PeerJ Computer Science

Efficient anomaly recognition using surveillance videos

Gulshan Saleem[1], Usama Ijaz Bajwa[1], Rana Hammad Raza[2], Fayez Hussain Alqahtani[3], Amr Tolba[4] and Feng Xia[5]

[1] Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[2] Electronics and Power Engineering Department, Pakistan Navy Engineering College (PNEC), National University of Sciences and Technology (NUST), Karachi, Pakistan
[3] Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[4] Computer Science Department, Community College, King Saud University, Riyadh, Saudi Arabia
[5] School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, Australia

ABSTRACT

Smart surveillance is a difficult task that is gaining popularity due to its direct link to human safety. Today, many indoor and outdoor surveillance systems are in use at public places and smart cities. Because these systems are expensive to deploy, these are out of reach for the vast majority of the public and private sectors. Due to the lack of a precise definition of an anomaly, automated surveillance is a challenging task, especially when large amounts of data, such as 24/7 CCTV footage, must be processed. When implementing such systems in real-time environments, the high computational resource requirements for automated surveillance becomes a major bottleneck. Another challenge is to recognize anomalies accurately as achieving high accuracy while reducing computational cost is more challenging. To address these challenge, this research is based on the developing a system that is both efficient and cost effective. Although 3D convolutional neural networks have proven to be accurate, they are prohibitively expensive for practical use, particularly in real-time surveillance. In this article, we present two contributions: a resource-efficient framework for anomaly recognition problems and two-class and multi-class anomaly recognition on spatially augmented surveillance videos. This research aims to address the problem of computation overhead while maintaining recognition accuracy. The proposed Temporal based Anomaly Recognizer (TAR) framework combines a partial shift strategy with a 2D convolutional architecture-based model, namely MobileNetV2. Extensive experiments were carried out to evaluate the model's performance on the UCF Crime dataset, with MobileNetV2 as the baseline architecture; it achieved an accuracy of 88% which is 2.47% increased performance than available state-of-the-art. The proposed framework achieves 52.7% accuracy for multiclass anomaly recognition on the UCF Crime2Local dataset. The proposed model has been tested in real-time camera stream settings and can handle six streams simultaneously without the need for additional resources.

Subjects Artificial Intelligence, Computer Vision, Neural Networks
Keywords Anomaly recognition, Crime detection, Video surveillance, Video analysis, Deep learning

Submitted 22 April 2022
Accepted 1 September 2022
Published 14 October 2022

Corresponding author
Usama Ijaz Bajwa,
usamabajwa@cuilahore.edu.pk

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.1117**

How to cite this article Saleem G, Bajwa UI, Hammad Raza R, Alqahtani FH, Tolba A, Xia F. 2022. Efficient anomaly recognition using surveillance videos. *PeerJ Comput. Sci.* **8**:e1117 http://doi.org/10.7717/peerj-cs.1117

INTRODUCTION

Surveillance systems continuously monitor ongoing activities in order to avoid or rescue any abnormal events. It is essential to improve the monitoring capability of such a system in order to ensure public safety (*Piza et al., 2019*). In recent years, a large number of CCTV cameras have been installed around the world to record ongoing activity or to serve as a record in the event of a mishap. Manual monitoring of such systems is costly, slow, human biased, and unreliable. Researchers are attempting to automate this process in order to facilitate security personnel and crime prevention authorities. Security surveillance based on computer vision is becoming more popular, and video-based activity recognition is one of them. A large amount of video data from CCTV cameras must be analysed in order to perform anomaly recognition. It is a difficult task because there is no clear distinction between normal and abnormal events because it is dependent on multiple factors and requires subjective definition. Other challenges include low-quality surveillance videos because CCTV typically captures events in low resolution, and most abnormal events occur at a distance. Abnormal events occur infrequently in nature whereas normal events involves huge variations, making annotation of each impossible. As a result of inability to provide adequate annotations and relying solely on one anomaly dataset may result in a high false positive rate. Another challenge is inter-class and intra-class variation problem as there is a thin boundary between normal and abnormal.

