Received: August 6, 2023. Revised: September 8, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Abnormal Crowd Behaviour Detection in Surveillance Videos Using Spatiotemporal Inter-Fused Autoencoder

Dhivya Karunya Sampath[1]* **Krishna Kumar[2]**

[1]*Department of Electronics & Communication Engineering, S.E.A. College of Engineering & Technology, Bengaluru, Visvesvaraya Technological University, Belagavi - 590018, India*
[2]*Department of Electronics & Communication Engineering, Gopalan College of Engineering and Management, Bengaluru, Visvesvaraya Technological University, Belagavi - 590018, India*
* Corresponding author's Email: dhivyaskarunya@gmail.com

Abstract: In crowded environments, the importance of automatic video surveillance cannot be overstated. It plays a vital role in detecting unusual incidents and averting accidents, particularly in areas teeming with pedestrians. Surveillance frameworks demonstrate their inherent strength when deployed in real-world scenarios. In this paper, an efficient algorithm that can identify abnormalities in videos has been proposed as a solution to the problems of high computing power and variation between negative and positive samples. To address the issue of fewer negative samples, the algorithm employs the spatiotemporal inter-fused autoencoder in a method of unsupervised learning to locate and extract the samples of negative from the dataset. A spatial-temporal convolutional neural network (CNN) is built using this methodology, and it has a basic structure and requires minimal computing power. To develop the model of detection, the spatiotemporal CNN is trained using the method of supervised training with negative and positive data. The UMN and UCSD datasets are used as benchmark datasets for this method. The outcomes of the experiment demonstrate that the proposed method is more accurate than the existing algorithms at both frame and pixel levels and it can locate anomalous behaviors in real-time. The proposed method achieves better accuracy of 99.76%, 99.92%, 99.15%, 98.45%, and 94.67% in UCSD ped1, UCSD ped2, and UMN scene 1, 2, 3 datasets compared to the hybrid CNN and RF classifiers, MCMS-BCN attention+densenet121/Efficientnetv2, gradient motion descriptor (PGD) and enhanced entropy classifier, and attention mechanism. Additionally, the proposed method achieves better AUC of 99.96%, 99.83%, 99.97%, 90.15%, and 99.72% in the datasets of UCSD ped1, UCSD ped2, and UMN scenes 1, 2, 3 compared to the low-rank and compact coefficient dictionary learning (LRCCDL), and hybrid CNN and RF classifiers.

Keywords: Abnormal behavior detection, Crowded scenes, Real-time, Spatiotemporal convolutional neural network, Surveillance videos.

1. Introduction

Video surveillance is a major concern in the design, long-term viability of urban areas and modern industrial, and operation. The effectiveness, security, and safety of the area, infrastructure, people, activities, and operations are all improved through video monitoring. The deployment of closed-circuit television (CCTV) camera systems has increased exponentially as a result of horizontal and vertical asset expansion and area utilization in both urban and industrial areas. However, for human observers would be ridiculous and impossible to precisely examine and evaluate every video stream [1]. Many computers vision-based initiatives have recently been proposed for tracking, detection, and identification of certain objects in videos of surveillance such as humans, animals, and cars to examine behavior. Contrarily, the analysis of crowd behavior is a new research field in computer vision with significant applications, including the automatic identification of panic and avoidance behaviors, natural disasters, violent occurrences, and chaotic, or riots crowd behavior [2]. Sharp surveillance structures' performance has changed abnormally is one of the most common and long-lasting errors in identifying PC vision and identifying changes in an image, various fundamental methods for using intelligent surveillance frameworks have been presented [3]. Two categories of typical diverging motion patterns were introduced: circular and straight motions and crowd divergence occurs when a crowd departs from a circular or straight walking path [4]. The primary focus of intrusion research detection is anomaly detection, which establishes the systems or users' normal behavior patterns and detects intrusions by comparing and matching those patterns with those of the monitoring system or the actual user [5]. The two fundamental problems that dominate the field of video anomaly detection: initially, positive and negative samples are uneven since normal events outperform abnormal ones by a significant margin. Secondly, video clips cannot list every type of abnormal event, whereas existing detection of anomaly databases only focus on a few key scenarios [6].

Anomaly detection and localization aim to identify abnormal activities in crowded surveillance videos, challenging conventional object-based methods due to high-density people and uncertain motions [7]. Traditional methods for detecting anomalous behavior include dynamic Bayesian networks (DBNs), cluster models, PTMs, and sparse methods. Hidden Markov models are the most commonly employed using probability to measure a behavior's match to a test sequence [8]. To provide security and safety as well as real-time supply and management of demand for public transportation, crowd behavior analysis can be a useful technique in the deployment of intelligent transportation systems [9]. The constraints established on humans can be overcome by the systems of automated computer vision-based surveillance [10]. Additionally, a large number of people travel daily through train stations, bus stations, metro stations, and airports. Controlling crowds and managing public meetings safely is crucial in these high-density crowd circumstances [11]. The common method employed all over the world is having a control room with various screens showing CCTV cameras live streams for crowd anomalies detection. These activities are supervised and recorded by human beings 24/7 [12]. The properties of both public and private assets can be protected and saved with the use of anomaly detection. Security-related uses for surveillance cameras involve identifying abnormal behavior or abnormalities in both public and private spaces [13]. To identify abnormal crowd detection, following are the primary contributions are summarized below:

- A deep learning method that identifies spatiotemporal patterns of typical activity from a stream of surveillance videos for abnormal detection and localization.
- The fuzziness accumulation of active learning results in the process of continuous learning for dynamic adaptation to changing behaviors of unknown/new normalities in the stream of surveillance video.
- The spatiotemporal CNN is trained using the method of supervised training with negative and positive data to create the model of detection. The dataset of UCSD pedestrians (Ped1 and Ped2) and the UMN are utilized as benchmark video surveillance datasets.

The rest of this research is organized as follows. Section 2 discusses the literature survey. Section 3 describes the proposed methodology. Section 4 describes the results. Section 5 discusses the conclusion.

