Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-14425-x

Robust learning for real-world anomalies in surveillance videos

Aqib Mumtaz[1] · Allah Bux Sargano[1] · Zulfiqar Habib[1]

Received: 26 May 2022 / Revised: 11 October 2022 / Accepted: 21 January 2023

Abstract
Anomaly detection has significant importance for developing autonomous surveillance systems. Real-world anomalous events are far more complex and harder to capture due to diverse human behaviors and a wide range of anomaly types. A key factor in defining activity is the temporal length or duration of the activity. The time period required for an anomalous activity to be completely understandable and meaningful depends on the nature and speed of the event. Some events are as fast to be captured within a few frames; however, some activities are slow and may require several thousands of video frames to define an activity. Deep learning architectures have a limited input temporal sequence length and suffer from learning very long sequences. There is a need to re-investigate the problem from the frame sequences perspective to better define an activity in the limited temporal length. In this research work, our contribution is two-fold. Firstly, a novel strategy of dynamic frame-skipping is proposed for producing meaningful temporal sequences for model learning. Secondly, a new deep learning model based on the Inflated Inception network (I3D) is proposed for learning spatial and temporal information from video frames. In order to evaluate the performance of the proposed model, experiments are performed on one of the most challenging real-world anomalies UCF-Crime dataset. The results confirm that the proposed model is robust and significantly outperforms state-of-the-art methods in terms of accuracy. In addition to this, the proposed model has achieved the highest F1 score for fast and slow activities, such as explosions, road accidents, robbery, and stealing, and the AUC score of 0.837.

Keywords Real-world anomalies · Anomaly detection · Surveillance videos · Deep learning · Inflated inception network · Dynamic frame-skipping

✉ Allah Bux Sargano
allahbux@cuilahore.edu.pk

✉ Zulfiqar Habib
drzhabib@cuilahore.edu.pk

[1] Department of Computer Science, COMSATS University Islamabad, Lahore 54000, Pakistan

1 Introduction

In the modern age, surveillance cameras have a pervasive need to fulfill the objective of surveillance for public security and safety. Hundreds and thousands of surveillance cameras are deployed in cities to ensure public security. The camera stream is manually monitored in the control rooms to monitor and track different abnormalities/anomalies. Anomalies are commonly referred to as the presence of unusual appearances or motion patterns that vary from the normal stream [44]. Video anomaly detection is the process of finding irregular patterns that do not conform to the expected behaviors in video sequences. Videos anomalies include a wide range of activities such as fights, road accidents, abuse, theft, crime scene, robbery, and shooting. These activities occur under different circumstances and infrequently happen compared to normal activities, making it nearly impossible for a human to monitor them 24/7. There is a significant need to develop automated video surveillance systems that automatically use modern video analysis techniques to identify and track anomalies in the video scene. In emergencies, automated systems can alarm the controlling authorities to take appropriate actions against the detected anomaly; this is impractical in a manual monitoring system and prone to errors due to human limitations. So, developing an anomaly detection system is a primary step toward building futuristic automated video surveillance systems.

Real-world anomalous events have complications and diversity. It is nearly impossible to list all possible anomalous events. Some events are complicated due to inter and intra-class variations, where two or more different classes possess similar characteristics. For example, robbery, shoplifting, and stealing have comparable appearance features; however, they belong to different anomalous events for crime scenes [54]. Another major challenge in a real-world anomalous event is the speed of action. Some actions are high-speed actions, as some events happen within seconds or in a fraction of a second, such as explosions and road accidents. On the other hand, some actions are of medium duration and take minutes to happen, such as abuse, arrest, and fight activity. The third category of events is slow actions that span over a long duration of time, such as robbery, shoplifting, and vandalism. These activities are sometimes so complex that even a human may not identify what a person is doing in a given video segment until the video is seen from start to end to understand the nature of the activity. The reason is the slowness of these real-world anomalous activities, such as abuse and robbery, compared to the explosion road accidents events.

Initially, researchers used various traditional feature descriptors along with traditional machine learning techniques for anomaly detection [6, 30, 50]. However, in recent years, deep learning architectures have been used to perform several computer-vision tasks successfully. With the advent of deep learning architecture 2D ConvNets are proposed to learn spatial features for image classification [25], object detection [13], activity recognition and abnormal behaviors identifications [35, 36, 45–47, 51]. The 3D-ConvNets [57] were proposed for activity recognition in videos as an end-to-end learning model for feature engineering and classification. ImageNet challenge has given birth to various deep architecture for image classifications for 1000 categories trained on millions of annotated images [25], such as Inception [20], VGG-16 [52], and ResNet [16]. The problem with ImageNet-based architectures is that they are specifically designed for image classification tasks only. These cannot be used to train on videos dataset to learn Spatio-temporal features. The 3D-ConvNets [57] is a popular video classification architecture to learn spatio-temporal features from video sequences; however, the network lacks the pretraining capabilities of the ImageNet dataset and has limited feature learning capabilities compared to deep complex ImageNet deep

architectures. Recently, Two-Stream Inflated 3D-ConvNets (I3D) [4] was proposed as a new benchmark network architecture using pre-trained ImageNet and Kinetics Human Action Video Dataset [22], with Inception-V1 [20] as a base architecture. This architecture achieved remarkable results in the domain of human action recognition on UCF-101 [53] and HMDB-51 [26] datasets due to Inflated Inception modules. However, the I3D network has yet to be tested on complex real-world anomaly detection problems. Real-world anomalies have complex human behaviors as compared to normal human simple actions. The proposed method focuses on real-world anomaly classification and offers the following two major contributions.

1. A novel technique of dynamic frame-skipping for generating meaningful temporal sequences. This technique helps us learn fast and slow activities and enables our model to generalize better over various activities such as robbery, shoplifting, stealing, explosion and road accidents.
2. A new deep learning-based framework inspired by the I3D model is proposed and is investigated on one of the most challenging benchmark UCF-Crime datasets for anomaly detection and classification.

The results show that the proposed method outperforms state-of-the-art methods for real-world anomaly detection on the UCF-Crime dataset. The rest of the paper is organized as Related work, Proposed Methodology, Datasets, Experimental Setup, Results, and Conclusion are presented in Sections 2, 3, 4, 5, 6, and 7, respectively.

2 Related work

Anomaly detection is one of the long outstanding and most challenging problems for researchers in computer vision. A lot of research has been done to investigate and develop anomaly detection algorithms such as [3, 15, 24, 27, 30, 34, 63, 67, 71], and survey papers such as [7, 38, 40]. There are three kinds of approaches generally adopted for video anomaly detection, i.e., unsupervised, weakly-supervised, and supervised settings based on the training data [72].

2.1 Unsupervised methods

In real-world scenarios, since anomalous events occur with a low probability as compared to normal events, it is often hard to capture/record all kinds of anomalies. In contrast, normal video data is broadly accessible from public surveillance cameras or social media. Unsupervised methods focus on learning by using normal videos in the training dataset. Although these approaches have more generalization abilities to detect anomalies in unseen anomalous patterns, however achieving high detection performance in real-world scenarios is the key challenge for unsupervised methods [72].

