Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19116-9

Deep learning based anomaly detection in real-time video

Ahmed Elmetwally[1,2] · **Reem Eldeeb**[1] · **Samir Elmougy**[1]

Received: 25 February 2023 / Revised: 15 March 2024 / Accepted: 22 March 2024

Abstract
Many security cameras have been put up in places like airports, roads, and banks for the safety of these public places. These cameras make a lot of video data, and most security camera recordings are only ever seen when something strange happens. This means that monitoring has to be done by people, which is time-consuming and often wrong, so automatic ways of monitoring have to be used. In this paper, we propose a system that automatically detects irregular events in videos based on the integration of Inflated 3D Convolution Network (I3D-ResNet50) and deep Multiple Instance Learning (MIL). This system considers both regular and unusual videos as negative and positive packets, respectively. Each video snippet is a case of that packet. An anomaly score is generated for each video snippet using a fully connected Neural Network (NN). After processing videos, we used an I3D-ResNet50 to extract features after applying 10-crop augmentations to the UCF-101 dataset that contains 130 GB of videos with 13 abnormal events such as fighting, stealing, abuse, etc., as well as normal events. Our experimental results show that the AUC is 82.85% with only 10,000 iterations compared with other approaches. This means that our model is better at spotting anomalies in real-time videos.

Keywords Anomaly Detection · I3D-ResNet50 · Multiple Instance Learning · Deep Learning · Video Processing

Ahmed Elmetwally, Reem Eldeeb, Samir Elmougy These authors are contributed equally to this work.

✉ Ahmed Elmetwally
ahmedeldemoksy@std.mans.edu.eg

Reem Eldeeb
Reemm_db@mans.edu.eg

Samir Elmougy
mougy@mans.edu.eg

[1] Department of Computer Science, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt

[2] Department of Computer Science, Misr Higher Institute for Commerce and Computers, Mansoura 35511, Egypt

1 Introduction

Recently, intelligent systems have played an important role in our lives. The digital transformation will be in everything because of the rapid advancement in Artificial Intelligence (AI) and its applications in different fields such as Computer Vision (CV), smart agriculture, physics, drug discovery, social network analysis, security, etc. An anomaly is an emergency or an event that is out of the ordinary or standard.

Anomaly Detection (AD) in video surveillance, which includes fighting, stealing, and robbery, among other crimes, is drawing an attention from CV researchers in real-world surveillance scenarios [1–3]. Humans now spend a lot of time looking at monitors to see if there are any unusual events that need to be quickly handled. This is time-consuming and may result in labelling errors due to exhaustion [4].

AD is considered an important phase in any system because of the need for a prospective device in many branches, such as automation for human behaviour characterization, human-computer interaction, and video surveillance structures. For example, smart cities need a safe way to keep track of everything, like people, cars, buses, and traffic, in case something goes wrong [5]. In the CV sector [3], AD in videos is considered a big challenge. Changes to Deep Learning (DL) models that have already been trained have a big impact on increasing the performance of AD systems.

Using a sparse coding AD method, which is based on a dictionary that is aware of normal events only, isn't very effective at determining what's wrong [6]. This leads to a lot of false warnings. Also, it's hard to spot strange behaviour in surveillance because there isn't a lot of annotated data, street cameras usually have a lower resolution, and there is a lot of image change within and between classes.

The key challenges for AD in surveillance videos could be summarized as follows [7]: (i) Anomaly events typically account for only a small fraction of a video, necessitating the removal of a vast amount of irrelevant data. It makes it harder to test how well an algorithm works and how well a device can compute, and it affects how well classifiers in models work. It also makes it difficult to deliver accurate detection results when the anomaly video is close to the normal video. (ii) Since anomalies can vary greatly from one another, it can be challenging to create features and analyse such massive amounts of video. (iii) Video-based tasks compared with image-based tasks are more difficult [8]. Besides the spatial information that both images and videos carry, such as the grey-scale histogram and RGB information, the management of temporal information should be included in video handling approaches.

In surveillance systems, enormous amounts of video data are always analysed using AD techniques. During the last ten years, there have been many different ways to deal with this important issue. This paper proposes an automated system that uses Inflated 3D Convolution Network (I3D-ResNet50) and deep Multiple Instance Learning (MIL) to find anomalous events in videos. Normal and abnormal videos are considered positive and negative packets, in which each video snippet is an case of that packet, which then predicts the score of each video snippet and applies a deep MIL ranking loss.

We introduce a weak-supervision enhancement strategy by presenting a new ranking function that uses deep MIL and I3D-ResNet50 to automatically spot strange things happening in videos that are used in both indoor and outdoor surveillance networks. This suggests that in practical applications, our system can achieve good real-time detection. A benchmark dataset, UCF-Crime [9], is used to show that our proposed method surpasses

the state-of-the-art in terms of the Area Under Curve (AUC), Receiver Operating Characteristic (ROC) curve, and False Alarm (FA).