2. Literature survey

Li [14] implemented a new algorithm for abnormal behavior identification in a crowded environment based on a new low-rank and compact coefficient dictionary learning (LRCCDL). A feature space was created based on the binarization of the background and reduction of surveillance videos by using the histogram of the maximal optical flow projection (HMOFP) foreground feature from a typical training set of frames. The optimization of joint achieves two outcomes: a learned low-rank dictionary and a vector of compact coefficient reconstruction training data that are centered around a mean value. However, the LRCCDL method's results were not superior because the amount of training data required by the LRCCDL method was significantly less than those of other methods, particularly for the datasets of UCSD and CUHK Avenue.

Fan [15] presented an efficient algorithm that identifies anomalies in videos and the model employs the spatiotemporal autoencoder to locate and remove the samples of negative that include behaviors of abnormal in the dataset using a method of unsupervised learning to address the issue of less negative samples. The spatiotemporal CNN was trained using the supervised training technique with negative and positive data to create the detection model. The spatiotemporal CNN-based model has a simple structure and achieves minimal computational resources. However, the presented method needs a deep learning method as a mainstream to improve abnormal behavior detection.

Tarik Alafif [16] implemented a HAJJv2, a large-scale crowd abnormal behavior Hajj dataset that has been labeled and annotated. Second, two techniques of convolutional neural network (CNN) and random forest (RF) were introduced for detecting and recognizing spatiotemporal abnormal behaviors in small and large-scale crowd videos. On two benchmark small-scale crowd datasets generally available to the public, UMN and UCSD, the implemented method achieves an average area under the curves (AUCs). However, the HAJJV2 dataset struggles with large-scale crowds due to distant camera view, along with partial, heavy, and full occlusions.

E. M. C. L. Ekanayake [17] presented a novel MCMS-BCNN attention network with a densenet121/ Efficientv2 architecture for identifying a variety of basic movements in public areas, particularly rapid motion changes, human flock movement, and panic occurrences in a variety of indoor and outdoor environments. With pre-processed morphological video frames, the significant spatial characteristics were retrieved from a bilinear and a multicolumn multistage CNN. In every scenario, the presented method achieves better performance consistency. However, to obtain superior video activity detection, the dense activity detection model needs to be improved with an autoencoder-based technique.

Patel [18] presented a multi-object tracking algorithm to generate reliable object tracks for examining crowd behavior in public places by reducing the occlusion of short-term objects, identity switches, and detection errors. The detection of the bounding box and velocity estimation of a linear object utilizing the Kalman filter were used to track the object's frame by frame and the missing detections and temporary object occlusion were handled by maintaining the predicted tracks in existence. The method was demonstrated in traffic environments and movement scenarios of pedestrians and achieves excellent accuracy for event detection. However, specific situations need enhanced separation criteria that require an evaluation of actual physical distance.

Faisal Abdullah & Ahmad Jalal [19] introduced a robust semantic segmentation method for tracking, estimating, and anomaly detection in pedestrian crowds. The weighted average technique, human motion analysis, and the attraction force model were combined to count and track a pedestrian while excluding non-human and non-pedestrian objects from the image. Over UMN and MED datasets, the introduced method detects anomalies accurately. However, the tracking system accuracy decreases slightly in dense crowds, due to full occlusions occurring during test videos.

Varghese [20] implemented a fuzzy deep learning method to identify nine distinct behaviors that represent nearly every kind of crowd behavior by examining the psychological and cognitive factors that influence the behavior of humans. The method combines two methods of cognitive deep learning with a psychological fuzzy computational framework that uses the five-factor personality model of OCEAN, OCC emotion theory, and visual attention to identify the behaviors of the crowd. The method accurately predicts crowd behavior for all crowd types and was independent of local feature descriptors and patterns of mobility. However, the suggested method requires data from numerous modalities to identify groups inside the crowds and detect suspicious activities in a crowd.

Faisal Abdullah [21] presented a particle gradient motion descriptor (PGD) and enhanced entropy classifier for a multi-person tracking system and human crowd behavior detection. Noise removal, contract correction, and edge detection were executed as pre-processing stages on the collected video frames before the detection of human/non-human was carried out utilizing multi-level thresholding and the operations of morphology. The presented method performs effectively in the detection of human crowd behavior. However, the tracking system performance marginally decreases with increasing scene numbers due to full occlusions in test videos.

Qianqian Zhang [22] introduced an attention mechanism based on a dynamic prototype unit (DPU) to improve the feature representation ability of the video anomaly detection model. To decrease the number of parameters and improves the model's accuracy, the depth separable convolution technique is developed. The attention mechanism captures all the spatial information in the video sequences which enables them to obtain better video frame features and higher accuracy. However, the attention mechanism eliminates too many distinguishing factors at each phase to distinguish anomalous detection.

There are some limitations with the existing methods that are mentioned above such as the attention mechanism eliminates too many distinguishing factors at each phase to distinguish anomalous detection. To obtain superior video activity detection, the dense activity detection model needs to be improved with an autoencoder-based technique. To overcome these issues, a