Early unsupervised algorithms used conventional handcrafted feature detectors and probability methods. These methods are based on expert-designed hand-engineered feature descriptors, for instance, optical flow features using the PCA (Principal Component Analysis) in association with space-time MRF (Markov Random Field) [23]. Moreover, the BoW (Bag of Words) and LDA (Latent Dirichlet Allocation) were used to estimate the interaction forces [34], and DT (Dynamic Textures) based model was used for temporal normalcy [27]. Another

approach used corner features with interaction flow to train a random forest classifier [58]. Sparse coding and a combination of sparse learning frameworks were used to attain a speed of 150 frames per second (fps) with handcrafted feature descriptors to classify normal and anomaly in video frames [30]. Appearance and motion-based features are investigated with a dictionary learned to reconstruct normal events for anomaly detection [10].

Recent advancements in deep learning techniques can benefit from large-scale datasets on high-end computation resources. Following unsupervised settings for anomaly detection, several research works are reported based on deep auto-encoder (AE) architecture [9, 15, 29] and in survey papers [8, 66]. A fully convolutional network (FCN) based method has been examined to learn motion features and regular discriminative patterns with FCN-based AE [15]. The FCN associated with long short-term memory (LSTM) as a Conv LSTM-AE was proposed to further enhance the AE performance [31]. Recurrent Neural Network (RNN) and sparse coding-based temporally-coherent sparse coding framework was suggested to introduce temporal information in the background of sparse coding for videos [32]. A memory-augmented (AE) approach has been investigated to memorize prototypical normal patterns with an attention-based memory access mechanism to reconstruct future frames for anomaly detection [14]. Stacked denoising autoencoders (SDAE) were used to learn both appearance and motion features [63, 64]. In contrast, dynamic video anomaly detection and localization were performed using sparse denoising autoencoder [37]; stacked denoising was originally proposed in [60, 61]. Spatiotemporal based AE was proposed for abnormal event detection [9, 11, 68]. Anomaly detection is generally achieved based on the reconstruction error for all discussed AE-based methods. In contrast, another paper formulated this problem as a multi-class classification problem using AE features by applying k-means clustering and one-versus-all Support Vector Machine (SVM) for abnormality detection [21]. A deep 3D autoencoder cascaded with a 3D CNN model is examined to detect anomalies [42]. Furthermore, Instead of using AE to compute the reconstruction error on the future frame to predict anomaly, an approach of future frame prediction based on the past frames is investigated to compute anomaly score based on the difference between the predicted future frame and ground truth [5, 12, 28, 41, 55, 65]. Furthermore, FCN as a generator framework predicts future frames in generative models [28, 65].

2.2 Weakly-supervised methods

Machine learning algorithms require a large amount of training data. Social Media has emerged as a popular source for large-scale data collection, such as videos on YouTube. Due to the increasing data on social media, it is possible to collect and annotate many datasets for anomaly detection. For well-defined anomaly activities, supervision information can significantly improve performance at the cost of laborious work. Following weakly-supervised settings, the video-level annotation has been adopted for training the model. The training set is annotated for normal and anomalous videos. However, the temporal location of the anomalous event in the videos is unknown, i.e., weak supervision [54]. In weakly-supervised settings, anomaly detection is investigated as multiple instance learning (MIL) problem. A graph-based MIL framework with anchor dictionary learning was proposed [17]. A C3D-based MIL method was evaluated on a benchmark UCF-Crime dataset [54]. Whereas another approach has studied weakly supervised learning as a noisy label learning problem. A graph convolutional network (GCN) based framework is designed to refine the noisy labels of a weakly labelled dataset for anomaly detection [69].

2.3 Supervised methods

There are certain scenarios in which supervised settings can significantly improve the classifier performance. These scenarios, where background objects are well defined in video scenes, i.e., the roads and cars for highway traffic accident detections. A frame-level temporal annotated training video is used for anomalous pattern identifications. The most common approach is to use prior geometric knowledge with object detection or semantic segmentation with further supervision from other publicly available datasets [2, 62]. In another approach, Faster-RCNN is first used to detect vehicles; then, the accident score is learned using the attention-based LSTM module [49].

Convolutional Neural Network (CNN) [18, 43, 48] based approaches are proposed for abnormal event detections in videos. A CNN model in compliance with multi-task Fast-RCNN is trained on large supervised datasets. The learned model corresponds to the generic model for video anomaly detection [18]. In addition to CNN, C3D features are generally more suitable for spatio-temporal feature learning in videos than standard CNNs [57]. Furthermore, a convolutional spatio-temporal auto-encoder-based method is also examined for anomaly detection in videos [11], and a deep 3D auto-encoder cascaded with a 3D CNN model is investigated to detect anomalies in the supervised domain [42]. The literature review shows that the C3D model-based Nearest Neighbor classifier has set initial benchmark results on the UCF-Crime dataset for anomaly classification on 14 different kinds of real-world anomalies [54]. Furthermore, TCNN [54], TCNN with motion [70], and C3D Fine-tuning [33] based approaches are investigated for anomaly detection. However, there is huge room to discover and improve the accuracy of the UCF-Crime dataset in this domain.

Furthermore, the UCF-Crime dataset is primarily available for weakly-supervised settings with video-level annotations only. Recently, the frame-level annotations of the UCF-Crime dataset for all 14 classes have been provided as a fully-supervised learning problem for anomaly classification [33]. In the literature, the majority of the researchers have formulated anomaly detection as an unsupervised or weakly-supervised learning problem on the UCF-Crime dataset [56, 59]. Whereas very few researchers have implemented anomaly detection as a fully-supervised learning problem on the dataset [33, 54, 70]. Fully-supervised implementation requires the laborious effort of frame-level annotations for each activity. Further, it is more challenging to organize technical implementation and model training due to the large-scale video data size for experimental purpose on the full dataset for all 14 classes. Some researchers have reported results on a subset of the UCF-Crime dataset on just 5 out of 14 classes to train in resource constraint environments producing relatively higher accuracy due to fewer anomalies selected for model training [59]. Therefore, the model training on large-scale datasets for anomaly detection in a fully-supervised mode is challenging and always compute demanding. However, a successfully trained model leads to a better generalization capability of the model to classify a wide variety of anomalies for developing real-world anomaly detection systems.

Recently, a 3D ConvNet model named Two-Stream Inflated 3D-ConvNets (I3D) was proposed [4]. This architecture has a C3D-like design with Inception-V1 [20] as base architecture pre-trained on ImageNet and Kinetics Human Action Video Dataset [22]. The I3D has inflated Inception-V1 modules from 2D to 3D, and the network has achieved excellent results for human action recognition on UCF-101 [53] and HMDB-51 [26] datasets. The proposed framework is inspired by the 3D learning capabilities of I3D and is investigated on

UCF-Crime as a fully-supervised learning problem for anomaly detection in surveillance videos.

3 Proposed methodology

This section defines the proposed methodology. First, the dynamic frame-skipping technique for generating meaningful temporal sequences is presented, then a proposed 3D deep learning-based model for anomaly detection is discussed.