The remainder of the paper is organized as follows: Section 2 introduces the related works. The proposed approach is described in depth in Section 3. Section 4 describes the experimental design and analysis, as well as the outcomes of our proposed framework, which are compared to those elicited by state-of-the-art methods. Section 5 concludes the paper.

2 Related Work

AD in surveillance videos has been studied for a long time, in which it is a challenge to solve because of the variances in visual features and changes between and within classes. This section explores the three well-known AD methods for videos that can be used in surveillance networks.

Sparse coding anomaly detection. Techniques based on sparse coding [10–14] have been suggested for the detection of abnormal behaviours. They trained a global dictionary using a low-level feature from the normal training samples. Only regular videos were used to train the network, and only the features that had been generated from the regular videos were used to generate the dictionary. During the testing phase of making the dictionary, a high reconstruction error shows that there are events that don't fit the pattern. These methods often give false alerts because it's hard to list all the different kinds of regular videos into one category.

Unsupervised anomaly detection. Some methods of data reconstruction use generative models to find examples of regular samples while minimizing the amount of error in the reconstruction [15–23]. These approaches make the assumption that unseen abnormal videos or images are frequently difficult to reconstruct accurately and treat examples with significant reconstruction mistakes as anomalies. However these methods can overfit the training data and can't tell the difference between normal and unusual events because they don't know what's unusual.

Weakly supervised anomaly detection. Compared to unsupervised methods, using some labelled anomalous examples has significantly enhanced performance[24–36]. However, it is very expensive to annotate a lot of frames at the frame level. So, the current AD methods are based on training with less supervision that uses less expensive video level annotations [24]. We are motivated by the advances in weakly supervised AD, so the following will be discussed in detail.

Sultani et al. [9] automated the learning of a deep anomaly ranking model using MIL, which forecasts high anomaly scores for abnormal video snippets. The MIL framework is the basic foundation for weakly-supervised video AD techniques. Deep Convolutional 3-Dimensional (C3D) features have been used for each bag containing 32 video snippets. They also present UCF-Crimes, a brand-new, massive dataset that is the first of its type and has 128 hours of recordings.It demonstrates that their suggested solution surpasses standard methods by a score of 75.41%.

Ullah et al. [1] presented a model that is based on Convolutional Neural Network (CNN) and can be used to get spatiotemporal features from video frames. These features can then be sent to a Multilayer Bi-Directional Long-Short-Term Memory (BDLSTM) model to classify events as either abnormal or normal. The research shows that the UCF-Crime and UCFCrime2Local datasets are 3.41% and 8.09% more accurate than cutting-edge methods.

Zhong et al. [37] came up with a cascade model that uses a Graph CNN (GCNN) to get rid of noisy labels before pixel reconstruction so that optical flow can be predicted. In order to choose the optimal model for AD, a generalisation ability evaluation based on pseudo-anomaly was also proposed. This evaluation gauges a model's capacity to represent anomalies. The selected model receives an AUC of 88.9% on Avenue, 82.6% on Ped1, 97.7% on Ped2, and 70.7% on ShanghaiTech datasets. The efficiency of our technology has been confirmed through extensive ablation trials.

Tian et al. [24] came up with the Robust Temporal Feature Magnitude (RTFM) model to tell the difference between aberrant and regular snippets in videos with weak labels. I3D or C3D is used for extracting features to train the snippet classifier. In tests on three data sets (UCF-Crime, UCSD-Peds, and XD-Violence), the RTFM model outperforms many other current methods. It also has a better ability to pick out subtle anomalies and get a lot of samples than many of the other methods.

Yan et al. [38] proposed a weakly supervised framework for spatiotemporal collaboration in video segmentation, named STC-Seg. They use images from optical flow and unsupervised depth estimation that work well together to create good fake labels for deep network training. in addition, a puzzle loss was created to enhance mask generation and enable end-to-end training with box-level annotations. Their approach is flexible enough to let image-level instance segmentation algorithms handle the video-level work. The KITTI MOTS and YouTube visualization datasets are used in their investigations.

3 The proposed approach

This section presents in detail the proposed approach that is composed of three phases: the video preprocessing phase; feature extraction and generation of anomaly score phase; and the MIL phase. Figure 1 shows the proposed real time AD approach. These three phases are defined as follows:

- **Video Preprocessing:** Each training video is divided into a fixed number of temporal snippets. Every 32 snippets makes a positive packet if it contains an anomaly, else it makes a negative packet. Each snippet is considered an case of a packet.
- **Feature Extraction:** We used I3D-ResNet50 [39], which was trained on Kinetics dataset [40], to extract spatial-temporal features from snippets after increasing the amount of data and reducing overfitting by applying 10-crop augmentation. The extracted features are fed into an Fully Connected Neural Network (FCNN), which generates the score of anomaly for every video snippet.
- **Multiple Instance Learning (MIL):** The network is trained using the deep MIL [41, 42], and the suggested ranking loss function. Weakly supervised training methods are used for the network. The following subsections contain a detailed description of each stage.

