Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.034805
Article

An Efficient Attention-Based Strategy for Anomaly Detection in Surveillance Video

Sareer Ul Amin[1], Yongjun Kim[2], Irfan Sami[3], Sangoh Park[1,*] and Sanghyun Seo[4,*]

[1]Department of Computer Science and Engineering, Chung-Ang University, Seoul, 06974, Korea
[2]Intelligent Convergence Research Lab., ETRI, DaeJeon, 34129, Korea
[3]Department of Electrical and Electronics Engineering, Chung-Ang University, Seoul, 06974, Korea
[4]College of Art and Technology, Chung-Ang University, Anseong, 17546, Korea
*Corresponding Authors: Sangoh Park. Email: sopark@cau.ac.kr; Sanghyun Seo. Email: sanghyun@cau.ac.kr
Received: 28 July 2022; Accepted: 21 November 2022

Abstract: In the present technological world, surveillance cameras generate an immense amount of video data from various sources, making its scrutiny tough for computer vision specialists. It is difficult to search for anomalous events manually in these massive video records since they happen infrequently and with a low probability in real-world monitoring systems. Therefore, intelligent surveillance is a requirement of the modern day, as it enables the automatic identification of normal and aberrant behavior using artificial intelligence and computer vision technologies. In this article, we introduce an efficient Attention-based deep-learning approach for anomaly detection in surveillance video (ADSV). At the input of the ADSV, a shots boundary detection technique is used to segment prominent frames. Next, The Light-weight Convolution Neural Network (LWCNN) model receives the segmented frames to extract spatial and temporal information from the intermediate layer. Following that, spatial and temporal features are learned using Long Short-Term Memory (LSTM) cells and Attention Network from a series of frames for each anomalous activity in a sample. To detect motion and action, the LWCNN received chronologically sorted frames. Finally, the anomaly activity in the video is identified using the proposed trained ADSV model. Extensive experiments are conducted on complex and challenging benchmark datasets. In addition, the experimental results have been compared to state-of-the-art methodologies, and a significant improvement is attained, demonstrating the efficiency of our ADSV method.

Keywords: Attention-based anomaly detection; video shots segmentation; video surveillance; computer vision; deep learning; smart surveillance system; violence detection; attention model

1 Introduction

Recently, the surge in the overall crime rate has become one of the leading causes of losing money and lives [1]. Advanced surveillance is the most effective method for promptly detecting such unusual events. A significant amount of attention is paid to anomaly event recognition in video surveillance due to its wide range of applications in a variety of fields, such as crime prevention, traffic safety, and intelligent video surveillance monitoring [2]. Globally, vast amounts of surveillance cameras have been installed in a variety of locations for public safety in recent years [3]. Due to the limits of manual monitoring, law enforcement agencies are unable to identify or prevent abnormal behavior. Unusual behavior must be identified using a computer vision-based system capable of classifying usual and unusual behavior without any individual involvement. Such an intelligent method is preferable for monitoring and reduces the amount of labor needed for 24-h manual observation. The literature [4–6] provides several approaches for defining anomalous activity as "the existence of variation in normal patterns." For instance, anomaly detection is addressed as a classification issue [7,8], whereby visual features are fed into the algorithm, which then learns the difference between usual and unusual events. In other contexts, such as the detection of violence and road traffic accidents [9], it is recognized as a binary classification issue. However, these solutions are restricted to two kinds of behaviors: violent behavior, and normal behavior, offering only a percentage of the solution for implementation in real-world scenarios. Until now, sparse coding-based anomaly detection methods have demonstrated promising results [10–12], and these methods are assumed to be the standard for anomaly classification. These techniques are trained so that the start segments of video clips (i.e., prior to an uncommon event) are utilized to develop a vocabulary of usual events. Despite that, this strategy is insufficient for correctly detecting aberrant activities from a dictionary of regular events. Anomaly detection strategies based on weakly-supervised Multiple-Instance Learning (MIL) are also investigated in [13]. During the training phase of this technique, the videos are separated into a predefined number of clips. These clips generate bag instances containing samples of usual and unusual events, and they can learn instance-level labels for the bags they create. Since surveillance system changes over time, sparse-coding schemes have particular problems. For example, transferring the dictionary learned from usual and unusual events, results in a high proportion of false alarms. In addition, distinguishing irregular activities in surveillance cameras is exceedingly challenging due to their low quality, high intra-class flexibility, and absence of labeled data, as anomalous activities are infrequently connected with normal appearance. Machines must rely on visual characteristics, whereas people can distinguish both typical and atypical events using their rational thinking. In general, visually robust features exceed visually weak features for event detection and recognition [14]. Existing approaches are predominantly plagued by a high rate of false alarms, resulting in poor performance. In addition, while these methods perform well with tiny datasets, their effectiveness is limited whenever applied to actual scenarios. In this research, we solve these issues by proposing a unique and optimal Light-weight technique to predict abnormalities in surveillance footage. The proposed ADSV method employs a windowing method and analyses a sequence of frames in chronological order to track motion as well as action in surveillance footage. The proposed ADSV method learns the visual unique characteristics from a series of frames per training sample by using the video's spatial and temporal characteristics. The prominent contribution of our research work may be summarized as follows:

- An effective and efficient shot segmentation-based pre-processing strategy is proposed, wherein shots are segmented from surveillance video having anomaly activity using a shot boundary detection algorithm.
- An efficient, novel and LWCNN with TimeDistributed 2D layers is proposed to extract and learn Spatial-temporal patterns from a series of frames per training sample.

- The ADSV method proposes a hybrid CNN and LSTM training mode that can efficiently process sequential data and assure training speed. In addition, the unnecessary features of data will degrade the model's performance; hence, the attention network is employed to reallocate feature weights to maintain the model's performance and enhance its generalization capability.
- The complex and challenging benchmark datasets UCF-Crime and CUHK-Avenue are utilized to evaluate our proposed ADSV methodology. In contrast to current anomaly detection approaches, we achieve state-of-the-art results employing our proposed ADSV system, which is accurate while requiring fewer model parameters and a smaller overall size of the model (54.1 MBs).

The remaining sections of the paper are structured as follows. Section 2 explains the review of existing techniques. The proposed methodology is explained in Section 3. The dataset's information, the quantitative assessment, and the explanation are provided in Section 4. The concluding thoughts and future work are presented in Section 5.

2 Related Work

The detection and recognition of anomalies in the surveillance environment have been widely researched in the past. Conventional feature-based approaches and deep feature-based approaches for abnormal event detection are the two primary categories mentioned in the literature on anomaly detection methods. In the following, prominent techniques in both categories are briefly discussed.

2.1 Conventional Feature-Based Approaches

Traditionally, anomaly identification mostly relied on manual, low-level feature-based approaches. These systems are grouped into three tiers: (1) Retrieval of features, which extracts low-level features from the training dataset; (2) pattern learning, which is differentiated by the dispensation of regular occurrences; and (3) anomaly identification, which identifies dissimilar patterns or deviations as abnormal occurrences. Zhang et al. [15] used the Markov-random field to depict frequent events by employing spatial and temporal information. In a similar way, Mehran et al. [16] proposed a social-interacting method whereby optical flow was used to recognize regular and aberrant actions and estimate cooperative forces. In addition, Nor et al. [17] presented a paradigm for interpretable anomaly detection that aids prognostic and health management PHM. Their methodology relies on a Bayesian learning algorithm with predetermined prior and probabilities [18]. It offers additional descriptions to produce explanations locally and globally for PHM tasks. A similar Attention-based LSTM is developed by Ullah et al. [19] for activity recognition in sports. They refined the spatial characteristics using convolution block attention. The refined extracted features are classified into various sports activities using a densely connected convolutional network with an activation function of SoftMax. Selicato et al. [20] tried to detect abnormalities in gene data. For instance, they develop a method based on ensemble for identifying regular and aberrant expression of gene patterns by employing traditional cluster algorithms as well as principal component analysis (PCA). Riaz et al. [21] suggested deep ensemble-based methods for anomaly detection in complicated scenarios. Furthermore, to identify human being joints, a position-based estimation technique is integrated. The identified joints are considered as features and sent to a neural network for the identification of anomalies. Lately, Zhao et al. [22] came up with the unsupervised method that combined a time-varying-based sparsely coding style, online querying inputs, as well as sparsely reconstruction capability derived using a trained lexicon of all activities to identify irregularities in videos. However, having the ability to

recognize abnormalities promptly has remained a difficulty, which has piqued the attention of experts in the field.