3.1 Dynamic frame-skipping

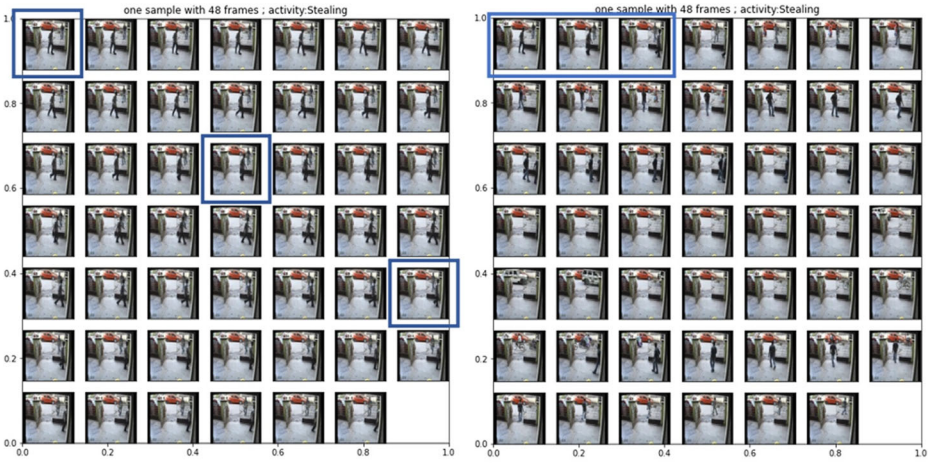
Real-world anomaly detection and classification are challenging due to the diversity in actions and human behaviors, which is hard to capture on surveillance cameras. A generalized model should detect a range of anomalous events regardless of the event's speed. Usually, models suffer from low accuracy due to insufficient discriminative feature learning for fast and slow events simultaneously from real-world surveillance videos. A normal surveillance camera records at 30 fps, which means we have a stack of 30 frames recorded each second, containing live-action. This frame rate is quite reasonable for fast actions such as explosions and road accidents. However, this frame rate is fairly high if we aim to detect slow activities such as stealing and robbery. These activities comprise complex human behaviors that can span over minutes to define an action. These slow activities are often so long that a full temporal sequence completely defining the activity would not be possible to feed to the limited temporal input size of the deep learning architectures for the classification tasks.

In order to learn fast and slow activities simultaneously, the dynamic frame-skipping method based on activity length or video length is proposed to skip video frames dynamically per activity. This method preserves the fast activities with minimal frame-skipping at approximately zero, further making slow activity virtually fast for the model to learn about slow activities efficiently. The dynamic frame-skipping (DFS) is computed as:

$$F = V/N, \tag{1}$$

Where F is the number of frames to skip based on the video length or activity length V by N. It is observed that fast anomalous events in the UCF-Crime dataset have a minimum average span of 200–400 frames covering the event. Hence, N is proposed at 200. F frame-skipping is applied to videos during data normalization; see more details in Section 5.1. The proposed method appropriately propagates the fast and slow actions' temporal features to let the model learn about real-world anomalous events, enhancing the better generalization capabilities of the model.

Figure 1 shows the dynamic frame-skipping in action on the stealing activity. Even the higher temporal length of 48 frames cannot define a person's behavior without frame-skipping. Figure 1a, without frame skipping, shows a person walking, but it does not provide any clue about the person's behavior regarding stealing. Whereas, in Fig. 1b, a more descriptive activity of the human behavior is captured in a single temporal sequence due to dynamic frame-skipping. Two persons suspiciously enter and exit the house depicting a suspected criminal activity.



(a) (b)

Fig. 1 Impact of dynamic frame-skipping on temporal sequence generation for slow activities: **a** Represents stealing activity temporal sequence sample with no frame-skipping. Anomalous activity is not visible; **b** Represents stealing activity temporal sequence sample with 16 dynamic frame-skipping. Anomalous activity is visible. A complete temporal sequence sample of (**a**) is represented in the first three frames in (**b**) due to dynamic frame-skipping

3.2 Proposed approach

ImageNet has proposed state-of-the-art deep architecture developed over the years for the task of image classifications for 1000 categories trained on 15 million annotated images [25], such as Inception [20], VGG-16 [52], and ResNet [16]. However, the video classification task is still an open challenge for deep architectures to learn very deep representations of real-world videos. Recently, state-of-the-art architectures performing well on the ImageNet challenge have been investigated in the video classification domain. As a result, a new Two-Stream Inflated 3D ConvNet (I3D) was proposed. I3D is based on 2D ConvNet inflation by starting with the 2D architecture of $N \times N$ and inflating filters and pooling kernels into 3D by endowing them with another temporal dimension of $N \times N \times N$. I3D is designed to work on videos by leveraging the well-designed architecture and parameters of the network on ImageNet [4]. The suggested model is based on I3D network with the proposed design of Conv-Top-Layers for better learning and generalization of anomaly detection, leveraging dynamic frame-skipping, using pre-trained weights on ImageNet and Kinetics [22] to train the model on the UCF-Crime dataset [54], after data normalization in various settings. The proposed model is further tested under fine-tuning and full network training settings for anomaly recognition.

I3D is trained and proposed with I3D designed Conv-Top-Layers for the UCF-Crime dataset. Conv-Top-Layers consist of *average pool, dropout (0.5), conv 1 × 1 × 1, batch normalization, reshape,* and *output* layer. Furthermore, the model is also investigated on Flatten-Top-Layers. Flatten-Top-Layers consist of *flatten, batch normalization, dense (1500), dropout (0.5) / l2 normalization, dense (128), dropout (0.5) / l2 normalization,* and *output* layer at the end. Figure 2 shows the diagram of the I3D network with the Inflated Inception-V1 module with proposed Conv-Top-Layers and examined Flatten-Top-Layers.

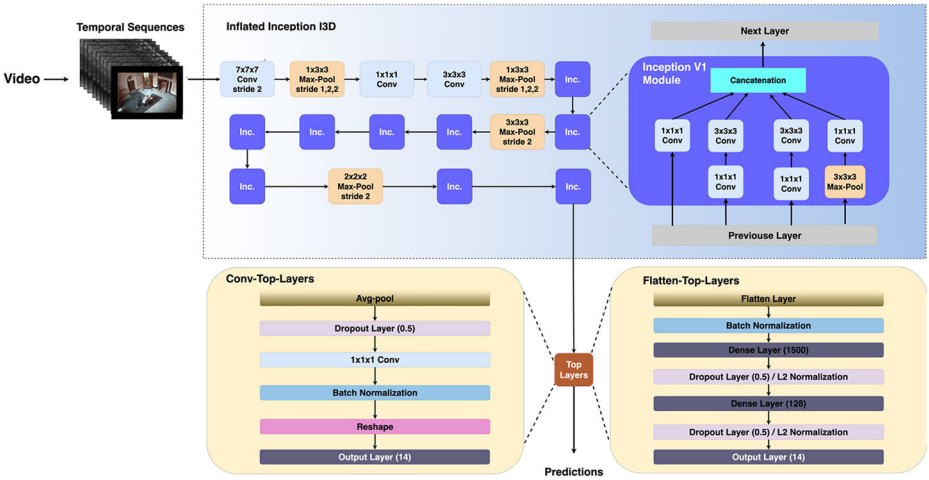


Fig. 2 Proposed model of Inflated Inception I3D network with Conv-Top-Layers. I3D is a C3D-like 3D ConvNet model with an Inflated 3D design of Inception-V1 as the base architecture. The inflated 3D version of the Inception-V1 module is represented on the top right, which is used to build the Inflated Inception I3D network. Conv-Top-Layers are proposed on the bottom left. Additionally, Flatten-Top-Layers investigated during experiments are presented on the bottom right