3.1 Video preprocessing

This subsection is designed to illustrate the video preprocessing. In this research, the UCF-Crime dataset [9] is used. It has videos with fixed parameters of 240*320 pixels and 30fps. As shown in Fig. 1, each training video was separated into 32 non-overlapping temporal snippets. Then each of the 32 pieces are put together into a positive or negative packet based

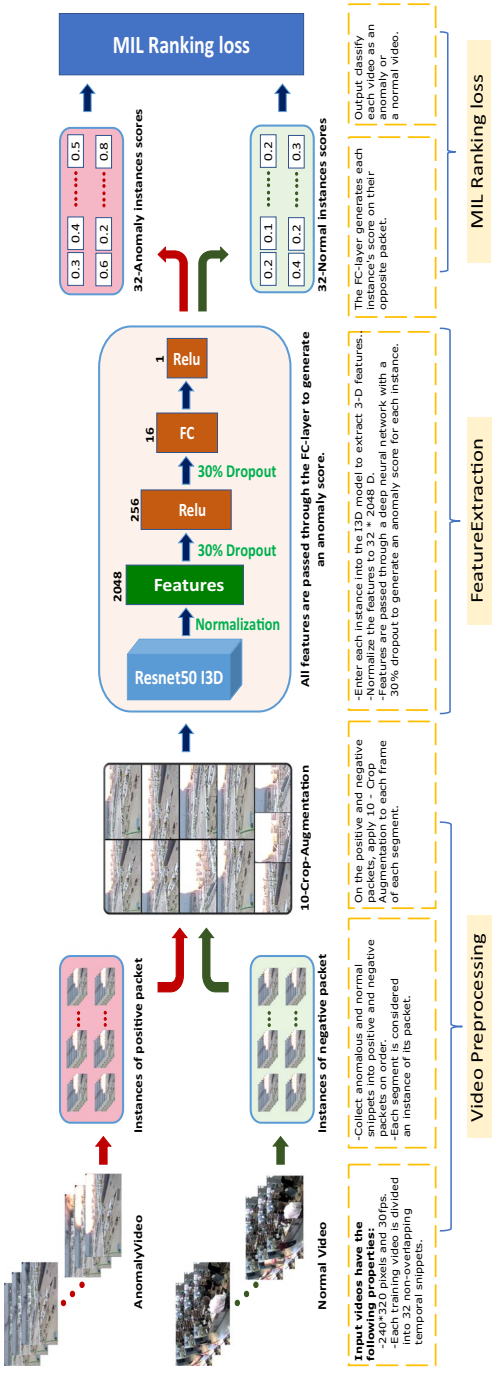


Fig. 1 The proposed Real Time Anomaly Detection (RTAD)

on whether or not it had any oddities. A positive video has one anomaly, whereas a negative video does not have. Thereafter, we express a positive movie as a positive packet P_a, in which different temporal snippets create temporal cases in the packet $(a^1, a^2, a^3.....a^k)$, where k represents the number of cases in the packet. We assume that the anomaly exists in at least one of these cases. A negative packet P_n , is used to represent negative video, with temporal segments of this packet forming negative cases$(n^1, n^2, n^3.....n^k)$. There is no anomaly in any of the cases in the negative packet.

3.2 Feature extraction

This subsection is designed to illustrate the feature extraction phase and generate an anomaly score for video snippets. Pretrained model I3D-ResNet50 was trained on the Kinetics dataset [43], and is based on 2D-ConvNet inflation, which involves expanding the filters and pooling kernels of very deep image classification convNets into 3D as in [39]. I3D-ResNet50 is an efficient extractor of temporary-spatial features for video frames. Furthermore, we are assured in our utilisation 3D ResNet. According to [44, 45], 3D-ResNet does better than all other algorithms in the AD challenge. After applying 10-crop augmentation, the average of all 16-frame clip features inside a video snippet is used to calculate the features for that snippet.This is followed by l_2 normalization.

These features with dimensions 2048D are passed into three layers of FCNN, as shown in Fig. 1, in which the first layer is made up of 256 units, the second layer is made up of 16 units, and the last layer is made up of one unit. In between FCNN layers, 30% dropout regularization is determined in [46].

3.3 Multiple instance learning (MIL)

To make a strong classifier, we need accurate annotations of both anomaly and normal data. Each video snippet's temporal annotations are required by a classifier in the context of supervised AD. On the other hand, obtaining temporal annotations for videos takes a lot of effort and time. The need for precise temporal annotations is diminished by MIL. The exact temporal place of anomalous events in videos is ambiguous in MIL. To determine whether a video abnormality is present, only video-level labels are necessary. We trained the network in a weakly supervised way using deep MIL in the same way that mentioned in [9]. The ranking loss function may be used in MIL to train the network by giving video-level labels. As discussed in Section 3.1, we have collected all cases.

Where P_a stands for the group of anomaly cases, since at least one of them has some kind of strange behavior, and an P_n is used to symbolise the collection of typical situations.