2.2 Deep Feature-Based Approaches

In the present era, deep feature-based approaches have achieved remarkable popularity in a range of unstructured multi-dimension data domains, such as activity recognition and video analysis, in comparison to conventional techniques. Luo et al. [23] designed a technique wherein video frames are encoded with a convolutional network and unusual activities are recognized with Conv-LSTM. Furthermore, the encoder captures video deviation to identify anomalies in monitoring systems. Luo et al. [23] developed a Conv-LSTM network with an auto-encoder for the detection of anomalies. In addition, they extended their approach by recognizing abnormalities with Recurrent Neural Network (RNN) and auto-encoder. Liu et al. [24] introduced a method for recognizing anomalies in the video by using the spatiotemporal detector. In this model, the discriminating prominence information and a group of temporal texture information were considered as usual data activities. Chang et al. [25] introduced a cluster-based auto-encoder to efficaciously capture valuable patterns from regular activities. Two phases have been utilized to understand spatiotemporal information consistency, the spatial-auto-encoder in the first phase is responsible for the final individual frame. However, the second phase's temporal-auto-encoder executes and generates the RGB difference among frames. Furthermore, abnormalities in videos are identified using generative models. Sabokrou et al. [26] developed Generative Adversarial Network (GANs) for identifying video anomalies. This network guides the normally distributed data employing GANs strategies. Recent video annotation techniques employ 3D Convolution (C3D) and MIL to detect abnormal occurrences [27]. For instance, Sultani et al. [12] introduced a system for identifying abnormal occurrences using weak video annotations as well as the MIL technique. This system was implemented on regular as well as irregular video data by generating two different bags of usual and unusual activities and further used the MIL approach to predict the probability scores of abnormal video activity. Landi et al. [28] suggested a method for tube extraction, which utilizes location information to construct a regression system for anomalies. Before passing the data to the regression model [29], the pooling layer integrates the spatial and temporal information of model inception and optic flow respectively. Zhong et al. [6] proposed an unsupervised method for identifying abnormalities and a supervised system for classifying actions having noise annotations. Due to the unpredictability of anomalous occurrences, the abnormal video annotations remained unclear. In addition, a graph convolution network was developed to remove the noise from these annotations, as well as an activity classifier was used to classify the activities. In comparison to the current strategies, this article Presented an efficient Attention-based deep-learning approach for anomaly detection in surveillance video. Section 3 discusses the prime components of the suggested ADSV method.

3 Proposed Methodology

This section elaborates on the proposed (ADSV) method and its fundamental component structure. Fig. 1 depicts a visual representation of our ADSV strategy. In brief, the technique of anomaly detection in surveillance video consists of three main components: Shot segmentation, feature extraction, sequence learning, and abnormality classifications. To begin with, a shots boundary detection technique is used to segment prominent frames. Furthermore, The LWCNN model receives the segmented frames in order to extract spatial and temporal information from the intermediate layer. Following that, spatial and temporal features are learned using LSTM cells and Attention Network from a series of frames for each sample of abnormal activity. To detect motion and action, the LWCNN

received chronologically sorted frames. Finally, the abnormal activities in video shots are identified using the proposed trained ADSV model.

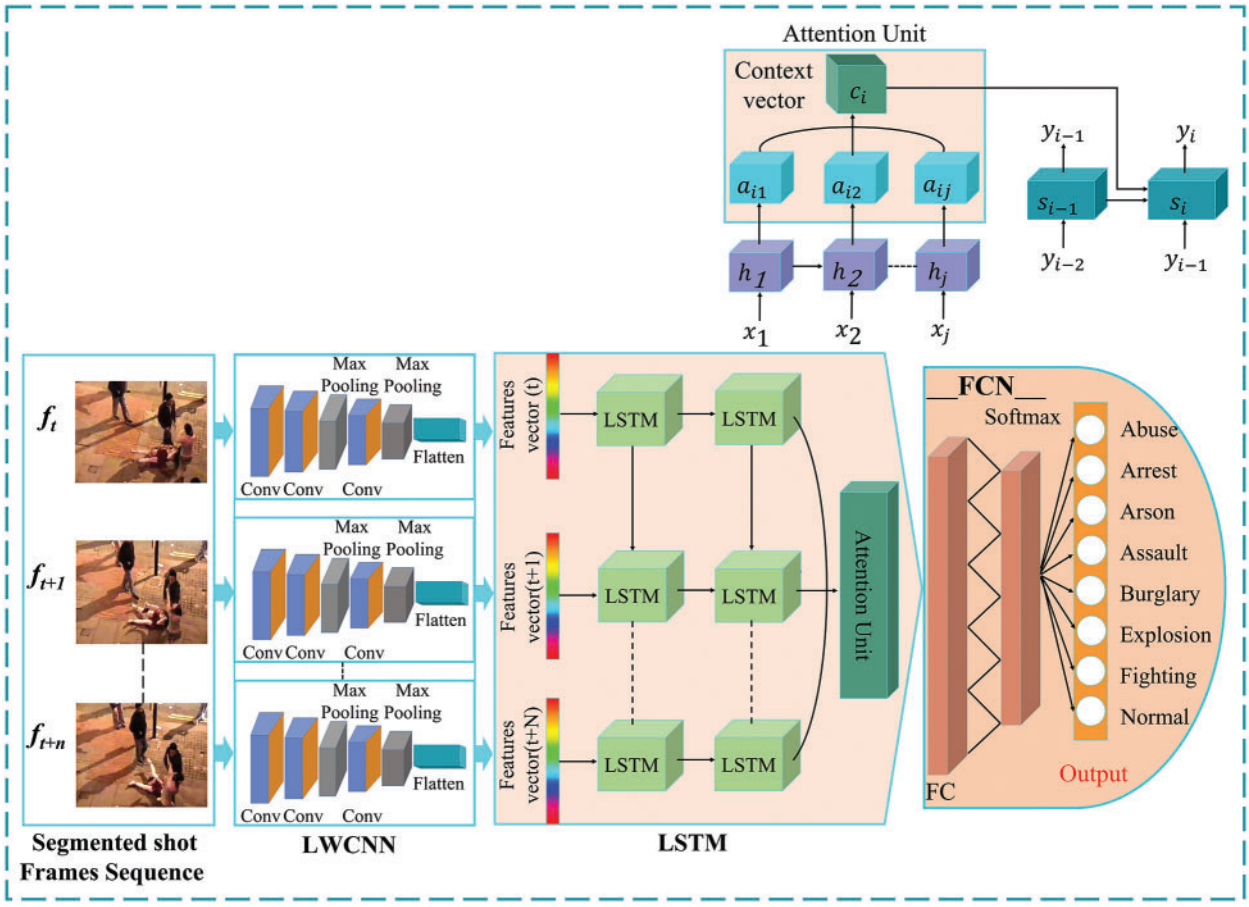

Figure 1: The proposed ADSV framework for video anomaly detection consists of three main components: shot segmentation, feature extraction, sequence learning, and abnormality classifications. At the input of the ADSV, a shots boundary detection technique is used to segment prominent frames. Next, The LWCNN model receives the segmented frames to extract spatial and temporal information from the intermediate layer. Following that, spatial and temporal features are learned using LSTM cells and Attention Network from a series of frames for each anomalous activity in a sample. To detect motion and action, the LWCNN received chronologically sorted frames. Finally, the anomaly activity in the video is identified using the proposed trained ADSV model