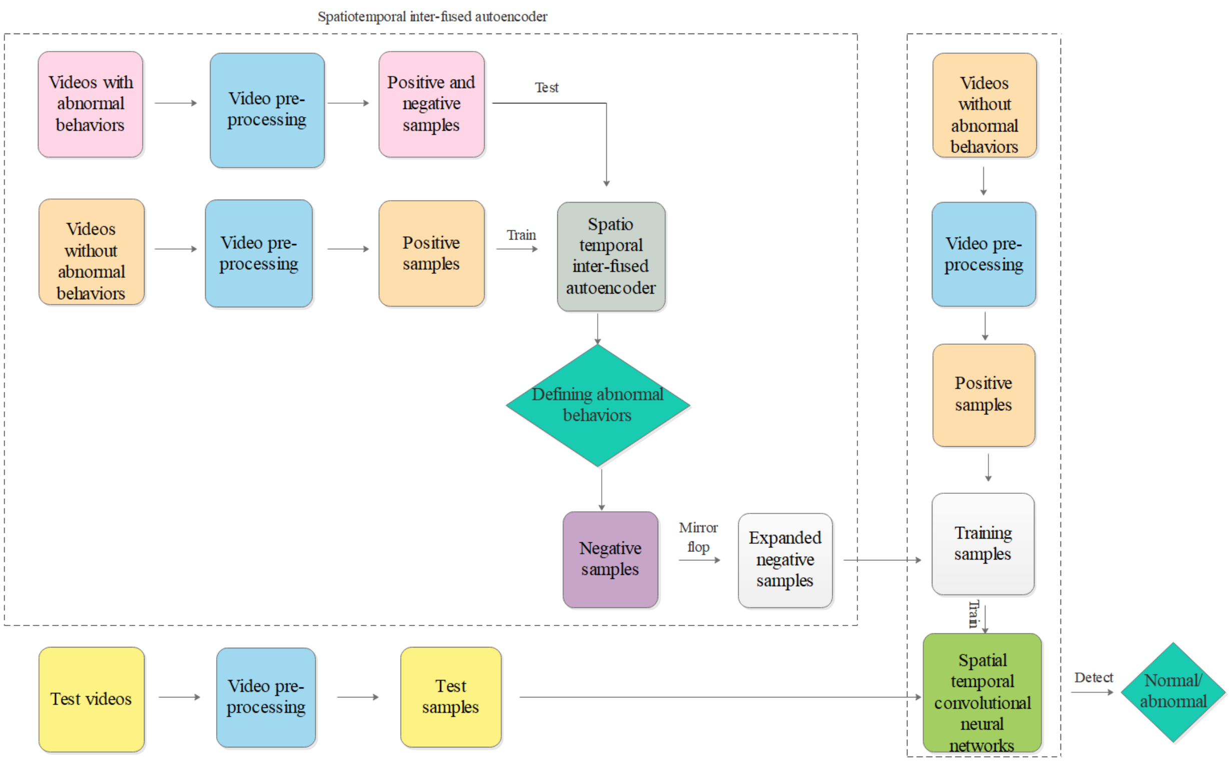

Spatiotemporal inter-fused autoencoder
Videos with abnormal behaviors
Video pre-processing
Positive and negative samples
Test
Videos without abnormal behaviors
Video pre-processing
Positive samples
Train
Spatio temporal inter-fused autoencoder
Defining abnormal behaviors
Negative samples
Mirror flop
Expanded negative samples
Videos without abnormal behaviors
Video pre-processing
Positive samples
Training samples
Train
Spatial temporal convolutional neural networks
Detect
Normal/ abnormal
Test videos
Video pre-processing
Test samples

Figure. 1 Block diagram of the proposed method

spatiotemporal inter-fused autoencoder is proposed for abnormal crowd behavior detection.

3. Proposed methodology

The abnormal crowd behavior detection in surveillance videos was executed by utilizing the representational strength of deep learning. The video block with normal behaviors provides the training data for the spatiotemporal autoencoder. The network is used to define and retrieve the abnormal video blocks after training. The abnormal and normal video blocks are employed to train the spatiotemporal CNN to create the model of detection. The overview of Abnormal crowd behavior detection in surveillance videos block diagram is shown in Fig. 1.

3.1 Datasets

3.1.1. UCSD dataset

The UCSD dataset is used to test the algorithm's performance using the detection of anomalies at the local scale, and it involves object localization tests. In Table 1. the UCSD dataset is split into two parts: ped1 and ped2. There are 34 videos in ped1 with no behavior of abnormal and 36 videos with any behavior of abnormal. Each video has 200 frames of the same size 238 x 158 and ten videos with real-world information are also included. In ped2 there are 12 videos of abnormal behavior and 16 videos without them and there are varying numbers of frames in each video. The pixel size is 360 x 240 and additionally, it includes 12 videos with real-world information. Both ped1 and ped2 involve abnormal behaviors like bicycles, cars, carts, skateboarders, people walking in the grass, and wheelchairs. Only a portion of the videos in ped1 is employed for training and in ped2 the model of detection will be applicable.

Table 1. UCSD dataset

Dataset	Normal behavior	Abnormal behavior	Frames	Pixel size
UCSD ped1	34 videos	36 videos	200	238 x 158
UCSD ped2	16 videos	12 videos	250	360 x 240

4. UMN Dataset

The crowd dataset of UMN includes both abnormal and normal behavior that has been observed at outdoor and indoor scenes. Each video begins with a period of ordinary crowd behavior and finishes with abnormal crowd behavior. In Table 2. three distinct scenarios make up the UMN dataset and consist of 1453, 4144, and 2142 frames. The pixels are 320 x 240 and the crowd dispersion is regarded as abnormal behavior in the dataset. The dataset's experimental outcomes are only analyzed at the level of the frame because the dataset lacks ground facts at

Table 2. UMN dataset

Dataset	Frames	Pixel Size
UMN scene1	1453	320 x 240
UMN scene 2	4144	320 x 240
UMN scene 3	2142	320 x 240

the level of the pixel.

4.1 Video pre-processing

The purpose of video pre-processing is to improve the data's fit to the model and, to a certain expand lower the computing expense of the method. Initially, each frame in the video is the same size at 180 x 120. Due to the pixel's values being unified, the created network continues to function when the video's scale changes. Second, the size of the video frames is reduced by converting the RGB-formatted video frames to greyscale pictures and there are two main causes. First, the target's spatial and temporal data is more crucial than its color qualities and grayscale pictures are used to extract the information mentioned above. The second reason is that lowering the image dimension makes the model structure simpler and maintains real-time performance. The frame of the video is then normalized so that the value of each pixel lies between [0,1] and the average video image is subtracted from it. The video sequence is then split into 15 x 15x 10 video blocks, where 15 stands for the dimension of spatial and 10 for the temporal dimension. The 15 x 15-pixel region holds the smallest amount of abnormal behavior. After pre-processing, the blocks of video are used as input in the testing and training stages.

4.2 Abnormal behavior detection and extraction

An unsupervised learning method is frequently utilized since abnormal behaviors are rare and difficult to identify in everyday life. Abnormal behavior is defined as deviation from normal behavior. Moreover, this technique eliminates the information of abnormal behavior in the dataset, causing data resource waste. Additionally, if a model is discovered to be inactive to the behavior of abnormal, it cannot be modified to better itself. The robustness of the method has been enhanced by employing the learning of supervised at the same time learning both abnormal and normal behavior. However, this algorithm needs manual classification of abnormal and normal behaviors, wasting a significant number of human resources. Therefore, it is essential to automatically identify abnormal behaviors.