The symbol t denotes the total number of cases in both collections. Since we don't know exactly what the positive cases are, we can optimize the objective function [42] based on the highest-scoring case in each packet, as shown below:

$$\min \frac{1}{m} \sum_{j=1}^{m} \overbrace{\max(0, 1 - Y_{P_j}(\max_{i \in P_j}(w.\phi(x_i)) - b))}^{A} + \frac{1}{2}\|w\|^2 \tag{1}$$

where A represent the hinge loss function, Y_{P_j} stands for each packet label, $\phi(X)$ represents the snippet's features, b stands for bias and m denotes the overall quantity of training samples.

3.4 Deep multiple instance learning (DMIL)

It can be difficult to precisely identify abnormal behaviour [47], because it varies considerably from person to person and is highly subjective. AD is often viewed as low-likelihood pattern identification rather than classification due to a lack of sufficient examples of abnormal behaviour. In the proposed approach, AD is given as a regression issue. The abnormal video snippets should have greater anomaly scores than the regular video portions. Therefore, ranking loss could be a method of teaching our model to give abnormal video snippets more ratings than normal ones, like:

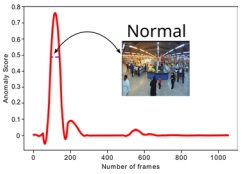
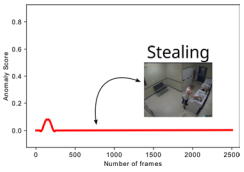
$$f(I_a) > f(I_n) \tag{2}$$

where I_a stands for anomalous video, and I_n stands for normal video, $f(I_a)$ and $f(I_n)$ stands for predicted anomaly scores, which range from 0 to 1, as appropriate.

If the training is conducted with knowledge of the snippet-level annotations, the previously proposed ranking method should function effectively. However, Equation (2) cannot be used in the absence of video snippet level annotations. So, we used a ranking loss function to train our model using MIL [48]. Before we can use the loss function in our task, we need to look at the possible false-warning situations shown in Table 1.

1. The first case (case 1) is a false anomaly warning, which happens when our model Predicts that a normal event will turn out to be abnormal.
2. The second case (case 2) is a false normal warning, which happens when our model predicts that an abnormal event will turn out to be normal.

Table 1 Examples of false warning cases

No	Cases	Description	Example
1	**False warning**	**Predicts that a normal event will turn out to be abnormal.**	
2	**False warning**	**predicts that an abnormal event will turn out to be normal.**	

The following objective function for multiple instance ranking is used to reduce all types of false alarms.

$$\max_{i \in P_a} f(I_a^i) > \max_{i \in P_n} f(I_n^i) \tag{3}$$

$$\max_{i \in P_a} f(I_a^i) > \min_{i \in Pa} f(I_a^i) \tag{4}$$

Equation (3) compares the highest-ranking examples from each packet [9], where the highest-ranking case from the positive packet is most likely a real positive and the highest-ranking case from the negative packet may possibly be a fake positive.

In Equation (4) The highest-ranking positive case is compared to the lowest-ranking positive case in the anomalous packet [48], where the highest-ranking case from the anomaly packet is more likely to be a real anomaly and the lowest-ranking case from the anomaly packet might be a false positive.

Equations (3) and (4) are used to avoid false warnings when max is applied to each packet's video cases. We only rank the two cases that have the highest anomaly scores in the anomaly and normal packets, instead of ranking every case in the packet and the two cases with the greatest and lowest anomaly scores in the anomaly packets. The highest anomaly score for the snippet in the anomaly packet is probably the actual anomaly case (anomalous snippet). The highest anomaly score for the snippet in the normal packet is that it appears the most like an anomalous snippet despite being a normal case. This normal case is seen as a difficult case that could result in a false warning in AD. We attempt to make the positive and negative events have significantly different anomaly ratings using Equation (3) and Equation (4). Thus, ranking loss using the hinge-loss algorithm is given by:

$$l(P_a,\ P_n) = l_1(P_a,\ P_n) + l_2(P_a,\ P_a) \tag{5}$$

where $l_1(P_a,\ P_n),\ l_2(P_a,\ P_a)$ are defined as below:

$$l_1(P_a,\ P_n) = \max(0, 1 - \max_{i \in P_a} f(\mathrm{I}_a^i) + \max_{i \in P_n} f(I_n^i)) \tag{6}$$

$$l_2(P_a,\ P_a) = \max(0, 1 - \max_{i \in P_a} f(\mathrm{I}_a^i) + \min_{i \in P_a} f(I_a^i)) \tag{7}$$

Similar to the study in [9], the loss function for keeping the smoothness and sparsity of the abnormality score over time is:

$$l(P_a, P_n) = l_1(P_a, P_n) + l_2(P_a, P_a) + \mu_1 \overbrace{\sum_{i}^{(n-1)} (f(I_a^i) - f(I_a^{i+1}))^2}^{A} + \mu_2 \overbrace{\sum_{i}^{n} f(I_a^i)}^{B} \tag{8}$$

where A is the term for temporal smoothness and B is the term for sparsity.