3.1 Segmentation of Shots by Using Boundary Detection Algorithm

Shots Boundary Detection is a prerequisite for most video applications that include the comprehension, indexing, characterization, or classification of video and temporal segmentation, making it a prominent research issue in content-based video analysis. Table 1 represents the parameter descriptions of shots segmentation by using a boundary detection algorithm. The concept of the boundary detection algorithm is implemented in the following key phases:

- Image Segmentation: Divide each frame from the video into $\mathfrak{m}$-rows and $\mathfrak{n}$-columns blocks at the image segmentation level. Then, the difference between two successive frames of the relevant blocks is calculated. Lastly, the gap between consecutive frames is calculated by adding the differences created by the various weights.
- Attention-Model: Attention, a neurobiological word, refers to the ability or capability to concentrate mental energies on an object through careful observation or attentive listening. The attention model suggests that, from a visual perspective, different contents are prioritized based on their relative significance; it also represents the relative importance of frames. On this basis, it is possible to conclude that pixel in different positions contributes differently to the detection of shot boundaries; pixels on the edge contribute more than pixels in other positions. As a result, distinct weights are assigned to blocks at various positions. The spatial distribution of pixels with varying grey values and the relative significance of pixels with varying positions are evaluated.
- Histogram Matching: There are several types of histogram matching. In most literature, the matching difference is calculated using a color histogram. However, after evaluating different types of histogram matching techniques, it was shown that the x^2 histogram surpassed the others in recognizing shot boundaries [30]. As a result, the x^2 histogram matching technique is used.
- Shot-Boundary Detection: Histogram differences are used to detect shot boundary and consequently extracted the key frame based on underlying activity. The threshold $\tau = \mathcal{MD} + \alpha \times \mathcal{STD}$ has a constant value of α which applies weight to the $\mathcal{STD}$ for the overall τ. If $\mathcal{D}(f_t, f_{t+1}) \geq \tau$, frame f_t indicates the Prior shot's end, while the frame f_{t+1} indicates the subsequent shot's end. Typically, the minimum feasible shot duration should be between one and two and a half seconds long. In addition, a frame rate of at least 25–30 fps is required to ensure smoothness, or a flashlight might be visible. Thus, a shot should consist of 30–45 frames. Therefore, a shot combining rule is provided [30]: If a captured shot contains less than 38 frames, it should be combined with the following shot or declared independent. In Table 3, the pseudo-code implementation of shot boundary detection is presented.

Table 1: The parameter descriptions for the algorithm of shot boundary detection

Symbols	Parameter descriptions
$\mathcal{BD}(f_t, f_{t+1})$	Block Difference is the inconsistency between two frames' relevant blocks.
$\mathcal{H}(f_t, j, c, b)$	Histogram measurement at the "b^{th}" block of frame "t" for the "j^{th}" grey value in "c" RGB channel.
$\mathcal{H}(f_{t+1}, j, c, b)$	Histogram measurement at the "b^{th}" block of frame "$t+1$" for the "j^{th}" grey value in "c" RGB channel.
$\mathcal{N}_{\mathcal{B}}$	Count of all the blocks.
$\mathcal{N}_{\mathcal{H}}$	The total amount of grayscale values that can be used.
$\mathcal{D}(f_t, f_{t+1})$	The difference in histograms between two successive frames.
W_k	The weight of the Block at (k).
$\mathcal{MD}$	Estimating the histogram's mean value.
$\mathcal{STD}$	Calculating the histogram's standard deviation.
x	A frame of reference.

3.2 Proposed LWCNN Method

CNNs were motivated by biological processes similar to the structure of the visual cortex in animals. The connection of the neurons in the convolution layers (*CL*) is arranged in a similar way to that of the visual system in an animal. Each neuron in the cortex responds to stimuli within a small portion of the input frame, known as the reception field. *CL* can maintain spatial relationships among input frames in video analysis by learning feature representations by applying filters whose values are learned throughout the training process [31]. The Proposed LWCNN method is comprised of three TimeDistributed 2D *CL* and two TimeDistributed 2D max pooling layers, with the number of channels, kernel, padding, and strides stated in Table 2. Furthermore, a kernel size of 3 × 3, a stride of 2 × 2, and an activation function of the Rectified Linear Unit (ReLU) are applied to each TimeDistributed 2D *CL*. Besides that, TimeDistributed 2D max pooling with a stride size of 2 × 2 is used to minimize the network's size following the second and third *CL*. Each convolutional process utilizes same-padding techniques to avoid losing information at the border of the input frame. In the first *CL*, we start with 64 feature maps, which are then followed by the identical feature maps in the second *CL*, and 128 activation maps are generated in the final *CL*.

Table 2: Descriptions of LWCNN technique used in the proposed ADSV method

LWCNN layer	No. of channels	Kernel (h × w)	Padding	Stride (h × w)	Output	Parameters
TimeDistributed-2D *CL*1 (ReLU)	64	3 × 3	Same	2 × 2	5,112 × 112 × 64	1792
TimeDistributed-2D *CL*2 (ReLU)	64	3 × 3	Same	2 × 2	5, 56 × 56 × 64	36928
TimeDistributed Max-Pooling-2D1	1	2 × 2	–	2 × 2	5, 28 × 28 × 64	0
TimeDistributed-2D *CL*3 (ReLU)	128	3 × 3	Same	2 × 2	5, 14 × 14 × 128	73856
TimeDistributed Max-Pooling-2D2	1	2 × 2	–	2 × 2	5, 7 × 7 × 128	0
TimeDistributed Flatten-1	–	–	–	–	5, 6272	0

Table 3: Pseudo-code implementation of shots boundary detection algorithm

Pseudo-code of shots segmentation by using a boundary detection algorithm

Required: The whole videos dataset $= \mathcal{T}_v \in \mathcal{R}^3$
Required: Divide each frame f from the video into sixteen blocks i.e., $\forall f \in \mathcal{B}^{16}$
for $\mathcal{T} \leftarrow 1\mathcal{T}_v$ **do**
  for $f \leftarrow 1\mathcal{T}$ **do**

$$\mathcal{BD}(f_t, f_{t+1}) \leftarrow \sum_{c=1}^{3} \sum_{b=1}^{\mathcal{N}_\mathcal{B}} \sum_{j=1}^{\mathcal{N}_\mathcal{H}} \frac{\mathcal{H}(f_t, j, c, b) - \mathcal{H}(f_{t+1}, j, c, b)}{x}$$

$$\mathcal{D}(f_t, f_{t+1}) \leftarrow \sum_{k=1}^{\mathcal{N}} W_k \mathcal{BD}_k(f_t, f_{t+1})$$

$$\mathcal{MD} \leftarrow \frac{\sum_{f_t=1}^{\mathcal{T}_{v-1}} \mathcal{D}(f_t, f_{t+1})}{\mathcal{T}_{v-1}}$$

$$\mathcal{STD} \leftarrow \sqrt{\frac{\sum_{f_t=1}^{\mathcal{T}_{v-1}} \left(\mathcal{D}(f_t, f_{t+1})\right)^2}{\mathcal{T}_{v-1}}}$$

(Continued)

Table 3: Continued

Pseudo-code of shots segmentation by using a boundary detection algorithm

```
        τ = MD + α × STD
        if D(f_t, f_{t+1}) ≥ τ then
            prior shot end frame f_t
            Succeeding shot end frame f_{t+1}
        else
            Print ("shot not found")
        end if
    end for
    return shot boundary detection for the input video
    Repeat loop till the end frame of the video
end for
```