4.2.1. Inter-fused autoencoder: spatiotemporal feature extractor

Autoencoders are built into deep neural networks to automatically learn data representations without supervision. An autoencoder, which is frequently used for dimensionality reduction, aims to learn how to analyze the representation of data. Recently, autoencoders have improved their ability to detect anomalies. The main benefit of utilizing an autoencoder for the proposed method is that can rebuild the data of input video with fewer errors. The method will exhibit high error when the behavior of abnormal occurs once it has been trained on normal video data. Encoder, decoder, and representation of latent space are the three parts of an autoencoder. To reconstruct the input X_i with the least amount of error, the transformation function of two nonlinear φ to $\emptyset$ operate on input vector X_i and code of latent h.

$$\varphi : X \rightarrow H \tag{1}$$

$$\emptyset : H \rightarrow X \tag{2}$$

By combining Eqs. (1), (2), we get (3)

$$\varphi, \emptyset = arg\, min \|X_i - (\varphi . \emptyset) X_i\|^2 \tag{3}$$

Where input $X_i \in R^n = X$
Code of latent $h \in R^m = H; \forall\, m < n$

The inter-fused autoencoder combines the layers of convolutional and LSTM. In Fig. 2, the model receives an input set of ten frames with various strides. Conv-LSTM learns the motion properties between frames after the layers of convolutional remove the most prominent and optimistic features of spatial. Following the encoding phase, the decoder employs backpropagation to reduce the error between the input and reconstructed frames to remove the same feature as the encoder.

Though both deconvolution and Conv-LSTM layers are combined in the decoder, convolution and Conv-LSTM make up the encoder. IFA's spatial encoder-decoder is made up of six batch levels of normalization: three deconvolutional layers, and three layers of convolutions based on the amount of input data. To improve the extraction of temporal characteristics of spatial representation, the temporal encoder-decoder includes five layers of Conv-LSTM. The first layer of convolution includes 256 filters with a size of 7 x7 stride 4 and offers 256 map features with a size of 64 x 64 pixels. Normalization of the batch has been utilized to speed up and improve the stability of the model's learning process. The

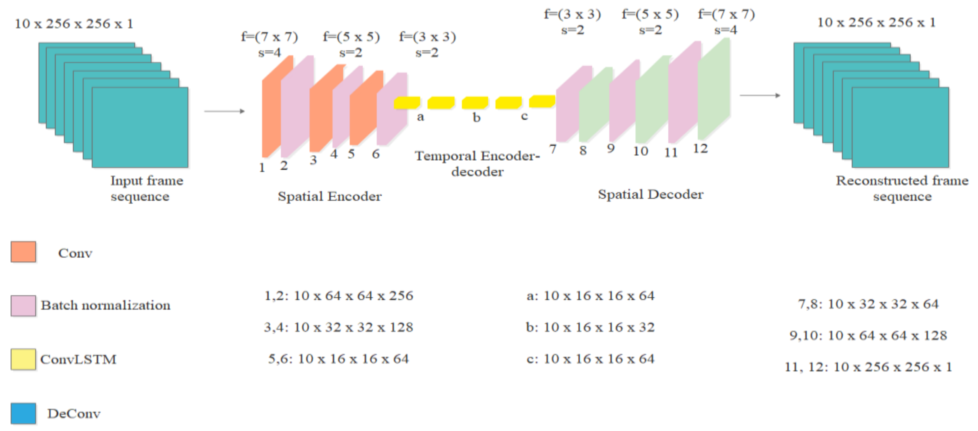


Figure. 2 Inter-fused autoencoder

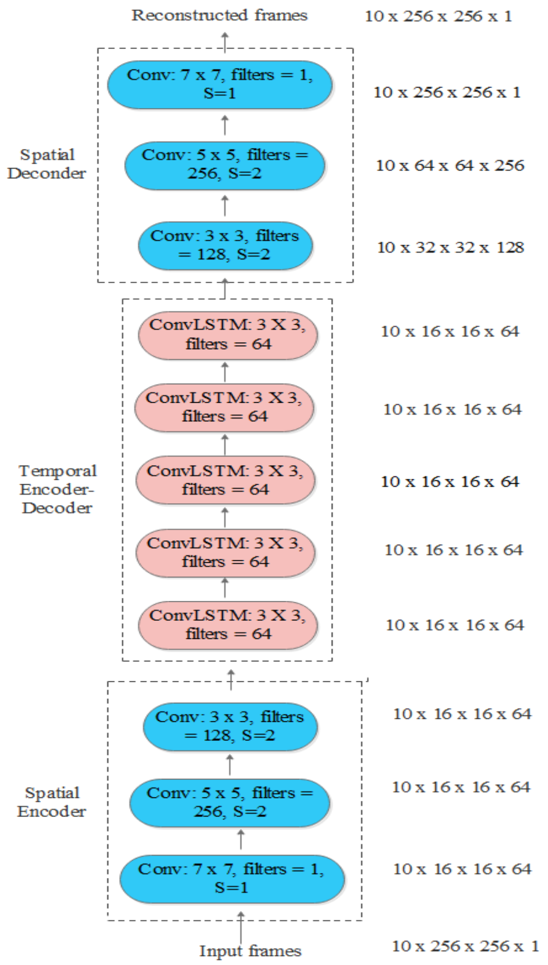

Reconstructed frames
10 x 256 x 256 x 1
Spatial Deconder
Conv: 7 x 7, filters = 1, S=1
10 x 256 x 256 x 1
Conv: 5 x 5, filters = 256, S=2
10 x 64 x 64 x 256
Conv: 3 x 3, filters = 128, S=2
10 x 32 x 32 x 128
Temporal Encoder-Decoder
ConvLSTM: 3 X 3, filters = 64
10 x 16 x 16 x 64
ConvLSTM: 3 X 3, filters = 64
10 x 16 x 16 x 64
ConvLSTM: 3 X 3, filters = 64
10 x 16 x 16 x 64
ConvLSTM: 3 X 3, filters = 64
10 x 16 x 16 x 64
ConvLSTM: 3 X 3, filters = 64
10 x 16 x 16 x 64
Spatial Encoder
Conv: 3 x 3, filters = 128, S=2
10 x 16 x 16 x 64
Conv: 5 x 5, filters = 256, S=2
10 x 16 x 16 x 64
Conv: 7 x 7, filters = 1, S=1
10 x 16 x 16 x 64
Input frames
10 x 256 x 256 x 1