Initially, classical machine learning-based approaches attempted to solve these problems (*Mehran, Oyama & Shah, 2009*; *Lu, Shi & Jia, 2013*), but they proved impractical due to the cost of handling large amounts of data as well as computation time. Following the success of deep networks in various computer vision tasks, researchers presented a solution to perform anomaly detection. Currently, there are variety of 3D convolutional neural networks, such as deep neural networks, recurrent neural networks, and autoencoders, which are facilitating security surveillance system (*Mansour et al., 2021*; *Ullah et al., 2021a*; *Ullah et al., 2019*; *Liu et al., 2018b*; *Sun et al., 2018*). The few popular frameworks which have proven their worth are C3D (*Tran et al., 2015*), I3D (*Carreira & Zisserman, 2017*), ResNet 50 (*Yao & Qian, 2018*), and TSN (*Wang et al., 2016*). These algorithms are used to address the different domain challenges, such as intraclass and inter-class variation, where two or more classes have similar characteristics but are diverse. These methods incorporate multiple frames at a time for learning and therefore are better at discriminating classes that are difficult to distinguish using 2D convolution. Such approaches incorporate multiple frames per cycle during learning and forecasting and require a much larger amount of memory and computational resources (*Canizo et al., 2019*; *Azizjon, Jumabek & Kim, 2020*). Some researchers have tried to overcome this by reducing the input frame size and number of frames or by reducing the depth of the convolutional neural network (CNN) architectures but at the cost of accuracy. *Ren et al. (2021)*, have highlighted opportunities and challenges of video anomaly detection and discussed various research directions. They have discussed some major issues, which includes ambiguity, dependency, privacy, noise, sparsity, and diversity.

Overall, majority of the systems are intended to achieve high accuracy on any specified dataset irrespective of computation cost. Recognition accuracy is important parameter to describe strength of a model but if accuracy comes at reduced cost then it becomes a practical solution. We proposed a 2D CNN based cost efficient anomaly recognition framework which is capable of extracting high dimensional data similar to 3D CNN models. It works through preprocessing the input data into frames, which produces normalized, resized, and augmented frames of the original data. These frames are then forwarded to the temporal feature extractor module, which extracts temporal behavior from the data through partial shift approach. Partial shift approach enables data exchange with neighboring frames over temporal dimension.It works by shifting learned features laterally with adjacent frames (*i.e.,* next, and previous frames). Temporal feature extractor is integrated into the residual block of the 2D CNN Baseline model so that Spatial features of input data can be extracted. We have used MobileNetV2 ResNet50 as the 2D backbone architecture in our experiment, which makes the anomaly detection problem more resource efficient. Figure 1 presents the general flow of our system using CCTV footage where Realtime camera stream is the basic source of information that is projected at all monitoring screens. Anomaly recognition model or system process that stream and detect anomalies whenever it occurred. This work used the UCF Crime dataset, which includes long, untrimmed videos with only video level annotation and no information on the temporal segment where the anomaly occurred. The available UCF Crime dataset involves few limitations, as discussed in *Maqsood et al. (2021)*. This work attempts to overcome the highlighted issues through frame level data annotations and spatial augmentation, which improved the performance of video based anomaly recognition. This study aimed at improving spatiotemporal feature based learning as these features provide useful information to process anomalous videos. The following are the contributions of this article:

- We attempted to address the issue of the high resource requirement of the anomaly recognition method and proposed a lightweight, resource-efficient real-time streaming TAR framework that can be embedded on a simple machine like a central processing unit (CPU).
- We proposed to use temporal learning *via* partial shift operation to improve spatiotemporal feature based learning. It enables frames to share their learning among adjacent frames and reduce the cost of processing. Moreover, it helps in building feature maps based on high activity areas that support the classification task of anomaly recognition.
- Our framework is capable of performing online anomaly recognition and it allows six simultaneous screens on a CPU-based system while using fewer parameters (2.2M), FLOPs (0.564GFLOPs), model size (0.6Mb) and low latency overhead which proves it to be resource efficient approach.
- Our model achieves 7.87% and 2.47% increased state-of-the-art accuracy with ResNet-50 and MobileNetV2 respectively on UCF-Crime dataset.