Table 1 shows the architectural details of the proposed network. The network is comprised of Conv3d_7 × 7 × 7, MaxPool_1 × 3 × 3, Conv3d_1 × 1 × 1, Conv3d_3 × 3 × 3, MaxPool_1 × 3 × 3, Inception_block_1, Inception_block_2, MaxPool_3 × 3 × 3, Inception_block_3, Inception_block_4, Inception_block_5, Inception_block_6, Inception_block_7, MaxPool_2 × 2 × 2, Inception_block_8, Inception_block_9 and Conv-Top-Layers for 14 classes output. The network has 12,286,984 total parameters. The loss function of categorical cross-entropy has been used for multi-class classification task. The categorical cross-entropy is a measure of the difference between two probabilities distributions as described:

$$L(p,q) = -\sum_{x=0}^{n} p(x)\ log\ q(x), \tag{2}$$

Where $p(x)$ and $q(x)$ are two probabilities and $L(p, q)$ is loss computed, measuring the difference between two probabilities $p(x)$ and $q(x)$. The network is trained and tested for both Conv-Top-Layers and Flatten-Top-Layers. Results confirm that the proposed model design of the I3D network with Conv-Top-Layers is outperforming all other benchmark accuracies.

4 Datasets

The UCSD Pedestrian is a classic dataset to discover anomalous patterns comprised of Ped1 and Ped2 datasets, containing recorded scenes of pedestrian walkways captured from two different static cameras. The UCSD contains a few anomalies like skaters, bikers, carts, walk-across, and others [27]. The Subway dataset has two long videos recording of a subway entrance and exit, capturing the activities of the people while entering and leaving through the

Table 1 Architecture details about the proposed model of Inflated Inception I3D network with Conv-Top-Layers

Block	Layers	Input	Kernel	Output
Input		48×199×199×3	N/A	48×199×199×3
Conv3d_7×7×7	Conv3D	48×199×199×3	7×7×7	24×100×100×64
	Batch Normalization	24×100×100×64	–	24×100×100×64
	Activation	24×100×100×64	–	24×100×100×64
MaxPool_1×3×3	MaxPooling3D	24×100×100×64	1×3×3	24×50×50×64
Conv3d_1×1×1	Conv3D	24×50×50×64	1×1×1	24×50×50×64
	Batch Normalization	24×50×50×64	–	24×50×50×64
	Activation	24×50×50×64	–	24×50×50×64
Conv3d_3×3×3	Conv3D	24×50×50×64	3×3×3	24×50×50×192
	Batch Normalization	24×50×50×192	–	24×50×50×192
	Activation	24×50×50×192	–	24×50×50×192
MaxPool_1×3×3	MaxPooling3D	24×50×50×192	1×3×3	24×25×25×192
Inception_block_1	Conv3D_1 / Batch Normalization / Activation	24×25×25×192	1×1×1	24×25×25×96
	Conv3D_1_1 / Batch Normalization / Activation	24×25×25×96	3×3×3	24×25×25×128
	Conv3D_2 / Batch Normalization / Activation	24×25×25×192	1×1×1	24×25×25×16
	Conv3D_2_1 / Batch Normalization / Activation	24×25×25×16	3×3×3	24×25×25×32
	MaxPooling3D_3	24×25×25×192	3×3×3	24×25×25×192
	Conv3D_3_1 / Batch Normalization / Activation	24×25×25×192	1×1×1	24×25×25×32
	Conv3D_4 / Batch Normalization / Activation	24×25×25×192	1×1×1	24×25×25×64
	Concatenate [Conv3D_1_1, Conv3D_2_1, Conv3D_3_1, Conv3D_4]	[24×25×25×128, 24×25×25×32, 24×25×25×32, 24×25×25×64]	–	24×25×25×256
Inception_block_2	Inception_block N Layers	24×25×25×256	N×N×N	24×25×25×480
MaxPool_3×3×3	MaxPooling3D	24×25×25×480	3×3×3	12×13×13×480
Inception_block_3	Inception_block N Layers	12×13×13×480	N×N×N	12×13×13×512
Inception_block_4	Inception_block N Layers	12×13×13×512	N×N×N	12×13×13×512
Inception_block_5	Inception_block N Layers	12×13×13×512	N×N×N	12×13×13×512
Inception_block_6	Inception_block N Layers	12×13×13×512	N×N×N	12×13×13×528
Inception_block_7	Inception_block N Layers	12×13×13×528	N×N×N	12×13×13×832
MaxPool_2×2×2	MaxPooling3D	12×13×13×832	2×2×2	6×7×7×832
Inception_block_8	Inception_block N Layers	6×7×7×832	N×N×N	6×7×7×832
Inception_block_9	Inception_block N Layers	6×7×7×832	N×N×N	6×7×7×1024
Conv-Top-Layers	AveragePooling3D	6×7×7×1024	2×7×7	5×1×1×1024
	Dropout (0.5)	5×1×1×1024	–	5×1×1×1024

Table 1 (continued)

Block	Layers	Input	Kernel	Output
	Conv3D	5 × 1 × 1 × 1024	1 × 1 × 1	5 × 1 × 1 × 14
	Batch Normalization	5 × 1 × 1 × 14	–	5 × 1 × 1 × 14
	Reshape	5 × 1 × 1 × 14	–	5 × 14
	Activation	5 × 14	–	5 × 14

turnstile. The dataset has anomalies, such as people jumping, squeezing around the turnstile, walking/going in the wrong direction, and a person cleaning the walls [1]. Another dataset, CUHK Avenue, has short video clips of the pedestrian walkway recorded from a single outdoor surveillance camera. The videos captured normal human behaviors, such as people entering or coming out of the building. The anomalies in the scenes are mostly staged actors doing activities, such as a person throwing a backpack or papers into the air, a child skipping across the walkway, etc. [30].

For video anomaly detection, although there are certain datasets widely used by the research community for anomalous activity recognition, however, there are some subjective deficiencies in these datasets to determine anomalies in real-world scenarios. The shortcoming is primarily related to the simplicity of the captured scenes, an insignificant number of anomalous events, the low resolution of videos in some datasets, staged actors/anomalous events, lack of spatial ground-truth, and inconsistencies in data annotations in some cases [39]. The UCF-Crime is a large dataset collected from online internet video sources. The diverse collection represents video recordings from hundreds of distinct surveillance cameras under different environmental circumstances. The dataset contains 1900 long untrimmed videos that contain 13 real-world anomalies, such as assault, accident, burglary, abuse, arson, arrest, fighting, explosion, shooting, shoplifting, robbery, stealing, and vandalism [54], see Fig. 3 Out of 1900 videos, 950 videos contain normal events, and the rest of the 810 videos contain 13 anomalous classes from real-world scenes. The frame-level annotations for the dataset were recently published for fully-supervised learning [33]. The dataset also contains fast and slow activities such as explosions, road accidents, robbery, stealing, and vandalism. The wide variety of the scenes captured from distinct cameras at different locations under different circumstances makes the dataset the most challenging one for anomaly classification to develop real-world surveillance systems.

5 Experimental setup

This section describes the experimental details for the proposed methodology, as discussed in Section 3.