As input, the suggested framework takes a sequence of the segmented frames that have been preprocessed. In addition, the proposed LWCNN is used to capture the spatial information of each frame, which is subsequently fed into the LSTM to extract temporal information. As shown in Fig. 1, we employed a frame-wise LWCNN to retrieve the spatial information of each frame. Furthermore, the input frames (f_t, f_{t+1}, f_{t+n}) are fed separately into an LWCNN, which converts every single frame into a particular series of spatial information representations, as shown in Eq. (1), where $\mathcal{F}_t$ represents a series of spatial information representations. Besides this, the temporal information $\mathfrak{h}_{t+1}$ is generated utilizing a series of spatial information representations as input to the LSTM model, as illustrated in Eq. (2). The LSTM layer's hidden state at time step t is represented by $\mathfrak{h}_t$, while the hidden state at time step $t-1$ is represented by $\mathfrak{h}_{t-1}$ [32]. The information from the previous time step is provided as input to the current time step.

$$\mathcal{F}_t = \texttt{LWCNN}(f_t, f_{t+1}, f_{t+n}) \tag{1}$$

$$\mathfrak{h}_{t+\mathfrak{n}} = \texttt{LSTM}(\mathcal{F}_t) \tag{2}$$

3.3 Sequence Learning and Anomalies Classification

The algorithms developed for applications that need sequential or temporal data are known as sequence learning. Moreover, RNNs are introduced to identify latent sequential patterns inside temporal and spatial sequential data. Video data is also a form of time series data in which changes in visual contents are represented by a large number of frames in such a way that the sequence of frames assists in knowing the context of behavior. In the case of long-term sequences, RNNs can comprehend such sequences, but they forget the sequence's prior inputs. This is referred to as the vanishing gradient issue, and it can be addressed using the LSTM, a type of RNN capable of tracking long sequences. Its distinctive structure, consisting of input, output, and forget gates, identifies long-term sequence patterns. During training, a sigmoid unit needs to adjust the gates by learning where to open and close them. Eqs. (3) to (8) outline the activities performed by the LSTM unit, where the input gate at time t is $\mathfrak{i}_t$ and the forget gate at time t is $\mathcal{F}_t$, which clears data from the memory cell as required and maintains a record of the preceding frame whose data must be erased from memory. The output gate $\mathcal{O}_t$ contains information regarding the next step, where g is the recurrent unit with activation function "$t\mathfrak{a}\mathfrak{n}\mathfrak{h}$" that is generated using the current frame's input and the state of the previous frame $\mathcal{S}_{t-1}$. Using

"$t\mathfrak{a}\mathfrak{n}\mathfrak{h}$" activation and the $\mathcal{C}_t$ memory cell, the LSTM step's hidden state is computed. An attention unit improves the capabilities of these LSTM layers [33]. The output of the final time step of $\mathfrak{h}_{t+\mathfrak{n}}$ is supplied as an input to the following attention unit. After receiving input from the attention network, the fully connected layer determines the final estimation score for each anomaly class.

$$\mathfrak{i}_t = \sigma\left(\left(\mathfrak{x}_t + \mathcal{S}_{t-1}\right)\mathcal{W}^{\mathfrak{i}} + b_{\mathfrak{i}}\right) \tag{3}$$

$$\mathcal{F}_t = \sigma\left(\left(\mathfrak{x}_t + \mathcal{S}_{t-1}\right)\mathcal{W}^{f} + b_{f}\right) \tag{4}$$

$$\mathcal{O}_t = \sigma\left(\left(\mathfrak{x}_t + \mathcal{S}_{t-1}\right)\mathcal{W}^{\mathcal{O}} + b_{\mathcal{O}}\right) \tag{5}$$

$$\mathrm{g} = t\mathfrak{a}\mathfrak{n}\mathfrak{h}\left(\left(\mathfrak{x}_t + \mathcal{S}_{t-1}\right)\mathcal{W}^{\mathrm{g}} + b_{\mathrm{g}}\right) \tag{6}$$

$$\mathcal{C}_t = \mathcal{C}_{t-1} \cdot f_t + \mathrm{g} \cdot \mathfrak{i}_t \tag{7}$$

$$\mathcal{S}_t = t\mathfrak{a}\mathfrak{n}\mathfrak{h}\left(\mathcal{C}_t\right) \cdot \mathcal{O}_t \tag{8}$$

3.4 Attention Mechanism

Attention has likely become the most essential idea in the fields of deep learning as well as computer vision. It is based on the biological systems of beings, which prefer to concentrate on the unique aspects of enormous amounts of data. The attention technique tackles extremely long-range dependency issues in LSTMs [33]. As seen in Fig. 1, the output at time i is given by $y_i = \mathrm{g}(y_{i-1}, s_i, c_i)$ where s_i is the decoder's hidden state. The context vector c_i is the weighted sum of the attention weight a_{ij} and the encoder's hidden state h_j, as given in Eq. (9). Attention weight a_{ij} is computed according to Eq. (10) and indicates the correlation between the *j-th* location of the input sequence and the *i-th* location of the result sequence. The attention score, e_{ij}, is dependent on the hidden state s_{i-1} of the decoder and the *j-th* annotate h_j of the input sequence, as indicated in Eq. (11), where v, w_1 and w_2 are the parameter matrices that are calculated by training the model. The technique of computing attention scores utilizing the preceding hidden state of the decoder is known as the Bahdanau style. In this study, the Bahdanau attention mechanism is utilized to enhance the sequential learning model.

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_{ij} \tag{9}$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{10}$$

$$e_{ij} = v^{\intercal} \tanh(w_1 s_{i-1} + w_2 h_j) \tag{11}$$

4 Experimental Results

Our proposed ADSV framework is evaluated using two publicly accessible datasets: CUHK-Avenue and UCF-Crime. Fig. 2 depicts a selection of usual and unusual event frames extracted from these datasets. Furthermore, the Three-Dimensional Uniform Manifold Approximation and Projection (3D-UMAP) Projection of the evaluated datasets is given in Fig. 3. The two quantitative measures were used to evaluate the proposed ADSV method: frame-based area under the curve (AUC) and receiver operating characteristic (ROC). Table 9 shows the training parameters of the proposed method. The algorithm is implemented employing Karas-supported Tensor Flow in the Python 3.7.1 platform. Experiments were performed on a computer (PC) equipped with an Nvidia-GeForce RTX-3070 GPU, 32 GB of RAM, the Windows 10 operating system, and the CUDA toolkit 11.0 with

cuDNN v8.0. The experimental results showed the efficacy of the suggested ADSV method as well as a sustained breakthrough in the state-of-the-art methods.

