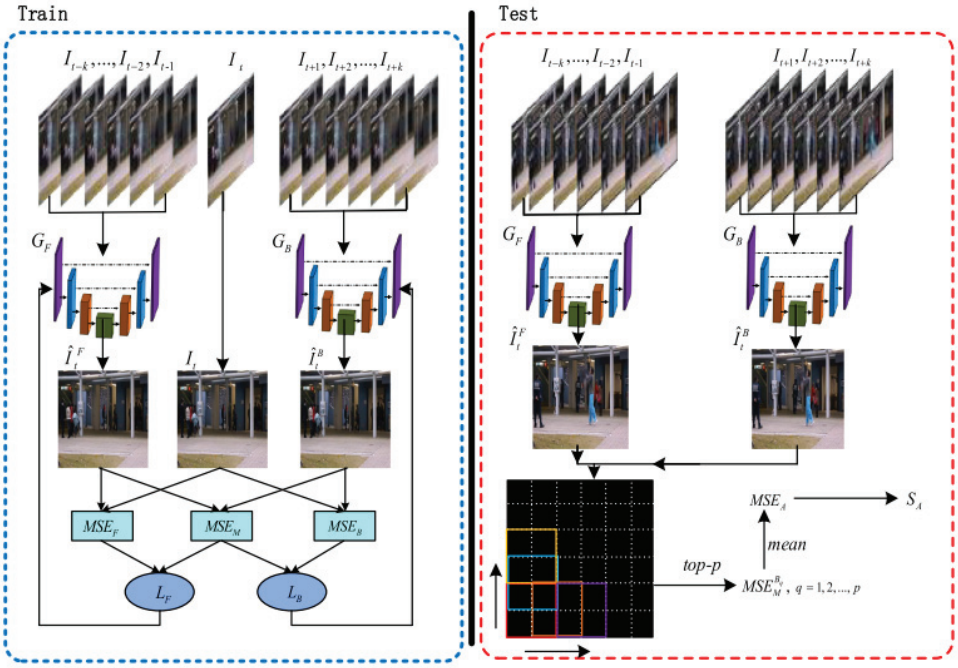

Fig. 4. The architecture of bidirectional prediction network.

frame MSE_M used as loss function of bidirectional prediction model, as shown in (6):

$$MSE_M = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I^F(i,j) - I^B(i,j)]^2 \tag{6}$$

Finally, as demonstrated in (7) and (8), L_F is the joint loss of forward prediction subnetwork G_F and L_B is the joint loss of backward prediction subnetwork G_B.

$$L_F = MSE_F + \lambda_F MSE_M \tag{7}$$

$$L_B = MSE_B + \lambda_B MSE_M \tag{8}$$

Where λ_F work as a weight for loss function MSE_M of the forward prediction subnetwork G_F. Similarly, λ_B has the same definition but for the G_B. Since the relationship between the two subnetworks is defined through the loss function, a better generalized model is expected.

C. Abnormality estimation

In this approach, the loss between the target frame and the predicted frames from bidirectional prediction network is used to estimate the score of abnormality. PSNR is used to calculate the anomaly score of the current frame I_t between $\hat{I}_t^F$ and backward prediction $\hat{I}_t^B$ in the testing process. (9) displays the formula used to calculate the PSNR score.

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

Where I_{max} is the highest pixel value possible for the current frame, which is typically 255. Higher the PSNR score calculated, lower the anomaly score going to be. Another approach for anomaly score estimation is the blocking strategy [31].

D. Findings of Bidirectional Prediction paper

To train and test the approach, researchers used UCSD [29] and CUHK Avenue [28] datasets, which are popular in the study of video anomaly detection. The bidirectional prediction network got 87.24% [31] 96.54% [31] AUC on UCSD Ped1 and UCSD Ped2 respectively. And achieved 87.39% [31] AUC on Avenue dataset. Apart from that, U-Net was not developed to extract temporal features so as a backbone of bidirectional approach to extract temporal features a better network is required. However, GAN and autoencoder approaches have one good thing which is more dataset can be prepared because they are trained on only normal data, (in testing time abnormal events are required too) and abnormal event occurrence is not frequent as a normal event.

VI. Conclusion

Researchers have shown many methods to detect anomaly from surveillance videos in the field of deep learning. This survey has systematically reviewed and summarized three of all available methods for deep neural networks in computer vision. Researchers will have a good understanding of a GAN approach, a fine-tuned approach, a key frame extraction with not too deep model approach and researchers will also understand the augmentation importance and how to tackle overfitting. This work will encourage a variety of potential developers to improve anomaly detection approaches and its results.