Figure. 3 Spatial view of IFA

second and third layers of convolution produce feature maps that are 32 x 32 and 16 x 16 pixels in size by using 128 and 64 filters of size 5 x 5 and 3 x 3 with stride 2. The four 64 and one 32-bit 3 x 3 kernels that make up the temporal encoder-decoder yield 16 x 16 map representation. After that, as shown in Fig. 3, the decoder of spatial rebuilds the input by batch normalization and deconvolution of each layer in the opposite order of the same size but with a distinct stride and feature map.

4.3 Detection models based on spatiotemporal CNN

In this section, a spatiotemporal CNN is built, and the final model for the detection of abnormality is created by employing a technique of supervised learning to learn the labelled negative and positive data. This model is strong because it fully uses

Received: August 6, 2023. Revised: September 8, 2023.

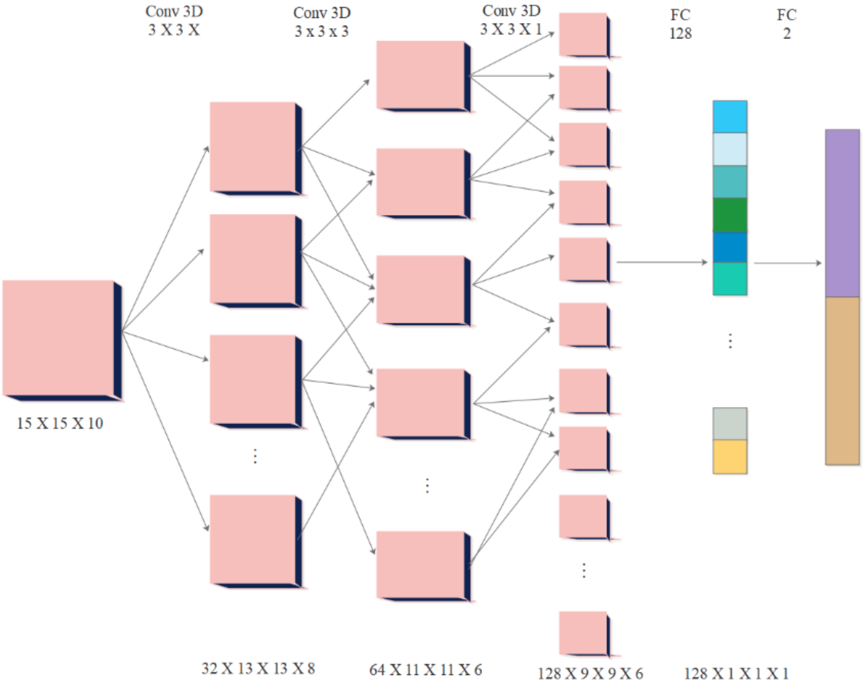


Figure. 4 Structure of a spatiotemporal CNN

information regarding both the behaviors of abnormal and normal. The model network structure is very simple, allowing for the quick identification and location of abnormal behavior.

4.3.1. Spatiotemporal convolution

The feature map and convolution kernel in two-dimensional convolution are both two-dimensional. As a result, only the image's features of spatial dimension are retrieved during the convolution process. The convolution of spatiotemporal is also known as the convolution of three-dimension, which corresponds to the convolution of two-dimension. Three-dimensional convolution can simultaneously remove the spatial and video temporal properties sequence since both the kernel of convolution and feature map are three-dimensional. The three-dimension convolution ($G[i,j,k]$) process using video blocks and kernels of convolution is defined by Eq. (4), where i,j,k denotes length, width, and convolution kernel's temporal dimension length. The size of the video block is represented by $x \times y \times t$ which equals $15 \times 15 \times 10$.

$$G[i,j,k] = \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} W^{ijk} V^{(x+i)(y+j)(t+k)} \tag{4}$$

4.3.2. Structure of spatiotemporal CNN

Spatiotemporal CNN is built as the detection model based on spatiotemporal convolution illustrated in Fig. 4.

A video block of 15 x 15 x 10 is the network's input, while the output divides the input into two groups. The network does not require the usage of a pooling layer because the size of the input is modest. Convolutional layers make up the first three network layers, which are utilized to extract the video block features. The input is convolved in the first layer employing 32 (3 x 3 x 3) kernels convolution to produce 32 (13 x 13 x 8) blocks of the feature. This layer's dropout rate is 0.25. The 64 (3 x 3 x 3) convolution kernels in the second layer are used to convolve the 64 (11 x 11 x 6) blocks of feature produced by the first layer. This layer's dropout rate is 0.25. The third layer convolves the block of features produced by the second layer using 128 (3 x 3 x 1) to create 128 (9 x 9 x 6) feature blocks. This layer has a dropout rate of 0.4 and the final two-layer networks are entirely interconnected layers. Based on the features retrieved from the first three layers, the video blocks are categorized. The ReLU activation function is used in the network's first four-layer and the function of SoftMax is used in the last layer for classification.

5. Experimental setup and results

In this paper, the proposed method is simulated using the requirements of hardware including a 2.8 GHZ genuine CPU and 8GB of RAM. The datasets of UCSD and UMN are used and the outcomes demonstrated that the proposed method evaluates better than the existing methods. The size of the batch is set to 256, the epoch is set to 30, and normal behaviors are used to train the spatiotemporal inter-fused autoencoder. To create the detection model, the spatiotemporal CNN is trained employing the labeled negative and positive data. The network epoch is set to 30 and the size of the batch is set to 300. This section presents the proposed method details of implementation and the analysis of results. Section 4.1 describes evaluation metrics, section 4.2 includes experimental results, and a comparative analysis with existing methodologies is described in section 4.3.