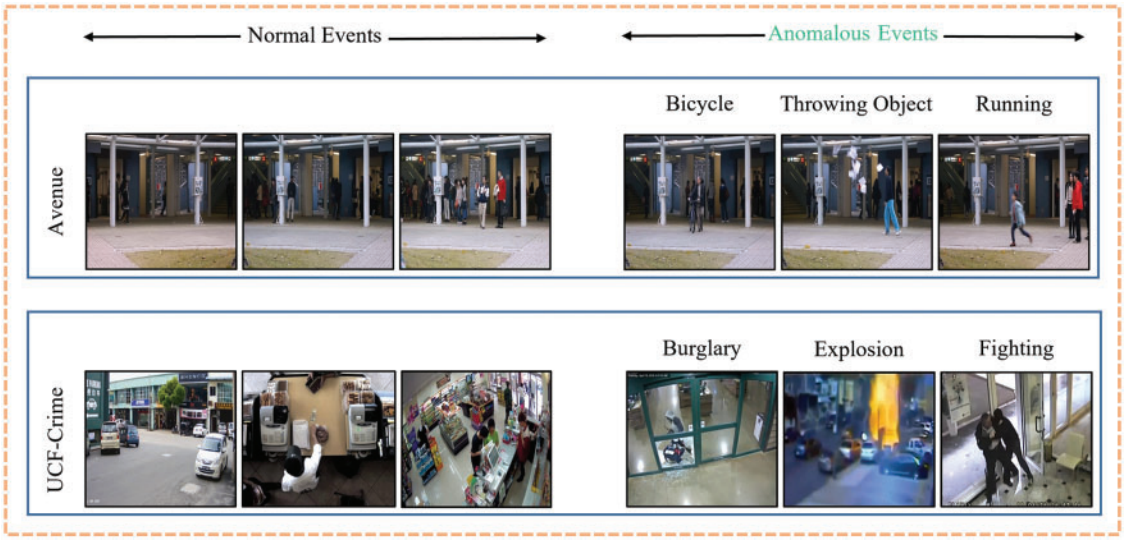

Figure 2: Samples of the anomaly detection dataset

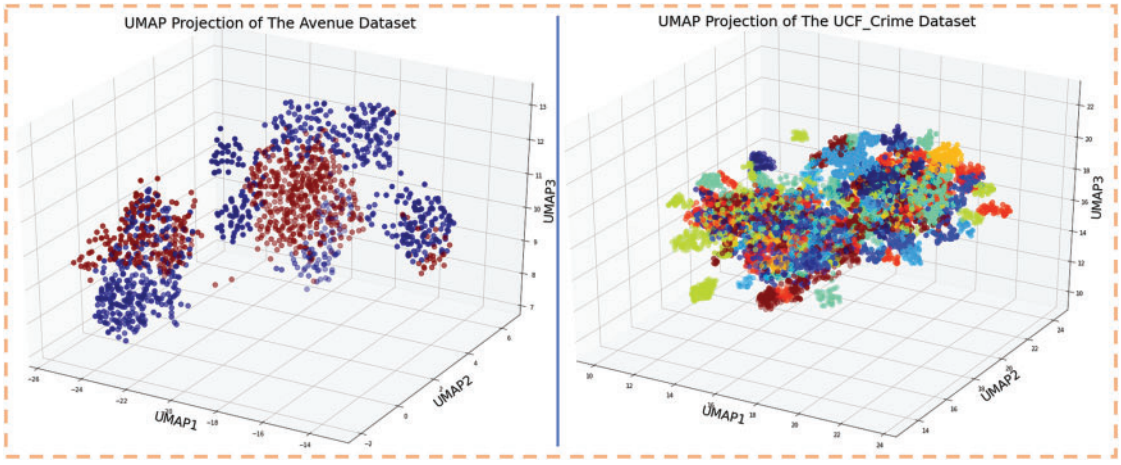

Figure 3: The 3D-UMAP Projection of the datasets Avenue and UCF-Crime

4.1 Datasets

CUHK-Avenue: the CUHK-Avenue dataset was collected at the Chinese University of Hong Kong using a static video camera with a clear resolution of 640 pixels by 360 pixels, which captured street activities. This dataset contains 16 train video clips representing regular human activity and 21 test video clips representing anomalous human activities and events. Walkers on the walkway and groups of walkers gathering on the pavement are examples of regular behavior. In contrast, persons wasting or discarding goods, loitering, going toward the camera, strolling on the grass, and leaving items are examples of anomalous activity.

UCF-Crime: The UCF-Crime dataset contains 1900 long, unedited videos for 13 real-world anomaly activities, such as abuse, arrest, assault, arson, burglary, explosion, fighting, road accident, robbery, stealing, shooting, shoplifting, and vandalism. The training set of the UCF-Crime dataset included 800 videos of normal events and 810 videos of abnormal events. The testing set consisted of the remaining 150 regular videos and 140 abnormal ones. In Table 4, the statistical information of the UCF-Crime dataset is presented.

Table 4: The statistical facts contained within the UC-Crime dataset

Types of anomalies	The number of videos	Training data	Testing data
Abuse.	50.	48.	02
Arrest.	50.	45.	05
Assault	50.	47	03
Arson	50.	41	09
Shooting	50.	27	23
Fighting	50	45	5
Explosion	50	29	21
Vandalism	50	45	05
Shoplifting	50	29	21
Stealing	100	95	05
Burglary	100	87	13
Robbery	150	145	05
Accident	150	127	23
Sum	950	810	140

4.2 Dataset Visualization Using UMAP

UMAP is a lately introduced strategy for learning manifolds that aims to accurately reflect local structures as well as effectively integrate global structures [34]. It offers a significant benefit over t-distributed Stochastic Neighbor Embedding (t-SNE). In contrast to t-SNE, UMAP has been shown to work effectively with huge datasets. UMAP is built on the following hypotheses:

- The data is transferred equally throughout a Riemannian manifold.
- The Riemannian metric is locally unchanging.
- The manifold is locally connected.

Based on these assumptions, the manifold can be represented as a fuzzy topological structure of the sample points with High Dimension (HD). The embedding manifold is determined by looking for a fuzzy topological structure in the Low-Dimension (LD) projection of the data. To construct the fuzzy topological structure, UMAP employs an HD graph to represent the sample points. The resultant HD graph is a weighted graph, where edge weights denote the probability that two points are related. UMAP calculates the similarities between HD sample points using an exponential probability distribution.

$$P_{i/j} = \exp\left(-\frac{\left(d\left(x_i, x_j\right) - \sigma_i\right)}{\sigma_i}\right) \tag{12}$$

The distance between the i^{th} and j^{th} sample points are denoted by $d\left(x_i, x_j\right)$, while the distance between the i^{th} data point and its first nearest neighboring are denoted by σ. The HD probability of data is symmetrized in UMAP:

$$P_{ij} = P_{(i/j)} + P_{(j/i)} - P_{\left(\frac{i}{j}\right)} P_{\left(\frac{j}{i}\right)} \tag{13}$$

The generated graph is a probability graph, as previously stated, and UMAP must identify the k^{th} adjacent neighbors:

$$k = 2^{\sum_i P_{ij}} \tag{14}$$

UMAP generates and optimizes an LD version of the HD graph that is as similar as possible. UMAP employs a probability measure to model distance in low dimensions:

$$q_{ij} = \left(1 + \alpha \left(y_i - y_j\right)^{2b}\right)^{-1} \tag{15}$$

where $\alpha \approx 1.93$ and $b \approx 0.79$ by default for UMAP. UMAP utilizes Binary Cross-Entropy (CE) as a cost function because of its capacity to capture the global data structure:

$$CE\left(P, Q\right) = \sum_i \sum_j P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right) + \left(1 - P_{ij}\right) \log\left(\frac{1 - P_{ij}}{1 - Q_{ij}}\right) \tag{16}$$

where P denotes the high-dimensional, and Q represents low-dimensional sample points' probabilistic similarity. To update the coordination of the LD sample points until convergence, cross-derivative entropy is utilized. In addition, Stochastic Gradient Descent was utilized by UMAP because of its quicker convergence and lower memory consumption. Several essential hyper-parameters influences UMAP's performance. The hyper-parameters are mentioned below:

- The target embedding's dimensionality.
- Selecting a low value for the k number of neighbors suggests that the interpretation will be extremely localized and extract fine-grained structural details. On the other hand, selecting a considerable value suggests that the prediction will be based on wider regions, which means some fine information structure will be missed.
- The smallest distance among points in the embedding space can be accepted. Lower minimum distance values will capture the underlying manifold structure more precisely, resulting in heavy clouds that make viewing difficult.

4.3 Evaluation and Comparison of the Avenue Dataset

The CUHK Avenue dataset is an open-source dataset that is commonly used to assess video anomaly detection techniques. We compared our proposed ADSV system to various current techniques on this dataset [6,11,23,24,35–38], including supervised and unsupervised strategies that also presented their frame-based AUC values on the CUHK Avenue dataset. Recent techniques, including Liu et al. [24] and Zhou et al. [39], revealed an AUC value of 84.9% and 86.1%, correspondingly, for the CUHK Avenue dataset. Zhou et al. [39] outscore earlier methods, as well as other recent techniques, including those described in [40], further reported high accuracy on the Avenue dataset. However, the ADSV system performs more effectively than most current methods, enhancing further state-of-the-art accuracy for the Avenue dataset. The proposed system outperforms state-of-the-art methods and accomplishes a frame-based AUC of 99%. Table 5 displays the quantitative comparison of the proposed ADSV system with the numerous current strategies in terms of frame-based AUC values.