The errors from the video snippets with the highest scores are received by both positive and negative packets in the used MIL ranking loss. By training on many anomalous and normal packets, we anticipate that the network will create a generalised model capable of forecasting high scores for anomalous snippets in positive packets. The final objective function is provided by:

$$l(W) = l(p_a, p_n) + \mu_3 \|W\| \tag{9}$$

where W stands for the weights of the model and μ_1, μ_2 and μ_3 represent hyper-parameters of the model.

4 Experimental results: Discussion and analysis

4.1 DataSet

There are several standard datasets available for the AD task. To test the performance of the proposed approach, we conducted extensive experiments on the most popular balanced AD datasets, namely UCF-Crime. It was created by Sultani et al. [9] and has a range of scene distributions for AD problems and captures a wide range of actual anomalies such as robbery, fighting, burglary, and so on. It's a sizable dataset, with 128 hours of video broken down into 290 testing movies and 1610 training videos, as seen in Table 2. Only video-level labels have been added to the dataset. Compared to other existing AD datasets, its size is noticeably larger, and its dataset is considerably more complex. The diversity of the inner-class population, the wide range of backgrounds, perspectives, and lighting makes it challenging.

4.2 Implementation details

Training Phase: There are 32 distinct, non-overlapping snippets taken from each video, each of which serves as a case in a packet. The amount of snippets (32) was determined via experiment. I3D-ResNet50 [39] is used to extract visual features from each video frame, which has a size of 240*320 pixels and a frame rate of 30 fps. After performing 10-crop augmentation, we generate I3D features for each 16-frame video snippets, which is followed by l_2 normalization.

The features for a video segment are calculated by taking the average of all the 16-frame clip features within that snippet. Three FCNN-layers received these features (2048D) as input. There are 256 units in the first layer, 16 units in the second layer, and one unit in the final layer. A dropout regularisation of 30% is employed between FCNN-layers [46]. ReLU activation function [49] is used for the first and last FCNN. As a minibatch, we randomly selected 30 positive and 30 negative packets.

Testing Phase: Each test video is resized to 240*320 pixels with 30 fps, then separated into individual 32 non-overlapping video snippets during the testing phase. Then, we pass the features of each video snippet through our suggested FCNN to get its "anomaly score" as shown in Fig. 2.

All experiments are conducted on a PC that has the characteristics listed in Table 3.

Table 2 UCF-Crime dataset details

	Total	**Train (85%)**	**Test (15%)**
Anomaly	950	810 (85%)	140 (15%)
Normal	950	800 (84%)	150 (16%)
Total	**1900**	**1610**	**290**

Table 3 PC characteristics

#	Pc characteristics	Values
Hardware		
1	**Processor**	Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz (8 CPUs),1.8GHz
2	**Memory**	8192 MB
3	**Display Devices**	Intel(R) HD Graphics 4600
Software		
4	**Operating System**	Windows 10