5.1 Evaluation metrics

Two of the following parameters can be determined using the ROC curve.

- Area under curve (AUC) – It calculates the area under the receiver operating characteristics curve and compares the true positive rate (TPR) and false positive rate (FPR) at various cut-off values.
- Recall – Recall or TPR is the number of accurately predicted positive cases among all positive samples presented in Eq. (5).

$$Recall = \frac{TP}{FN+TP} \tag{5}$$

- False positive rate – FPR is described as the proportion of fraudulent alerts that were incorrectly predicted out of all legal transactions presented in Eq. (6).

$$FPR = \frac{FP}{TN+FP} \tag{6}$$

- Accuracy – Accuracy is the proportion of accurate predictions to all input samples and it is calculated using the below Eq. (7).

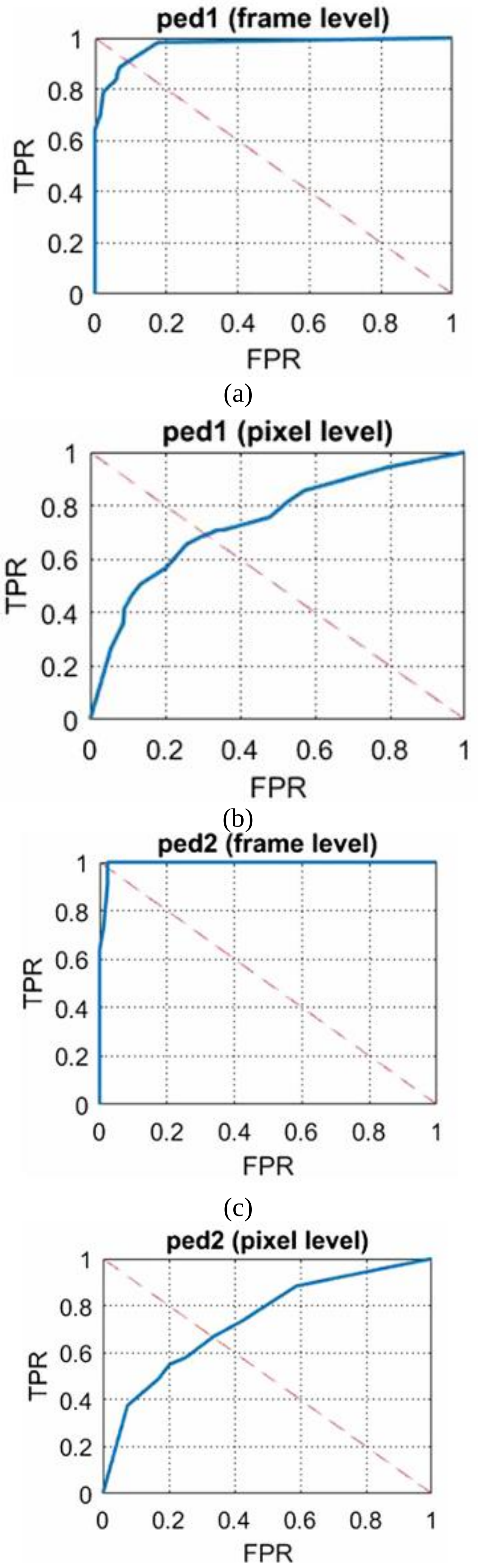
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

- Precision - The precision measures the percentage of actual data records versus expected data records is expressed in Eq. (8).

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

- F1-Score – It is also known as the harmonic mean, which seeks a balance between recall and precision is expressed in Eq. (9).

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \tag{9}$$



(a)

(b)

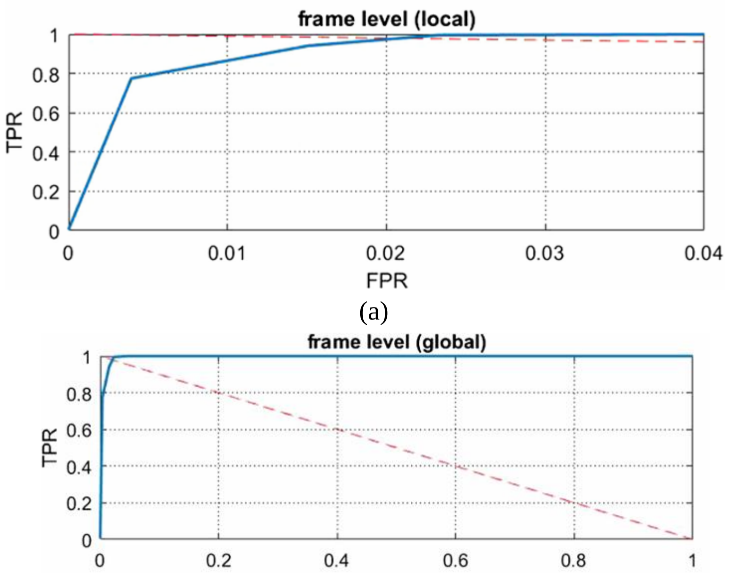
(c)

(d)

Figure. 5 Proposed method's ROC curve for the dataset of UCSD

5.2 Experimental results

The experiment outcomes are evaluated by using both the level of frame and pixel. Frame level is considered abnormal if at least one pixel is found to



(a)

(b)

Figure. 6 Proposed method’s ROC curve for the dataset of UMN

be out of the ordinary. The frame is considered true positive (TP) if the corresponding ground truth is also anomalous.

If not, it is regarded as a false positive (FP) and it is unable to assess the algorithm’s accuracy in locating the behavior of abnormal. At the level of the pixel, if more than 40% of anomalous pixels in the ground truth are found in a structure, it is said to be true positive (TP) if not it is regarded as an FP. This method is significantly stronger than the criterion at the frame level, and evaluates the way accurately the algorithm determines the location of the abnormal behavior.