Figs. 6c and 6d depict the training accuracy and loss graphs for the proposed system using Avenue Dataset respectively.

Table 5: A frame-based (AUC) evaluation of the proposed ADSV technique to existing approaches for the Avenue Dataset

Method	CUHK-Avenue Dataset
Hasan et al. [35]	70.2
Luo et al. [38]	80.3
Chong et al. [36]	80.6
Tudor Ionescu et al. [37]	81.7
Liu et al. [24]	84.9
Zhou et al. [39]	86.1
Proposed ADSV	99.0

4.4 Evaluation and Comparison of the UCF-Crime Dataset

The UCF-Crime dataset is an open-source dataset that is commonly used to assess algorithms for detecting anomalies in surveillance videos. We compared our ADSV method for video anomaly to other existing techniques [6,11,12,35,38] using the UCF-Crime dataset. The ROC and AUC are the evaluation metrics used to assess the performance of our system. We established the AUC and ROC curves using the test set from the UCF-Crime dataset to compare them to the results in [6,11,12,35,38]. The experimental research results shown in Fig. 4 reveal that our framework surpasses the techniques described in [6,11,12,35,38]. Table 6 compares the frame-based AUC values for the UCF-Crime dataset with other existing techniques. It can be seen that the proposed framework obtained the highest frame-based AUC value of 97%, a 14.88% increase over the approach of Zhong et al. [6], which achieved the last highest AUC value of 82.12%. Following, other research in [11,12,35] obtained AUC values of 75.41%, 50.6%, and 65.51%, respectively, revealing that our framework performs effectively with the UCF-Crime dataset. Fig. 5 illustrates the frame-based anomalies detection efficiency of the suggested ADSV method with the anomaly scores for the following test video sequences: (a) Arrest-48, (b) Fighting-47, (c) Assault-51, and (d) Normal Videos-27. Figs. 6a and 6b depict the training accuracy and loss graphs for the proposed model using UCF-Crime Dataset respectively.

4.5 In Comparison to Contemporary Methods

In this subsection, the performance of our proposed ADSV framework for anomaly detection is compared to that of existing methodologies utilizing the UCF-Crime dataset. The authors of [41] analyzed a range of deep learning algorithms using multi-layer Bidirectional Long Short-Term Memory Neural Network (BLSTM) integration, such as VGG-19, InceptionV3, and ResNet-50. Among all these methods, the technique with the smallest model size is ResNet-50+multi-layer BLSTM. In addition, Deep learning models have become increasingly sophisticated in the current era, requiring massive quantities of storage; they also have enhanced computation costs and restrictive installation protocols over the connected device. A delay in responses can lead to the loss of human life or property in an anomaly detection system; hence, efficient model selection is essential to every anomaly detection system. In comparison to other existing approaches [6,41–43], the suggested ADSV method for anomaly detection has a minimum storage capacity, fewer learning parameters, and a quicker processing time. Table 7 compares the effectiveness of the proposed ADSV method

to existing contemporary strategies in terms of model size (MB), time complexity, and parameter count. Furthermore, a strong and effective anomaly detection system with low false alarm rates is highly desired since, in everyday life surveillance, the frequency of normal actions occurring is higher than the frequency of abnormal actions. Consequently, we made a comparison between the proposed framework and recently published methodologies, as given in Table 8, and the results suggest that the proposed ADSV framework has had the minimum false alarm rates. In comparison to earlier methods, the suggested ADSV framework can process a sequence of 32-frames in 0.26 s [6,41–43]. Figs. 7a and 7b depict the class-wise accuracy of the CUHK-Avenue and UCF-Crime Dataset respectively.

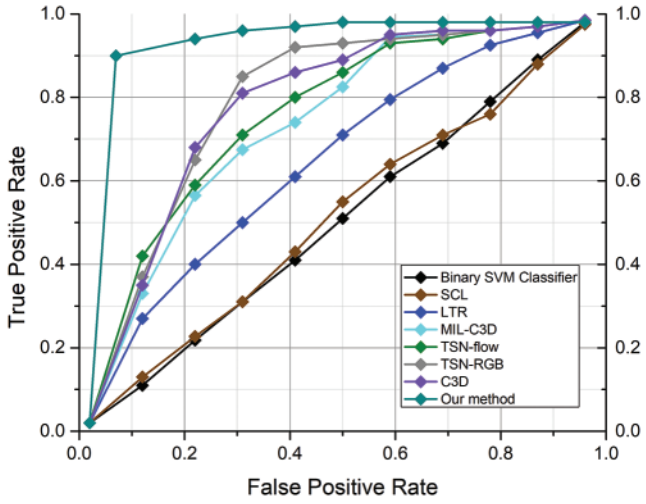

Figure 4: For the UCF-Crime dataset, the proposed strategy is compared with alternative methodologies already in use. The colors black, dark brown, blue, cyan, green, gray, purple, and teal, respectively, are used to depict the ROC curves for the binary SVM classifier [12], SCL [11], LTR [35], MIL-C3D [12], TSN-flow [6], TSN-RGB [6], C3D [6], and ADSV approach

Table 6: A frame-based (AUC) evaluation of the proposed ADSV technique to existing approaches for the UCF-Crime Dataset

Method	UCF-Crime Dataset [12]
Binary SVM classifier [12]	50.0
LTR [35]	50.6
Spatiotemporal [44]	63.0
SCL [11]	65.51
MIL-C3D without constraints [12]	74.44
MIL-C3D with constraints [12]	75.41
TSN-Optical flow [6]	78.08
C3D [6]	81.08
TSN-RGB [6]	82.12
Proposed ADSV	97.0