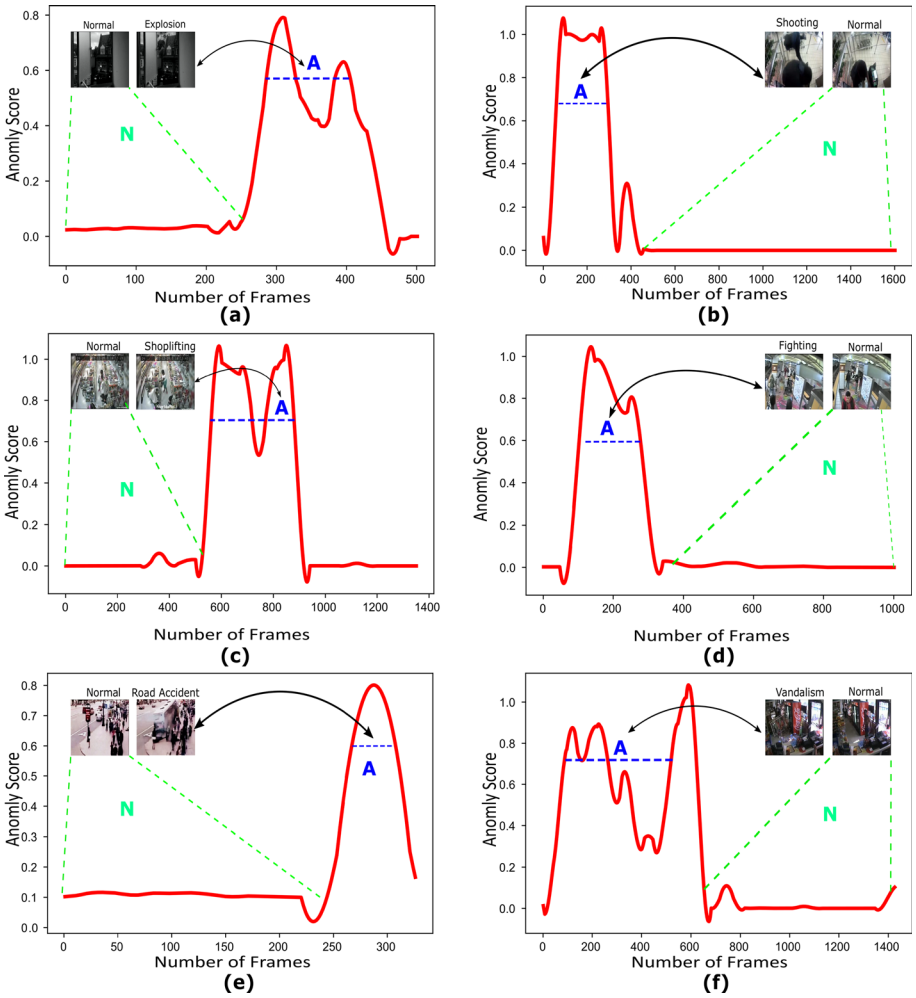


Fig. 2 Visualization of testing results on UCF-Crime

4.3 Evaluation metrics

As Follow the previous work on AD [9, 50]. We use a frame-based Receiver Operating Characteristic (ROC) curve, Area Under Curve (AUC), and False Alarm (FA) to figure out how well our approach works. More information is given about FA in Section 3.4.

4.4 Parameters setting

Setting values for parameters is thought to be the most important part because it directly affects how accurate the model is. As shown in Table 4, the numerical values for each parameter are set based on different experiments. The features were sent to three FCNN-layers. The first layer has 256 units, the second layer has 16 units, and the third layer has one unit. A dropout regularisation of 30% is used in between FCNN layers [46]. ReLU [46] activation function is used for the first and last layers. We choose 30 positive and 30 negative packets at random as a minibatch. Using Theano [51], with the Adagrad optimizer, we trained our suggested model with a starting learning rate of 0.001. All hyper-parameter values are set as in [9]: $\mu_1 = \mu_2 = 8 * 10^{-5} and\ \mu_3 = 0.01$.

The network is trained for 10000 epochs using the deep MIL and the proposed ranking loss function, described in Equation (9).

4.5 Comparison with the State-of-the-art

To evaluate the proposed approach and assess its efficiency, we compare it with state-of-the-art methods as shown in Table 5, including the recently published contributions. M. Hasan et al. [52], C. Lu et al. [10], W. Sultani et al. [9], Zhong et al. [53], Zaheer et al. [54], Kamoona et al. [55], and SVM binary classifier [56].

It could be noticed that our proposed approach passes the state-of-the-art methods in terms of AUC and FA, with values of 82.85 and 0.2, respectively. The diff ratio column is the difference between the proposed approach and the state-of-the-art methods. Notably, the proposed approach has the benefit of being able to analyze a 16-frame video clip in only 0.76 seconds. The data indicate that the proposed approach has the potential to effectively perform real-time detection in practical scenarios. Based on the reproducible code that is available, the AUC curves for our method and the state-of-the-art approaches were also plotted as shown in Fig. 3.

Based on the previous results, we can say that the video processing phase starts by dividing each video into a group of snippets and then collecting all 32 snippets into packages. This had the effect of accelerating the speed of the video classification process. The process of increasing the features of the single clip through 10-crop augmentation before extracting the features with the use of the I3D-Resnet50 pretrained model had the effect of improving the extracted features used to generate high scores for abnormal snippets, which are used in training the FCNN. We can not deny the role of the adagrad optimizer and the activation function Relu in improving the score of each video clip. The 40% dropout had the effect of identifying the most important features used in the training process and avoiding the overfitting problem. We also avoided false warnings, which happens when our model predicts that a normal event will turn out to be abnormal by comparing the highest-ranking examples from each packet, as described in Equation (3), where the highest-ranking case from the positive packet is most likely

Table 4 Model's parameters selection

Experiment#	Input Video Properties	Number of Segments	Features	Normalization	Sparsity and Smoothness	Model's Parameters Input Dimensions	Number of Units and Activation Function First Layer	Second Layer	Last Layer	Drop out	Optimizer	Mini Batch	Noofiterations	Total Time	AUC
1	240*320	32	C3D-FC6	EQ1	EQ2	32*4096	512-Tanh	32	1-Tanh	60%	Adagrad	30	100	0:33:47.817865	73.99
2	pixels									40%			200	1:03:30.414635	74.70
3	and									40%			500	2:40:14.528501	75.22
4	30fps						512-Relu		1-Sig	20%			200	1:06:14.996219	77.48
5			I3D		32*2048		256-Sigmoid		1-Relu	10%			500	1:37:23.939074	78.57
6							256-Tanh		1-Tanh	40%			500	1:57:45.022679	79.14
7							256-Relu	16	1-Sigmoid	40%			1000	3:24:52.304633	80.12
8			I3D with 10-Crop Aug				256-Relu	16	1-Relu	30%			1000	26:13:46.941854	82.85