Despite all this, as per the future scope, a better temporal features extraction method is required. Model takes frames as input and do the needed maths to predict whether given frames have anomaly or not. But to identify which frame is dependent on another frame to classify video and which frame is more contributing in classification for that a self-attention mechanism and vision transformer could help. Another possibility is to take anomaly recognition problem as a video caption problem and then based on text prediction a text classification model can be used to classify anomaly category. During training time, video description as label will help the model to extract specific features from corresponding frames.

REFERENCES

[1] Piza, E.L.; Welsh, B.C.; Farrington, D.P.; Thomas, A.L. CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis. Criminol. Public Policy 2019, 18, 135–159.

[2] Cheng, K.W.; Chen, Y.T.; Fang, W.H. An efficient subsequence search for video anomaly detection and localization. Multimed. Tools Appl. 2016, 75, 15101–15122.

[3] He, C.; Shao, J.; Sun, J. An anomaly-introduced learning method for abnormal event detection. Multimed. Tools Appl. 2018, 77, 29573–29588.

[4] Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2720–2727.

[5] Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488.

[6] Huo, J.; Gao, Y.; Yang, W.; Yin, H. Abnormal event detection via multi-instance dictionary learning. In International Conference on Intelligent Data Engineering and Automated Learning; Springer: Berlin/Heidelberg, Germany, 2012; pp. 76–83.

[7] Zhang, D.; Gatica-Perez, D.; Bengio, S.; McCowan, I. Semi-supervised adapted hmms for unusual event detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 611–618.

[8] Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 935–942.

[9] Nor, A.K.M.; Pedapati, S.R.; Muhammad, M.; Leiva, V. Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data. Mathematics 2022, 10, 554.

[10] Ullah, M.; Mudassar Yamin, M.; Mohammed, A.; Daud Khan, S.; Ullah, H.; Alaya Cheikh, F. Attention-based LSTM network for action recognition in sports. Electron. Imaging 2021, 2021, 302-1–302-6.

[11] Selicato, L.; Esposito, F.; Gargano, G.; Vegliante, M.C.; Opinto, G.; Zaccaria, G.M.; Ciavarella, S.; Guarini, A.; Del Buono, N. A new ensemble method for detecting anomalies in gene expression matrices. Mathematics 2021, 9, 882.

[12] Riaz, H.; Uzair, M.; Ullah, H.; Ullah, M. Anomalous Human Action Detection Using a Cascade of Deep Learning Models. In Proceedings of the 2021 9th European Workshop on Visual Information Processing (EUVIP), Paris, France, 23–25 June 2021; pp. 1–5.

[13] Zhao, B.; Fei-Fei, L.; Xing, E.P. Online detection of unusual events in videos via dynamic sparse coding. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3313–3320.

[14] Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional lstm for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 439–444.

[15] Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.

[16] Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection–a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.

[17] Chang, Y.; Tu, Z.; Xie, W.; Yuan, J. Clustering driven deep autoencoder for video anomaly detection. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 329–345.

[18] Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.

[19] Ullah, W.; Ullah, A.; Hussain, T.; Khan, Z.A.; Baik, S.W. An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos. Sensors 2021, 21, 2811.

[20] Tomar, D.; Agarwal, S. Multiple instance learning based on twin support vector machine. In Advances in Computer and Computational Sciences; Springer: Berlin/Heidelberg, Germany, 2017; pp. 497–507.

[21] Landi, F.; Snoek, C.G.; Cucchiara, R. Anomaly locality in video surveillance. arXiv 2019, arXiv:1901.10364.

[22] Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1237–1246.

[23] Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.

[24] Li, W.; Mahadevan, V.; Vasconcelos, N. Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 36, 18–32.

[25] Ul Amin S, Ullah M, Sajjad M, Cheikh FA, Hijji M, Hijji A, Muhammad K. EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos. Mathematics. 5 May 2022; pp. 1555.

[26] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. in Proceedings of the IEEE international conference on computer vision, ICCV: 4489 – 4497.

[27] Maqsood R, Bajwa UI, Saleem G, Raza RH, Anwar MW. Anomaly recognition from surveillance videos using 3D convolution neural network. Multimedia Tools and Applications. May 2021; pp. 18693-716.

[28] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 fps in matlab, Proceedings of the IEEE International Conference on Computer Vision (2013) 2720–2727.

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

[30] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351 (2015) 234–241.

[31] Chen D, Wang P, Yue L, Zhang Y, Jia T. Anomaly detection in surveillance video based on bidirectional prediction. Image and Vision Computing. 1 Jun 2020; pp. 103915.

[32] Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.

[33] Chong, Y.S.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In International Symposium on Neural Networks; Springer: Berlin/Heidelberg, Germany, 2017; pp. 189–196.

[34] 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. 2017, pp. 117–127.

[35] D. Koshti, S. Kamoji, N. Kalnad, S. Sreekumar and S. Bhujbal, "Video Anomaly Detection using Inflated 3D Convolution Network," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 729-733.