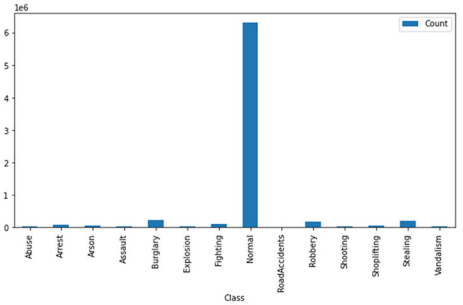
5.1 Data normalization

UCF-Crime is a challenging dataset due to diverse anomalies collected from real-world scenes. It is equally challenging to manage such a huge dataset to perform successful experiments on the dataset. The dataset has approximately 10 million total video frames for model training and

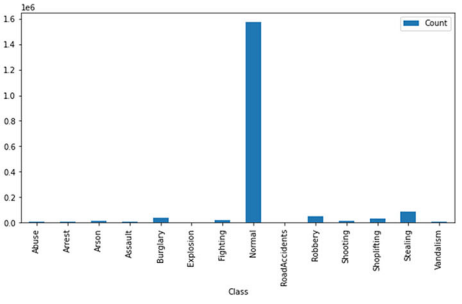
Fig. 3 Sample anomalous classes frames from UCF-Crime Dataset: **a** Abuse; **b** Arson; **c** Assault; **d** Explosion; **e** Fighting; **f** Road Accidents

testing, comprised of 14 classes with 13 anomalous classes and 1 class for normal events. The frame-level dataset annotation is available for anomaly classification [33].

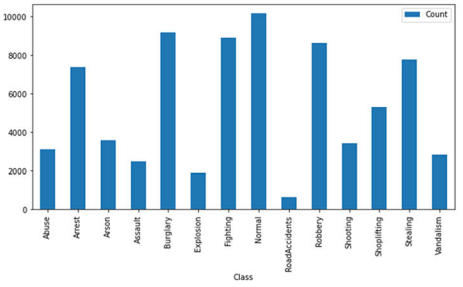
However, when by looking at the annotations, dataset videos are converted into frames for each category, it is evident that normal videos have clear domination with 6,298,945, 1,570,919 frames out of 7,385,800, 1,859,718 frames, over other anomalous classes in Trainset and Testset as presented in Fig. 4a and b respectively. It is impossible to train a model on such a biased, unbalanced dataset to learn anomalous events without data normalization. Next, a normalized Trainset and Testset with zero frame-skipping is produced by random selection of several normal samples from all normal videos, as close to the number of other class samples, as shown in Fig. 4c and d, respectively. The dataset is normalized; however, a few classes, for instance, explosion and road accidents, are not balanced due to short-length videos compared to other classes. Figure 4e and f represents the extraction of another normalized Trainset and Testset with fixed-length one frame-skipping in all videos. The visualization shows that fixed-length frame-skipping does not participate in data normalization; instead, it decreases the sample space for all classes. The dataset is still not sufficiently balanced as short-length video samples are even further reduced by half due to one frame-skipping. Next, Fig. 4g and h represents the visualization of normalized Trainset and Testset with the proposed dynamic frame-skipping strategy as discussed in Section 3.1. Dynamic frame-skipping positively contributes to data normalization across all classes, making the dataset more balanced and ordered for classifier training. A more balanced sample space for long videos is generated based on video length using dynamic frame-skipping. Furthermore, short-length video samples are accurately preserved as compared to all other classes for better-generalized classifier learning. The sample distribution is based on the temporal length of 48 and temporal stride of 12 per sample for the proposed model design, as discussed in Section 3.2.


1e6
Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

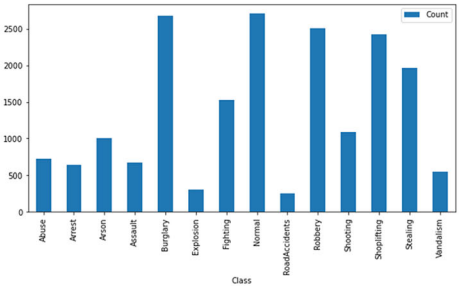
(a)


1e6
Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

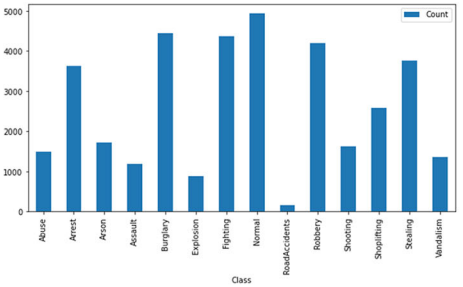
(b)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

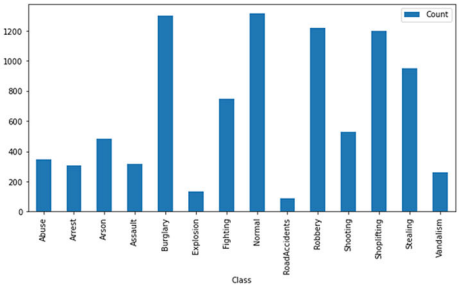
(c)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

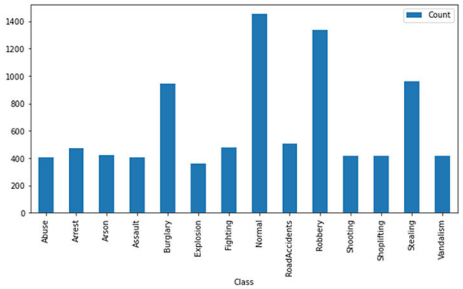
(d)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

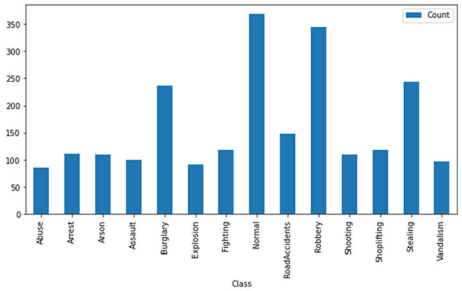
(e)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

(f)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

(g)


Count
Abuse
Arrest
Arson
Assault
Burglary
Explosion
Fighting
Normal
RoadAccidents
Robbery
Shooting
Shoplifting
Stealing
Vandalism
Class

(h)

◀ **Fig. 4** Data normalization steps applied to UCF-Crime Dataset for model training: **a** Trainset without normalization; **b** Testset without normalization; **c** Trainset normalized without frame-skipping; **d** Testset normalized without frame-skipping; **e** Trainset normalized with fixed-length one frame-skipping; **f** Testset normalized with fixed-length one frame-skipping; **g** Trainset normalized with dynamic frame-skipping; **h** Testset normalized with dynamic frame-skipping

5.2 Parameters settings

Hyperparameter tuning is performed to achieve superior accuracy results, and optimal parameters are discovered during a wide range of experiments. Several extensive experiments are performed by examining dataset normalization, frame-skipping, temporal length, temporal stride, and random runtime augmentation on image sequences with architectural changes in the model. Based on the experiments, batch size 8 is set for training with temporal length 48, temporal stride 12, and dynamic frame-skipping applied. Input image dimension is set to 199 × 199 for the model, and random runtime augmentation with a probability of 0.2 is applied to image sequences before feeding the samples to the model. Data shuffle is set to true at all levels for sample generation for train, validation, and test sets. Progressive data loading is implemented carefully, and the dataset is loaded to GPU in an organized fashion for training and testing for model evaluations in a computationally efficient way. Each batch is provided to the model containing temporal image sequences for model training. The network is trained on *Adam* with a *learning rate of* 0.0001 on NVIDIA 1080ti GPU, Intel Xeon X5670 @ 2.93 GHz (2 processors) with 96GB RAM.