Table 5 The AUC comparison for various methods on the UCF-Crime dataset

Methods	Features	AUC	Diff Ratio	False Alarm
Binary classifier	-	49.99	32.86	-
Hassan et al. [52]	-	50.66	32.19	27.20
Lu et al. [10]	-	65.51	17.24	3.10
Sultani et al. [9]	C3D	75.41	7.44	1.9
Zhong et al. [53]	C3D	81.08	1.77	2.8
Zaheer et al. [54]	ResNext	79.84	3.01	-
Kamoona et al. [55]	C3D	79.49	3.36	0.5
Ours	I3D	82.85	-	0.2

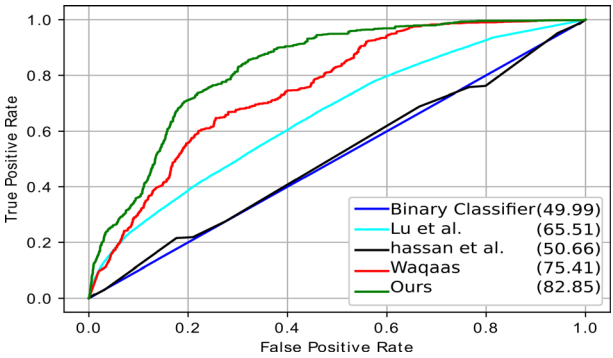


Fig. 3 The ROC Values for various methods on the UCF-Crime dataset

a real anomaly while the highest-ranking case from the normal packet may possibly be a fake anomaly. We also avoided the false warning when our model predicts an abnormal event will turn out to be normal by comparing the maximum ranked case from the anomaly packet with the minimum ranked case from the anomaly packet, as described in Equation (4), where the case with the lowest rank in the anomaly packet could be a false positive and the case with the highest rank in the anomaly packet is most likely to be a true positive. We also keep the smoothness and sparsity of the abnormality score over time, as in the study by Sultani et al. [9], as described in Equation (8). For comparison, we revisited most of the detection studies,and we used performance metrics like AUC and ROC on the benchmark dataset to assess the proposed approach. According to our study, the proposed approach could be used in different environments like smart cities, smart universities, smart companies, or generally in any smart-based environment, which could be helpful for detecting any anomaly effectively.

Significantly, the suggested approach has the advantage of processing a 16-frame clip during inference in just 0.76 seconds on an RTX 2080 Ti. This period includes the time required for I3D extraction. These findings indicate that our technology is capable of achieving effective real-time detection in practical scenarios.

5 Conclusion

In this paper, the problem of AD in video surveillance was tackled. We have proposed a deep FCNN for detecting anomalies in video surveillance analysis. In the first phase, the surveillance video, which has properties of 240 * 320 pixels and 30 *fps*, is divided into 32 snippets and fed into I3D-ResNet50 to extract features. Second, the FCNN generates an anomaly score for each snippet. Finally, we used the deep MIL and suggested a novel ranking function. Weakly supervised training methods are used for the network. With an improvement of 7.44% over the second method in the dataset, when compared to state-of-the-art techniques, experimental results of our proposed method conducted over the UCF-Crime dataset demonstrate higher accuracy levels in terms of anomalous event recognition. In the future, meta heuristic algorithms will be used to optimise rank-score functions to make algorithms that work better.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data Availability The supporting data for the study's findings is freely accessible at https://www.dropbox.com/sh/75v5ehq4cdg5g5g/AABvnJSwZI7zXb8_myBA0CLHa?dl=0

Code availability The supporting source code for the study's findings is freely accessible at https://github.com/AhmedEldemoksy/AnomalyDetection

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW (2021) Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks. Multimedia Tools and Applications 80(11):16979–16995
2. Jin P, Mou L, Xia GS, Zhu XX (2022) Anomaly detection in aerial videos with transformers. IEEE Trans Geosci Remote Sens 60:1–13
3. Chen S, Li Z, Tang Z (2020) Relation r-cnn: A graph based relation-aware network for object detection. IEEE Signal Process Lett 27:1680–1684
4. Li N, Zhong JX, Shu X, Guo H (2022) Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning. Neurocomputing
5. Chackravarthy S, Schmitt S, Yang L (2018) Intelligent crime anomaly detection in smart cities using deep learning. In: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp. 399–404. IEEE
6. Ullah W, Ullah A, Hussain T, Muhammad K, Heidari AA, Del Ser J, Baik SW, De Albuquerque VHC (2022) Artificial intelligence of things assisted two-stream neural network for anomaly detection in surveillance big video data. Futur Gener Comput Syst 129:286–297