The UCSD dataset’s ROC is shown in Fig. 5. Evaluations of ped1 and ped2 occurred at the frame and pixel levels, respectively. The comparison of ROC curves for the criterion of frame level using the UCSD ped1 and ped2 datasets are shown in (a) and (c) respectively. In the dataset of ped1, the proposed method is significantly less effective than supervised methods, according to the evaluation of frame-level. In the dataset of ped2, the method evaluates better than the models of deep learning and comes closer to supervised learning techniques. As a result, the frame-level criterion produces periodic false alarms from deep learning models. Fig. 5 (b), (d) compares the curves of ROC for the criterion of pixel-level using the datasets from the UCSD ped1 and ped2 appropriately.

A crowd anomalous activity detection is run to compare the effectiveness of various methods. The university of minnesota (UMN) surveillance videos are employed. The ROC of the UMN dataset (a) and (b) is displayed in Fig. 6. The proposed method’s classification performance on the UMN dataset is comparable to the global and local frame levels for anomaly detection. The UMN dataset’s significant impacts on the average motion intensity of the scene are an apparent feature. The later frames will change greatly from the earlier ones when anomalies occur, such as when the crowd quickly splits.

5.3 Comparative analysis

The comparative analysis includes methods, datasets, accuracy, precision, recall, and f1-score. Table 3. shows the comparison of the proposed method with the existing methods such as Hybrid CNN and RF classifiers [16], MCMS-BCN Attention+densenet121/Efficientnetv2 [17], and attention mechanism [22] on the UCSD dataset. Table 4. shows the comparison of the proposed method with the existing methods such as Hybrid CNN and RF classifiers [16], MCMS-BCN Attention+densenet121/Efficientnetv2 [17], and PGD and enhanced entropy classifier [21] on UMN dataset. Table 5. shows that the comparison of proposed method AUC values with the existing methods such as LRCCDL [14], Hybrid CNN and RF classifiers [16], on UMN and UCSD datasets.

In Table 3, UCSD ped 1, ped 2 achieves better accuracy of 99.76%, 99.92%, better precision of 99.82%, 99.90%, better recall of 99.92%, 99.90%, and better f1-score of 99.85%, 99.89% compared to the existing methods. In table 4, UMN scene 1,2 and 3 achieves better accuracy of 99.15%, 98.45, 94.67%, better precision of 99.57%, 99.32%, 99.82%, better recall of 98.85%, 97.82%, 93.50% and better f1-score of 99.20%, 98.91%, 98.47% compared to the existing methods. In table 5. UMN scene1, scene2, scene3,

Table 3. Comparison of proposed method with the existing methods on UCSD dataset

Author	Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Tarik Alafif [16]	Hybrid CNN and RF classifiers	UCSD ped 1	99.49	99.60	99.88	99.74
		UCSD ped 2	99.62	99.76	99.86	99.81
E. M. C. L. Ekanayake [17]	MCMS-BCNN-Attention+ densenet121/ Efficientnetv2	UCSD ped 1	98.62	99.51	97.24	98.61
		UCSD ped 2	98.95	99.54	98.36	98.95
Qianqian Zhang [22]	Attention mechanism	UCSD ped 1	80.5	N/A	N/A	N/A
		UCSD ped 2	97.9	N/A	N/A	N/A
Proposed method	Spatiotemporal inter-fused autoencoder	UCSD ped 1	99.76	99.82	99.92	99.85
		UCSD ped 2	99.92	99.90	99.90	99.89

Table 4. Comparison of the proposed method with the existing methods on the UMN dataset

Author	Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Tarik Alafif [16]	Hybrid CNN and RF classifiers	UMN scene 1	88.85	99.34	87.35	92.96
		UMN scene 2	81.07	99.06	76.23	86.16
		UMN scene 3	93.33	99.40	93.32	96.26
E. M. C. L. Ekanayake [17]	MCMS-BCNN-Attention+ densenet121/ Efficientnetv2	UMN plaza 1	99.10	99.41	98.78	99.09
		UMN plaza 2	98.38	99.12	97.63	98.37
Faisal Abdullah [21]	PGD and enhanced entropy classifier	UMN scene 1	87.43	N/A	N/A	N/A
		UMN scene 2	83.21	N/A	N/A	N/A
		UMN scene 3	90.63	N/A	N/A	N/A
Proposed method	Spatiotemporal inter-fused autoencoder	UMN scene 1	99.15	99.57	98.85	99.20
		UMN scene 2	98.45	99.32	97.82	98.91
		UMN scene 3	94.67	99.82	93.50	98.47

Table 5. Comparison of proposed method AUC values with the existing methods on UMN and UCSD datasets

Author	Methods	Datasets	AUC (%)
Li [14]	LRCCDL	UMN scene 1	99.94
		UMN scene 2	99.55
		UMN scene 3	99.93
		UCSD ped 1	90.01
		UCSD ped 2	95.20
Tarik Alafif [16]	Hybrid CNN and RF classifiers	UMN scene 1	99.73
		UMN scene 2	99.79
		UMN scene 3	99.77
		UCSD ped 1	88.87
		UCSD ped 2	98.55
Proposed method	Proposed method Spatiotemporal inter-fused autoencoder	UMN scene 1	99.96
		UMN scene 2	99.83
		UMN scene 3	99.97
		UCSD ped 1	90.15
		UCSD ped 1	99.72

UCSD ped 1, and ped2 datasets achieves better AUC values of 99.96%, 99.83%, 99.97%, 90.15%, and 99.72% compared to the existing methods.

5.4 Discussion

This section provides a discussion about the proposed spatiotemporal inter-fused autoencoder and compares those results with existing methods such as LRCCDL [14], hybrid CNN and RF classifiers [16], MCMS-BCN Attention+densenet121/Efficientnetv2 [17], PGD and enhanced entropy classifier [21], and attention mechanism [22] in comparative analysis 4.3.

The major goal of this study is to identify spatiotemporal patterns of typical activity from a stream of surveillance videos for abnormal detection and localization. The spatiotemporal CNN is trained using the method of supervised training with negative and positive data to develop the model of detection. The UMN (scenes 1, 2, and 3) and UCSD (ped1 and ped 2) datasets are used as benchmark datasets for this method. The outcomes of the experiment demonstrate that the proposed method is more accurate than the existing algorithms at both frame and pixel levels and it can locate anomalous behaviors in real-time. In the result analysis, UMN and UCSD datasets achieves better accuracy, precision, recall, f1-score, and AUC when compared to the existing methods.