The results of the experiments are reported as per the standard anomaly detection evaluation metrics to compare with other benchmark published methods. The results are reported using the evaluation metrics of accuracy, confusion matrix, and area under the curve (AUC).

6 Results

This section describes the results of the proposed methodology. The proposed Methodology, Dataset, and Experimental Setup are discussed in Sections 3, 4, and 5, respectively.

The proposed methodology is evaluated on the benchmark dataset of UCF-Crime [54] for anomaly detection in real-world environments. A comparative analysis is performed with the reported benchmark accuracies from the literature review. Previously, researchers have mostly used C3D network variations for learning spatio-temporal features for anomaly detection. This research work has investigated and proposed a model design based on I3D [4] Inflated Inception model with Conv-Top-Layers with batch normalization, specifically to learn anomalous events leveraging the dynamic frame-skipping for fast and slow activities learning to achieve the highest model accuracy. I3D [4] Inflated Inception model is extensively investigated for both Conv-Top-Layers and Flatten-Top-Layers for fine-tunning and full network training with dynamic frame-skipping under different training settings.

The proposed strategy of dynamic frame-skipping not only performs data normalization for the Training set and Test set, as discussed in Section 5.1. Furthermore, it creates a fast virtual representation of slow activities for a model to learn complex human behaviors for long-duration slow anomalous activities such as robbery, shoplifting and stealing. Figure 5 illustrates the confusion matrix, and Table 2 shows the statistical metrics Precision, Recall, and F1-measure for 13 anomalous classes and 1 normal event class on the UCF-Crime dataset.

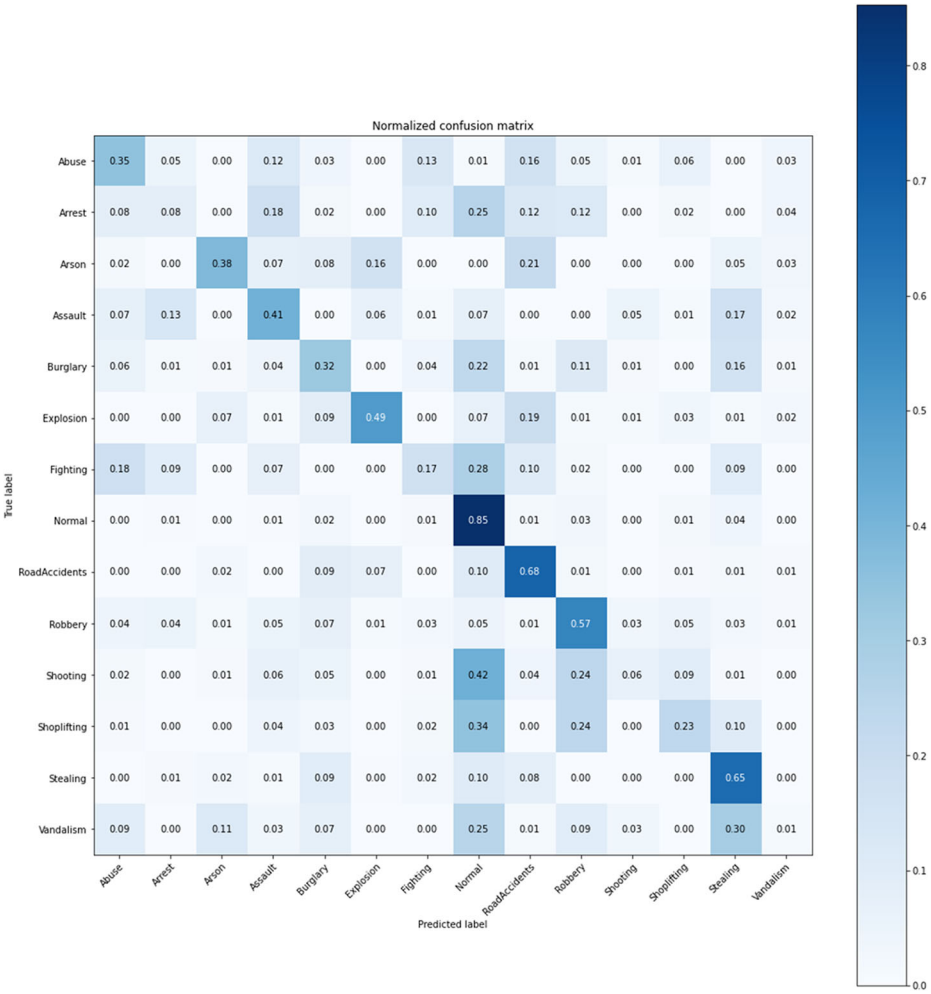


Fig. 5 Confusion matrix generated for proposed method I3D with Conv-Top-Layers on UCF-Crime dataset

The result shows that the model is not missing even a single class producing zero precision. The model is able to learn generalized representation to capture all anomalous classes with precision. Furthermore, the F1-measure score suggested that the model is successfully able to learn fast and slow activities with the highest F1-measure score, such as explosion, road accidents from fast events have the highest F1-measure score of 0.51, 0.55 respectively and robbery, stealing from the slow and long activities have also achieved highest F1-measure score of 0.60, 0.59 respectively. The F1-measure score of robbery stealing is even higher than explosion and road accident events. This suggests that dynamic frame-skipping directs the model to learn the realistic temporal features for fast and slow activities simultaneously to achieve better generalization capabilities.

Sultani et al. [54] proposed a UCF-Crime dataset for real-world anomalies detection in surveillance videos and published their benchmark results on C3D [57] with nearest neighbor classifier and Tube Convolutional Neural Network (TCNN) [19] with accuracies of 23%, 28.4% respectively for anomaly detection. Zhu et al. [70] achieved an accuracy of 31% using

Table 2 Evaluation metrics generated for proposed method I3D with Conv-Top-Layers on UCF-Crime dataset

Classes	Precision	Recall	F1-measure
Abuse	0.28	0.35	0.31
Arrest	0.15	0.08	0.11
Arson	0.55	0.38	0.45
Assault	0.31	0.41	0.35
Burglary	0.42	0.32	0.37
Explosion	0.54	0.49	0.51
Fighting	0.27	0.17	0.21
Normal	0.51	0.85	0.64
Road Accidents	0.46	0.68	0.55
Robbery	0.62	0.57	0.60
Shooting	0.23	0.06	0.10
Shoplifting	0.38	0.23	0.28
Stealing	0.54	0.65	0.59
Vandalism	0.05	0.01	0.02
Micro Average	0.47	0.47	0.47
Macro Average	0.38	0.38	0.36
Weighted Average	0.43	0.47	0.44