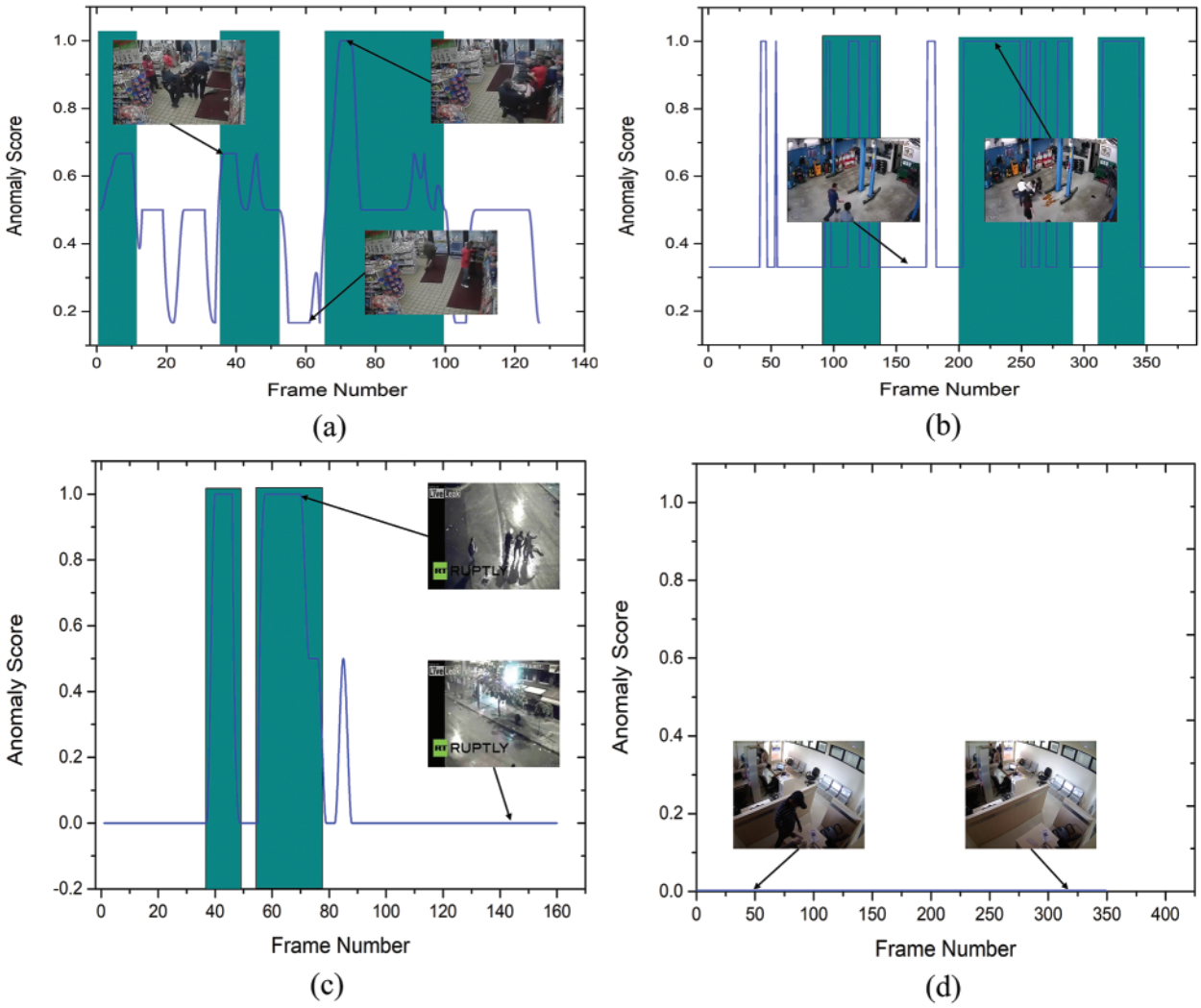

Figure 5: The abnormality score curves for four test videos of the UCF-Crime Dataset, including (a) Arrest-48, (b) Fighting-47, (c) Assault-51, and (d) Normal-Videos-27. The ground truth anomalous frames are highlighted by teal patches. For a more accurate depiction, we normalized anomaly values for each video to the interval [0,1]. It indicates that as abnormalities occur, anomaly scores grow

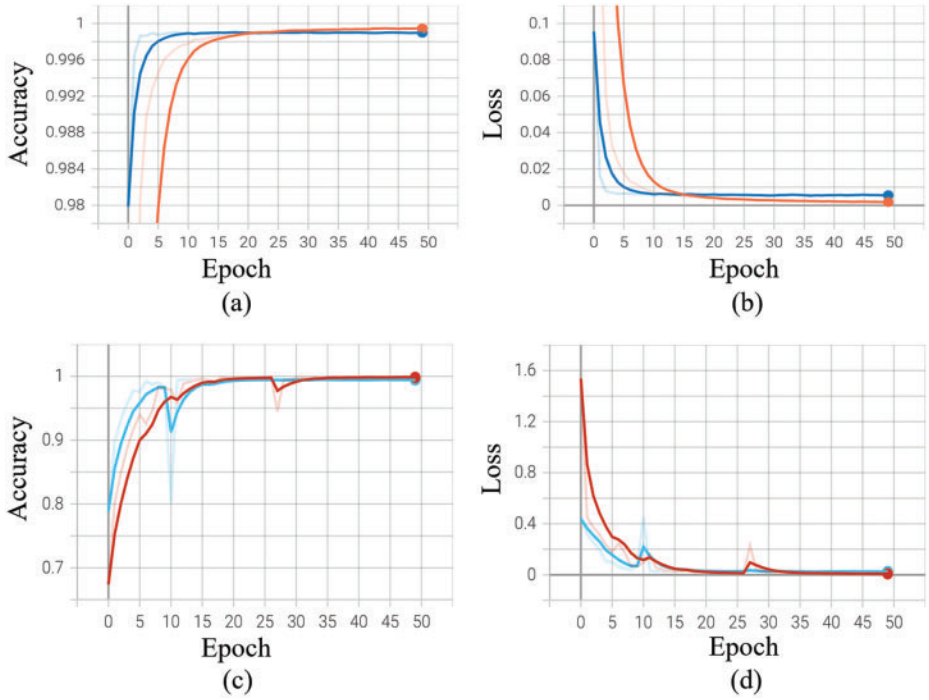

Figure 6: The training accuracy and loss graphs for the ADSV method using UCF-Crime Dataset are plotted in (a) and (b), while using the Avenue dataset the proposed method generates training accuracy and loss graphs as shown in (c) and (d) respectively

Table 7: Comparison of the parameters, model size, and computational complexity of the proposed approach with the state-of-the-art

Method	Number of parameters (million)	Size of model (MB)	Time complexity/Sequence (s)
C3D	–	313	–
VGG-19+multi-layer BLSTM	143	605.5	0.22
Inception-V3+multi-layer BLSTM	23	148.5	–
ResNet-50+multi-layer BLSTM	25	143	0.20
Proposed ADSV method	14.17	54.1	0.26

Table 8: False alarm rate comparison between the proposed ADSV method and recent existing approaches

Method	UCF-Crime Dataset [12]	Avenue Dataset [11]
MIL-C3D-with-constraints [12]	1.9	–
Hasan et al. [35]	27.2	–

(Continued)

Table 8: Continued

Method	UCF-Crime Dataset [12]	Avenue Dataset [11]
Lu et al. SCL [11]	3.1	–
C3D [6]	2.8	–
TSN-RGB [6]	0.1	–
TSN-flow [6]	1.1	–
Proposed ADSV	0.07	0.03

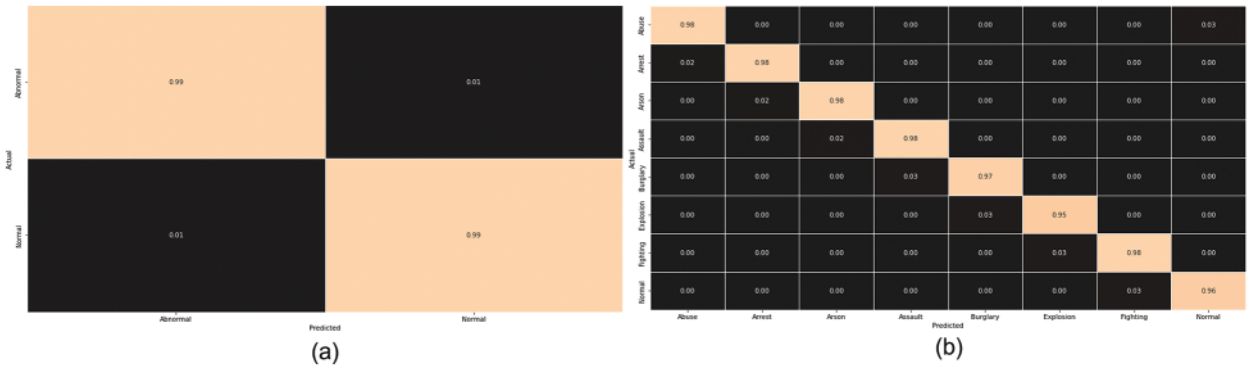

Figure 7: The confusion matrix of the proposed model using both the Avenue and UCF-Crime Dataset are plotted in (a) and (b) respectively

Table 9: The training parameter of the proposed method

Model's training parameters	
Batch size	32
Epochs	50
Step size	1200
Initial learning rate	0.001
Optimizer	Stochastic Gradient Descent (SGD)
Loss	Categorical crossentropy
Training time per step	267 millisecond/step
Training time per epoch	320 s
Total training time	266.7 min

5 Conclusion and Future Direction

Intelligent surveillance systems are popular among computer vision specialists, primarily for security objectives. They assist in rapid response and countermeasures to any anomalous events occurring in a surveillance scene. However, such algorithms are data-hungry and require robust processing systems for effective and efficient analysis. This research utilized several benchmark anomaly detection datasets and presented the ADSV method for video abnormality detection with state-of-the-art performance. In brief, the technique of anomaly detection in surveillance video

consists of three main components: Shot segmentation, feature extraction, sequence learning, and abnormality classifications. To begin with, a shots boundary detection technique is used to segment prominent frames. Furthermore, The LWCNN model receives the segmented frames to extract spatial and temporal information from the intermediate layer. Following that, spatial and temporal features are learned using LSTM cells and Attention Network from a series of frames for each sample of abnormal activity. To detect motion and action, the LWCNN received chronologically sorted frames. Finally, the abnormal activities in video shots are identified using the proposed trained ADSV model. Several evaluation criteria were used to validate our ADSV framework. It has been demonstrated that the ADSV method is more accurate than recently published approaches. The experimental research results show that the suggested ADSV approach significantly lowers false alarm rates when compared to the anomaly detection literature and increases accuracy for the Avenue and UCF-Crime datasets by 12.9% and 14.88%, respectively. However, there is potential for improvement in our proposed system's real-time precision and efficacy. We intend to include channel and temporal Attention-based deep learning networks in the future to improve the precision and effectiveness of the currently proposed ADSV architecture.