6. Conclusion

In this paper, a spatiotemporal inter-fused autoencoder is proposed for detecting abnormal behavior. The development network of the spatiotemporal inter-fused autoencoder to remove samples of anomalous crowd behavior from the dataset assists in the better definition of anomalous crowd behavior. When the CPU is used, the spatiotemporal CNN-based model's basic structure allows for high-accuracy real-time detection. Both networks are capable of extracting advanced features of semantics from videos, which improves the algorithm's applications. The proposed method achieves better accuracy of 99.76%, 99.92%, 99.15%, 98.45%, and 94.67% in UCSD ped1, UCSD ped2, and UMN scene 1, 2, 3 datasets compared to the Hybrid CNN and RF classifiers, MCMS-BCN Attention+densenet121/Efficientnetv2, PGD and enhanced entropy classifier, and attention mechanism. Additionally, the proposed method achieves better AUC of 99.96%, 99.83%, 99.97%, 90.15%, and 99.72% in the datasets of UCSD ped1, UCSD ped2, and UMN scenes 1,2, 3. In the future, intend to combine information from many modalities to identify groups within crowds and to recognize suspicious behavior in a crowd.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] R. Nawaratne, D. Alahakoon, D. D. Silva, and X. Yu, "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance", *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 1, pp. 393-402, 2020.

[2] E. Hatirnaz, M. Sah, and C. Direkoglu, "A novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos", *Multimedia Tools and Applications*, Vol. 79, No. 25-26, pp. 17579-17617, 2020.

[3] M. Murugesan and S. Thilagamani, "Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network", *Microprocessors and Microsystems*, Vol. 79, p. 103303, 2020.

[4] M. U. Farooq, M. N. M. Saad, and S. D. Khan, "Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd", *The Visual Computer*, Vol. 38, No. 5, pp. 1553–1577, 2022.

[5] S. Xie, X. Zhang, and J. Cai, "Video crowd detection and abnormal behavior model detection based on machine learning method", *Neural Computing and Applications*, Vol. 31 (1-Supplement), pp. 175-184, 2019.

[6] T. Li, X. Chen, F. Zhu, Z. Zhang, and H. Yan, "Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection", *Neurocomputing*, Vol. 439, pp. 256-270, 2021.

[7] X. Zhang, S. Yang, J. Zhang, and W. Zhang, "Video anomaly detection and localization using motion-field shape description and homogeneity testing", *Pattern Recognition*, Vol. 105, p. 107394, 2020.

[8] L. Xia and Z. Li, "A new method of abnormal behavior detection using LSTM network with temporal attention mechanism", *The Journal of Supercomputing*, Vol. 77, No. 4, pp. 3223-3241, 2021.

[9] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets", *Neurocomputing*, Vol. 371, pp. 188-198, 2020.

[10] Z. Ilyas, Z. Aziz, T. Qasim, N. Bhatti, and M. F. Hayat, "A hybrid deep network based approach for crowd anomaly detection", *Multimedia Tools and Applications*, Vol. 80, No. 16, pp. 24053-

24067, 2021.

[11] P. Pattan and S. Arjunagi, “A human behavior analysis model to track object behavior in surveillance videos”, *Measurement: Sensors*, Vol. 24, p. 100454, 2022.

[12] A. A. Khan, M. A. Nauman, M. Shoaib, R. Jahangir, R. Alroobaea, M. Alsafyani, A. Binmahfoudh, and C. Wechtaisong, “Crowd Anomaly Detection in Video Frames Using Fine-Tuned AlexNet Model”, *Electronics*, Vol. 11, No. 19, p. 3105, 2022.

[13] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri, and S. S. Ullah, “Anomaly detection in traffic surveillance videos using deep learning”, *Sensors*, Vol. 22, No. 17, p. 6563, 2022.

[14] A. Li, Z. Miao, Y. Cen, X. P. Zhang, L. Zhang, and S. Chen, “Abnormal event detection in surveillance videos based on low-rank and compact coefficient dictionary learning”, *Pattern Recognition*, Vol. 108, p. 107355, 2020.

[15] Z. Fan, J. Yin, Y. Song, and Z. Liu, “Real-time and accurate abnormal behavior detection in videos”, *Machine Vision and Applications*, Vol. 31, No. 7, p. 72, 2020.

[16] T. Alafif, A. Hadi, M. Allahyani, B. Alzahrani, A. Alhothali, R. Alotaibi, and A. Barnawi, “Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds”, *Electronics*, Vol. 12, No. 5, p. 1165, 2023.

[17] E. M. C. L. Ekanayake, Y. Lei, and C. Li, “Crowd Density Level Estimation and Anomaly Detection Using Multicolumn Multistage Bilinear Convolution Attention Network (MCMS-BCNN-Attention)”, *Applied Sciences*, Vol. 13, No. 1, p. 248, 2022.

[18] A. S. Patel, R. Vyas, O. P. Vyas, M. Ojha, and V. Tiwari, “Motion-compensated online object tracking for activity detection and crowd behavior analysis”, *The Visual Computer*, Vol. 39, No. 5, pp. 2127-2147, 2023.

[19] F. Abdullah and A. Jalal, “Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system”, *Arabian Journal for Science and Engineering*, Vol. 48, No. 2, pp. 2173-2190, 2023.

[20] E. B. Varghese, S. M. Thampi, and S. Berretti, “A psychologically inspired fuzzy cognitive deep learning framework to predict crowd behavior”, *IEEE Transactions on Affective Computing*, Vol. 13, No. 2, pp. 1005-1022, 2022.

[21] F. Abdullah, Y. Y. Ghadi, M. Gochoo, A. Jalal, and K. Kim, “Multi-person tracking and crowd behavior detection via particles gradient motion descriptor and improved entropy classifier”, *Entropy*, Vol. 23, No. 5, p. 628, 2021.

[22] Q. Zhang, H. Wei, J. Chen, X. Du, and J. Yu, “Video Anomaly Detection Based on Attention Mechanism”, *Symmetry*, Vol. 15, No. 2, p. 528, 2023.