TCNN with motion-aware features. Maqsood et al. [33] published their accuracy of 45% on C3D fine-tunning. Some researchers have reported results as a weakly-supervised learning approach on a partial UCF-Crime dataset of just 5 classes, which makes the dataset much simpler, producing relatively higher accuracy due to simpler anomalies selected for model training [59], whereas we have reported results on the full dataset as a whole for all 14 classes with full dataset complexity, to produce a real-world anomaly detection model that can practically be used in real-world scenarios. Furthermore, there is a difference in the evaluation matrix for weakly-supervised reported results [56, 59] and fully-supervised results with frame-level annotation in comparison. Table 3 shows results with all published state-of-the-art papers reporting results as fully-supervised learning problems on challenging UCF-Crime datasets. Results show that I3D with Flatten-Top-Layers using one frame-skipping on full network training has achieved 30% accuracy. I3D with Flatten-Top-Layers using dynamic frame-skipping and fine-tuning has achieved 39% accuracy. I3D with Flatten-Top-Layers using dynamic frame-skipping with end-to-end network learning has achieved 40% accuracy as compared to one static frame-skipping with 30% accuracy. There is a sharp 10% accuracy gain

Table 3 Accuracy comparison of the proposed method with different approaches on the UCF-Crime dataset

Author/Year	Method	Frame-Skipping	Accuracy (%)
Sultani et al. (2018) [54]	C3D + Nearest Neighbor	Zero	23%
Sultani et al. (2018) [54]	TCNN	Zero	28.4%
Zhu et al. (2019) [70]	TCNN + Motion	Zero	31%
Maqsood et al. (2021) [33]	C3D Fine-tunning	Zero	45%
Network with Flatten-Top-Layers	End-to-end network learning	One	30%
Network with Flatten-Top-Layers	Fine-tunning	Dynamic	39%
Network with Flatten-Top-Layers	End-to-end network learning	Dynamic	40%
Network with Conv-Top-Layer	Fine-tunning	Dynamic	45%
Network with Conv-Top-Layers (Proposed Method)	End-to-end network learning	Dynamic	**47%**

using dynamic frame-skipping by generating a much-formulated meaningful temporal sequence for fast and slow activities learning for the model learning.

Furthermore, I3D with Conv-Top-Layers is examined. I3D with Conv-Top-Layers using dynamic frame-skipping and fine-tuning has achieved 45% accuracy. Finally, the proposed I3D with Conv-Top-Layers using the proposed dynamic frame-skipping with suggested experimental settings in an end-to-end networking learning scheme has achieved the highest accuracy of 47%, surpassing all previously reported accuracies, to our knowledge. The results suggested that dynamic frame-skipping has generated a unique set of temporal features for fast and slow activities image sequences for robust learning of the internal convolutional layer of the model in an end-to-end learning scheme.

Figure 6 shows the final overall AUC curve plot with an AUC score of 0.837 for the proposed method of I3D with Conv-Top-Layers using dynamic frame-skipping in an end-to-end learning scheme on the UCF-Crime dataset.

The results confirm the proposed approach of I3D with Conv-Top-Layers leveraging dynamic frame-skipping outperforms state-of-the-art accuracies on the challenging UCF-Crime dataset for the task of real-world anomaly detection in video surveillance.

7 Conclusion

The major challenge to build a real-world surveillance system is to capture a wide range of anomalies. Real-world anomalous events are diverse and more complex than a simple human action recognition task. Some events are fast and happen rapidly within seconds, i.e., explosions and road accidents. Some activities are slow and happen over a longer duration, i.e., robbery, shoplifting, stealing, and vandalism comprise complex human behaviors to be

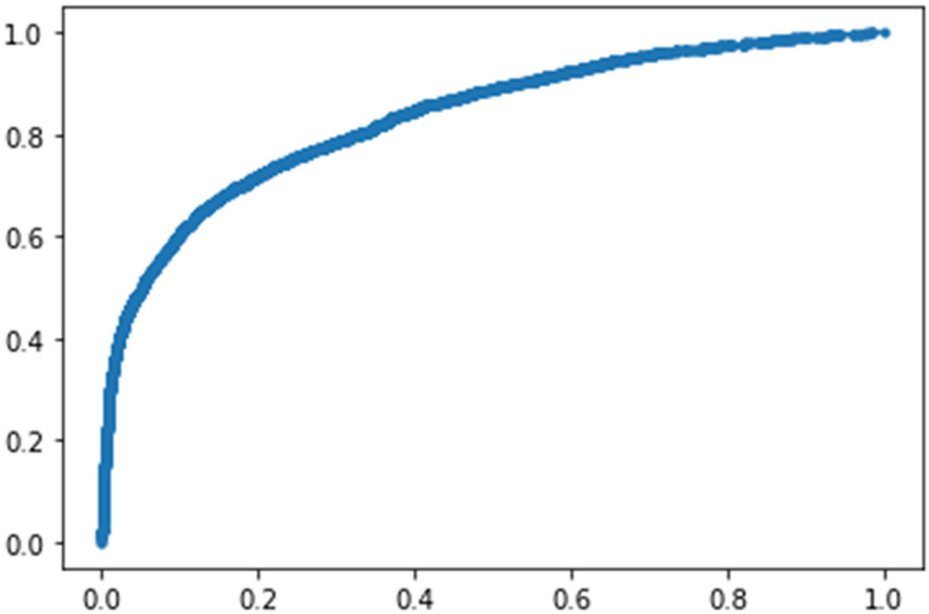


Fig. 6 Proposed method AUC plot with AUC score of 0.837 on the UCF-Crime Dataset

detected in real-world scenarios. This research has investigated the fast and slow activities' temporal features for the videos captured at 30 frames per second (fps), which is a good frame rate to capture fast activities per frame. However, the 30 fps frame rate is too high for slow activities. The slow activities will essentially be recorded in several thousand frames for a single anomalous activity such as robbery, or stealing. In such activities, a person initially looks suspicious and then performs anomalous activity over time. In order to learn fast and slow activities simultaneously from real-world scenarios, a novel strategy of dynamic frame-skipping is investigated and proposed for anomalous event recognition. Such that, deep learning architectures can better learn long-term temporal features for slow and fast activities simultaneously for better model generalization.

Furthermore, a new deep learning model inspired by the I3D Inflated Inception network with Conv-Top-Layers is proposed to learn spatial and temporal features in video frames. The experiments are performed on the UCF-Crime real-world complex benchmark dataset containing 13 real-world anomalous classes and normal events. The results demonstrate that the proposed model has learnt and outperformed in terms of the F1-measure score for both fast and slow anomalous activities. The developed model is robust and has gained the highest accuracy with better generalization capabilities for real-world anomaly detection in surveillance videos.

Author contribution A.M., A.B.S. and Z.H. conceived and designed the research direction; A.M. proposed/implemented methodology and performed the research experiments; A.B.S. and Z.H. analyzed the data; A.B.S. and A.M. contributed reagents/materials/analysis tools; A.M. wrote the research paper. All authors have read and agreed to the published version of the manuscript.

Funding This research is supported by the PDE-GIR project, which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778035.

Data availability Not applicable.

Declarations

Institutional review board statement Not applicable.

Informed consent Not applicable.

Conflict of interest The authors declare no conflict of interest.