Funding Statement: This research was supported by the Chung-Ang University Research Scholarship Grants in 2021 and the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports, and Tourism in 2022 (Project Name: Development of Digital Quarantine and Operation Technologies for Creation of Safe Viewing Environment in Cultural Facilities, Project Number: R2021040028, Contribution Rate: 100%).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. L. Piza, B. C. Welsh, D. P. Farrington, A. L. J. C. Thomas and P. Policy, "CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis," *Criminology & Public Policy*, vol. 18, no. 1, pp. 135–159, 2019.

[2] J. J. P. Suarez and P. C. J. A. P. A. Naval Jr, "A survey on deep learning techniques for video anomaly detection," ArXiv Preprint ArXiv:2009.14146, vol. 2, 2020.

[3] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan and S. W. J. S. Baik, "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors*, vol. 21, no. 8, pp. 2811, 2021.

[4] Y. Cong, J. Yuan and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR 2011*, Colorado Springs, CO, USA, pp. 3449–3456, 2011.

[5] T. Wang, M. Qiao, A. Zhu, Y. Niu, C. Li *et al.,* "Abnormal event detection via covariance matrix for optical flow based feature," *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 17375–17395, 2018.

[6] J. -X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li *et al.,* "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 1237–1246, 2019.

[7] K. -W. Cheng, Y. -T. Chen, W. -H. J. M. T. Fang and Applications, "An efficient subsequence search for video anomaly detection and localization," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 15101–15122, 2016.

[8] D. Xu, Y. Yan, E. Ricci, N. J. C. V. Sebe and I. Understanding, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 100, no. 156, pp. 117–127, 2017.

[9] T. Zhang, W. Jia, B. Yang, J. Yang, X. He *et al.,* "MoWLD: A robust motion image descriptor for violence detection," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 1419–1438, 2017.

[10] C. He, J. Shao, J. J. M. T. Sun and Applications, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29573–29588, 2018.

[11] C. Lu, J. Shi and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Sydney, NSW, Australia, pp. 2720–2727, 2013.

[12] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6479–6488, 2018.

[13] J. Huo, Y. Gao, W. Yang and H. Yin, "Abnormal event detection via multi-instance dictionary learning," in *Int. Conf. on Intelligent Data Engineering and Automated Learning*, Natal, Brazil, pp. 76–83, 2012.

[14] Y. Zhu and S. J. A. P. A. Newsam, "Motion-aware feature for improved video anomaly detection," ArXiv Preprint ArXiv:1907.10211, 2019.

[15] D. Zhang, D. Gatica-Perez, S. Bengio and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, pp. 611–618, 2005.

[16] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 935–942, 2009.

[17] A. K. M. Nor, S. R. Pedapati, M. Muhammad and V. J. M. Leiva, "Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data," *Mathematics*, vol. 10, no. 4, pp. 554, 2022.

[18] P. Shukla, S. Nasrin, N. Darabi, W. Gomes, A. R. Trivedi *et al.,* "Compute-in-memory with monte-carlo dropouts for Bayesian edge intelligence," ArXiv Preprint ArXiv:2111.07125, 2021.

[19] M. Ullah, M. Mudassar Yamin, A. Mohammed, S. Daud Khan, H. Ullah *et al.,* "Attention-based LSTM network for action recognition in sports," *Electronic Imaging*, vol. 2021, no. 6, pp. 302–1, 2021.

[20] L. Selicato, F. Esposito, G. Gargano, M. C. Vegliante, G. Opinto *et al.,* "A new ensemble method for detecting anomalies in gene expression matrices," *Mathematics*, vol. 9, no. 8, pp. 882, 2021.

[21] H. Riaz, M. Uzair, H. Ullah and M. Ullah, "Anomalous human action detection using a cascade of deep learning models," in *2021 9th European Workshop on Visual Information Processing (EUVIP)*, Paris, France, pp. 1–5, 2021.

[22] B. Zhao, L. Fei-Fei and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR 2011*, Colorado Springs, CO, USA, pp. 3313–3320, 2011.

[23] W. Luo, W. Liu and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *2017 IEEE Int. Conf. on Multimedia and Expo (ICME)*, Hong Kong, China, pp. 439–444, 2017.

[24] W. Liu, W. Luo, D. Lian and S. Gao, "Future frame prediction for anomaly detection a new baseline," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6536–6545, 2018.

[25] Y. Chang, Z. Tu, W. Xie and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *European Conf. on Computer Vision*, Glasgow, United Kingdom, pp. 329–345, 2020.

[26] M. Sabokrou, M. Khalooei, M. Fathy and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3379–3388, 2018.

[27] D. Tomar and S. Agarwal, "Multiple instance learning based on twin support vector machine," *Advances in Computer and Computational Sciences*, vol. 1, no. 553, pp. 497–507, 2017.

[28] F. Landi, C. Snoek and R. J. A. P. A. Cucchiara, "Anomaly locality in video surveillance," ArXiv:1901.10364, vol. 1, 2019.

[29] E. Javaheri, V. Kumala, A. Javaheri, R. Rawassizadeh, J. Lubritz *et al.,* "Quantifying mechanical properties of automotive steels with deep learning based computer vision algorithms," *Metals*, vol. 10, no. 2, pp. 163, 2020.

[30] G. I. Rathod, D. A. J. I. J. O. E. T. Nikam and A. Engineering, "An algorithm for shot boundary detection and key frame extraction using histogram difference," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 8, pp. 155–163, 2013.

[31] R. Nawaratne, D. Alahakoon, D. De Silva and X. J. I. T. O. I. I. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

[32] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen *et al.,* "Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New Orleans, Louisiana, USA, pp. 1703–1710, 2018.

[33] D. Bahdanau, K. Cho and Y. J. A. P. A. Bengio, "Neural machine translation by jointly learning to align and translate," ArXiv Preprint ArXiv:1409.0473, 2014.

[34] L. McInnes, J. Healy and J. J. A. P. A. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," ArXiv Preprint ArXiv:1802.03426, vol. 3, no. 29, 2018.

[35] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, pp. 733–742, 2016.

[36] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Advances in Neural Network—ISSN 2017*, 2nd ed., Muran, Hokkaido, Japan: Springer International Publishing, pp. 189–196, 2017.

[37] R. Tudor Ionescu, S. Smeureanu, B. Alexe and M. Popescu, "Unmasking the abnormal events in video," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2895–2903, 2017.

[38] W. Luo, W. Liu and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 341–349, 2017.

[39] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu *et al.,* "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019.

[40] R. Hinami, T. Mei and S. I. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 3619–3627, 2017.

[41] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad *et al.,* "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.

[42] S. Ul Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji *et al.,* "EADN: An efficient deep learning model for anomaly detection in videos," *Mathematics*, vol. 10, no. 9, pp. 1555, 2022.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, pp. 2818–2826, 2016.

[44] U. Gianchandani, P. Tirupattur and M. Shah, "Weakly-supervised spatiotemporal anomaly detection," University of Central Florida Center for Research in Computer Vision REU, 2019.