7. Nayak R, Pati UC, Das SK (2021) A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis Comput 106:104078
8. Yan L, Ma S, Wang Q, Chen Y, Zhang X, Savakis A, Liu D (2022) Video captioning using global-local representation. IEEE Trans Circuits Syst Video Technol 32(10):6642–6656
9. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488
10. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE International Conference on Computer Vision,pp. 2720–2727
11. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 341–349
12. Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: CVPR 2011, pp. 3449–3456. IEEE
13. Mo X, Monga V, Bala R, Fan Z (2013) Adaptive sparse representations for video anomaly detection. IEEE Trans Circuits Syst Video Technol 24(4):631–645
14. Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: CVPR 2011, pp. 3313–3320. IEEE
15. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel Avd (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1705–1714
16. Morais R, Le V, Tran T, Saha B, Mansour M, Venkatesh S (2019) Learning regularity in skeleton trajectories for anomaly detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11996–12004
17. Ionescu RT, Khan FS, Georgescu MI, Shao L (2019) Object-centric auto encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7842–7851
18. Nguyen TN, Meunier J (2019) Anomaly detection in video sequence with appearance-motion correspondence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1273–1283
19. Burlina P, Joshi N, Wang I, et al (2019) Where's wally now? deep generative and discriminative embeddings for novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11507–11516
20. Tudor Ionescu R, Smeureanu S, Alexe B, Popescu M (2017) Unmasking the abnormal events in video. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2895–2903
21. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545
22. Venkataramanan S, Peng KC, Singh RV, Mahalanobis A (2020) Attention guided anomaly localization in images. In: European Conference on Computer Vision, pp. 485–503. Springer
23. Park H, Noh J, Ham B (2020) Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14372–14381
24. Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G (2021) Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4975–4986
25. Liu W, Luo W, Li Z, Zhao P, Gao S, et al (2019) Margin learning embedded prediction for video anomaly detection with a few anomalies. In: IJCAI, pp. 3023–3030
26. Pang G, Shen C, van den Hengel A (2019) Deep anomaly detection with deviation networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 353–362
27. Ruff L, Vandermeulen RA, Görnitz N, Binder A, Müller E, Müller KR, Kloft M (2019) Deep semi-supervised anomaly detection. arXiv:1906.02694
28. Zaheer MZ, Lee Jh, Astrid M, Mahmood A, Lee SI (2021) Cleaning label noise with clusters for minimally supervised anomaly detection. arXiv:2104.14770
29. Wu J, Zhang W, Li G, Wu W, Tan X, Li Y, Ding E, Lin L (2021) Weakly-supervised spatio-temporal anomaly detection in surveillance video. arXiv:2108.03825
30. Lv H, Zhou C, Cui Z, Xu C, Li Y, Yang J (2021) Localizing anomalies from weakly-labeled videos. IEEE Trans Image Process 30:4505–4515
31. Wu P, Liu J (2021) Learning causal temporal relation and feature discrimination for anomaly detection. IEEE Trans Image Process 30:3513–3527

32. Zaheer MZ, Mahmood A, Astrid M, Lee SI (2020) Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: European Conference on Computer Vision, pp. 358–376. Springer
33. Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, Yang Z (2020) Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European Conference on Computer Vision, pp. 322–339. Springer
34. Zaheer MZ, Mahmood A, Shin H, Lee SI (2020) A self-reasoning framework for anomaly detection using video-level labels. IEEE Signal Process Lett 27:1705–1709
35. Feng JC, Hong FT, Zheng WS (2021) Mist: Multiple instance self training framework for video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14009–14018
36. Wan B, Fang Y, Xia X, Mei J (2020) Weakly supervised video anomaly detection via center-guided discriminative learning. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE
37. Zhong Y, Chen X, Jiang J, Ren F (2022) A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos. Pattern Recogn 122:108336
38. Yan L, Wang Q, Ma S, Wang J, Yu C (2022) Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio temporal collaboration. IEEE Trans Circuits Syst Video Technol 33(1):393–406
39. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308
40. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al (2017) The kinetics human action video dataset. arXiv:1705.06950
41. Babenko B (2008) Multiple instance learning: algorithms and applications. View Article PubMed/NCBI Google Scholar 1–19
42. Andrews S, Tsochantaridis I, Hofmann T (2002) Support vector machines for multiple-instance learning. Advances in neural information processing systems 15
43. Zisserman A, Carreira J, Simonyan K, Kay W, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, et al (2017) The kinetics human action video dataset
44. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555
45. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 3154–3160
46. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15(1):1929–1958
47. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM computing surveys (CSUR) 41(3):1–58
48. Dubey S, Boragule A, Jeon M (2019) 3d resnet with ranking loss function for abnormal activity detection in videos. In: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 1–6. IEEE
49. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Icml
50. Li W, Mahadevan V, Vasconcelos N (2013) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32
51. Team TTD, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, et al (2016) Theano: A python framework for fast computation of mathematical expressions. arXiv:1605.02688
52. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–742
53. Zhong JX, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1237–1246
54. Zaheer MZ, Mahmood A, Khan MH, Segu M, Yu F, Lee SI (2022) Generative cooperative learning for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14744–14754
55. Kamoona AM, Gostar AK, Bab-Hadiashar A, Hoseinnezhad R (2023) Multiple instance-based video anomaly detection using deep temporal encoding-decoding. Expert Syst Appl 214:119079

56. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497