References

1. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30(3):555–560
2. Bai S et al (2019) Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In: Proc. CVPR Workshops, pp 117–124
3. Basharat A, Gritai A, Shah M (2008) Learning object motion patterns for anomaly detection and improved object detection. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp 1–8
4. 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

5. Chalapathy R, Toth E, Chawla S (2019) Group anomaly detection using deep generative models. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol 11051 LNAI, pp 173–189
6. Cheng KW, Chen YT, Fang WH (2015) Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. IEEE Trans Image Process 24(12):5288–5301
7. Chidananda K, Kumar S (2022) Human anomaly detection in surveillance videos: a review. Inf Commun Technol Compet Strateg:791–802
8. Chong YS, Tay YH (2015) Modeling representation of videos for anomaly detection using deep learning: a review. arXiv Prepr. arXiv1505.00523
9. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: International symposium on neural networks, pp 189–196
10. Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3449–3456
11. Dhole H, Sutaone M, Vyas V (2019) Anomaly detection using convolutional spatiotemporal autoencoder. In: 2019 10th international conference on computing, communication and networking technologies, ICCCNT 2019
12. Dong F, Zhang Y, Nie X (2020) Dual discriminator generative adversarial network for video anomaly detection. IEEE Access 8
13. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 580–587
14. Gong D et al (2019) Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE international conference on computer vision, pp 1705–1714
15. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 733–742
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778
17. He C, Shao J, Sun J (2018) An anomaly-introduced learning method for abnormal event detection. Multimed Tools Appl 77(22):29573–29588
18. Hinami R, Mei T, Satoh S (2017) Joint detection and recounting of abnormal events by learning deep generic knowledge. In: Proceedings of the IEEE international conference on computer vision
19. Hou R, Chen C, Shah M (2017) Tube Convolutional Neural Network (T-CNN) for action detection in videos. In: Proceedings of the IEEE international conference on computer vision, vol 2017-Octob, pp 5822–5831
20. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456
21. 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 computer society conference on computer vision and pattern recognition, pp 7842–7851.
22. Kay W et al (2017) The kinetics human action video dataset. arXiv Prepr. arXiv1705.06950
23. Kim J, Grauman K (2009) Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE conference on computer vision and pattern recognition, pp 2921–2928
24. Kratz L, Nishino K (2009) Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE conference on computer vision and pattern recognition, pp 1446–1453
25. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
26. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2556–2563
27. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32
28. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection - a new baseline. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 6536–6545
29. Liu Y, Liu J, Lin J, Zhao M, Song L (2022) Appearance-motion united auto-encoder framework for video anomaly detection. IEEE Trans. Circuits Syst. II Express Briefs

30. 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
31. Luo W, Liu W, Gao S (2017) Remembering history with convolutional LSTM for anomaly detection. In: IEEE International Conference on Multimedia and Expo (ICME), pp 439–444
32. 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
33. Maqsood R, Bajwa UI, Saleem G, Raza RH, Anwar MW (2021) Anomaly recognition from surveillance videos using 3D convolution neural network. Multimed Tools Appl 80(12):18693–18716
34. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: IEEE conference on computer vision and pattern recognition, pp 935–942
35. Mumtaz A, Sargano AB, Habib Z (2018) Violence detection in surveillance videos with deep network using transfer learning. In: 2nd European Conference on Electrical Engineering and Computer Science (EECS), pp 558–563
36. Mumtaz A, Sargano AB, Habib Z (2020) Fast learning through deep multi-net CNN model for violence recognition in video surveillance
37. Narasimhan MG, Sowmya Kamath S (2018) Dynamic video anomaly detection and localization using sparse denoising autoencoders. Multimed Tools Appl 77(11):13173–13195
38. Nayak R, Pati UC, Das SK (2020) A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis Comput 106:104078
39. Ramachandra B, Jones M (2020) Street scene: a new dataset and evaluation protocol for video anomaly detection. In: The IEEE winter conference on applications of computer vision, pp 2569–2578
40. Ramachandra B, Jones MJ, Vatsavai RR (2020) A survey of single-scene video anomaly detection. IEEE Trans Pattern Anal Mach Intell 44:1–18
41. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) Abnormal event detection in videos using generative adversarial nets. In: Proceedings - International Conference on Image Processing, ICIP, pp 1577–1581
42. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process 26(4):1992–2004
43. Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput Vis Image Underst 172:88–97
44. Saligrama V, Konrad J, Jodoin PM (2010) Video anomaly identification. IEEE Signal Process Mag 27:18–33
45. Sargano AB, Angelov P, Habib Z (2016) Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines. Appl Sci 6(10):309
46. Sargano AB, Wang X, Angelov P, Habib Z (2017) Human action recognition using transfer learning with deep representations. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp 463–469
47. Sargano A, Angelov P, Habib Z (2017) A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Appl Sci 7(1):110
48. Se SAP, Ravanbakhsh M, Nabi M, Mousavi H, Sangineto E, Sebe N (2018) Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. In: Proceedings - 2018 IEEE winter conference on applications of computer vision, WACV 2018
49. Shah AP, Lamare JB, Nguyen-Anh T, Hauptmann A (2019) CADP: a novel dataset for CCTV traffic camera based accident analysis. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–9
50. Shao J, Loy C-C, Kang K, Wang X (2016) Slicing convolutional neural network for crowd video understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5620–5628
51. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos, pp 1–9
52. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition, pp 1–14
53. Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv Prepr. arXiv1212.0402
54. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 6479–6488
55. Tang Y, Zhao L, Zhang S, Gong C, Li G, Yang J (2020) Integrating prediction and reconstruction for anomaly detection. Pattern Recogn Lett 129:123–130
56. 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

57. 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
58. Ullah H, Ullah M, Conci N (2014) Dominant motion analysis in regular and irregular crowd scenes. In: International workshop on human behavior understanding, pp 62–72
59. Ullah W, Ullah A, Hussain T, Khan ZA, Baik SW (2021) An efficient anomaly recognition framework using an attention residual lstm in surveillance videos. Sensors 21(8):2811
60. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, pp 1096–1103
61. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11(12):3371–3408
62. Wang G, Yuan X, Zhang A, Hsu H-M, Hwang J-N (2019) Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In: AI City Challenge Workshop, IEEE/CVF Computer Vision and Pattern Recognition (CVPR) Conference, Long Beach, California, pp 382–390
63. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. In: In British Machine Vision Conference (BMVC), pp 1–3
64. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst 156:117–127
65. Ye M, Peng X, Gan W, Wu W, Qiao Y (2019) AnoPCN: Video anomaly detection via deep predictive coding network. In: Proceedings of the 27th ACM international conference on multimedia, pp 1805–1813
66. Yuan FN, Zhang L, Shi JT, Xia X, Li G (2019) Theories and applications of auto-encoder neural networks: a literature survey. Jisuanji Xuebao/Chinese J Comput 42(1):203–230
67. Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3313–3320
68. Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua XS (2017) Spatio-temporal AutoEncoder for video anomaly detection. Proceedings of the 25th ACM international conference on multimedia, pp 1933–1941
69. 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
70. Zhu Y, Newsam S (2019) Motion-aware feature for improved video anomaly detection 30th Br. Mach. Vis. Conf. 2019, BMVC 2019
71. Zhu Y, Nayak NM, Roy-Chowdhury AK (2013) Context-aware activity recognition and anomaly detection in video. IEEE J Sel Top Signal Process 7(1):91–101
72. Zhu S, Chen C, Sultani W (2020) Video anomaly detection for smart surveillance. arXiv Prepr. arXiv2004.00222