The rest of this article is arranged as follows. In Section II, related work is discussed, and Section III represents the proposed anomaly recognition framework, where the method

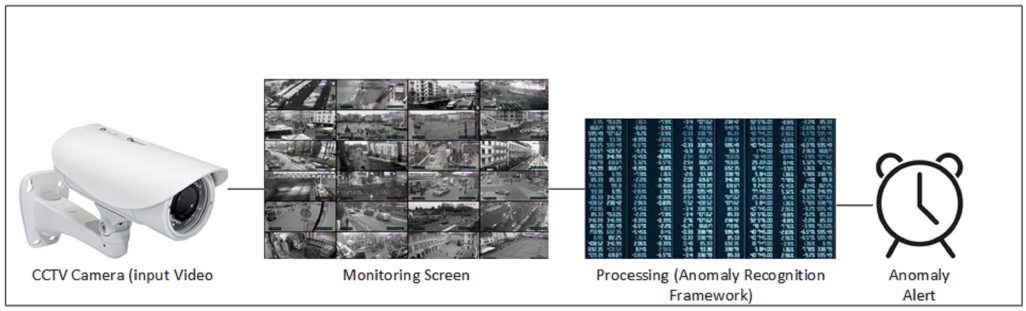

Figure 1 **General data flow of anomaly detection framework.**
Full-size DOI: 10.7717/peerjcs.1117/fig-1

and dataset augmentation is explained. The experimental results and discussion are given in Section IV and the conclusion is given in Section V.

RELATED WORK

A lot of work has been done to perform anomaly detection tasks to support modern security surveillance systems, which include computer vision-based methods and deep learning-based frameworks. Most of these methods are focused on providing high anomaly recognition accuracy using 3D based deep networks which requires huge computation cost. Along with 3D networks, few researchers have worked on improving feature modeling to achieve higher accuracy. The majority of these methods provide reasonable accuracy, but computation cost and time are frequently overlooked, which is the primary focus of this study. We have discussed about some existing methods that are either based on 3D CNN models or describe feature modeling schemes. Both categories are related because we improved the spatiotemporal feature modelling and asserted that our proposed framework outperforms 3D CNN-based methods.

3D convolutional neural networks

3D convolutional neural network based approaches gained popularity due to their promising performance such as in *Sultani, Chen & Shah (2018)*, where the authors have suggested a deep learning-based method for detecting anomalies in real-world surveillance footage. The authors proposed and developed a combination of two methods for detecting anomalies in a real-world video. Multiple instance learning (MIL), which automatically trains the model to construct a deep anomaly-ranking model that predicts high anomaly scores for anomalous video segments inside a video, is used to train their model.The primary contribution of the authors is the preparation of an anomaly dataset (UCF Crime) containing thirteen types of anomalous surveillance videos. *Liu et al. (2018a)* employed a Temporal Convolutional 3D Network (T-C3D), which is capable of real-time action recognition in videos because of its optimized network complexity and, as a result, substantially lower computing cost needs. Their framework for deploying T-C3D directs the network to learn video action representations in a hierarchical multi-granularity way. *Maqsood et al. (2021)* presented a framework for detecting real-world anomalies in

video footage. They have proposed a straightforward but effective method for learning spatiotemporal features by training deep three-dimensional convolutional networks on the UCF Crime video dataset. Whereas, we repurposed a 2D CNN based architecture for our anomaly recognition framework which provides performance of 3D CNN.

Feature modeling

Spatiotemporal modeling

Another approach is to improve spatiotemporal feature extraction process to perform anomaly recognition. *Nawaratne et al. (2019)* have targeted real-time anomaly detection through active/online learning and proposed the incremental spatial-temporal learner (ISTL), which can overcome the issues of both anomaly detection and localization. The proposed learner is evaluated by using a temporal threshold rather than focusing only on spatial threshold to save the contextual information from video feed. In 2019, *Landi, Snoek & Cucchiara (2019)* have worked by considering locality for anomaly detection of real-world videos. They have worked through enriching a considerable portion of an existing dataset with spatial and temporal domains with annotations, using spatiotemporal tubes instead of whole-frame video segments. *Liu et al. (2020)* have performed real time action representation using temporal encoding along with deep compression network. Their framework extracts video representations in a hierarchical granularity manner and use residual 3D CNN to extract appearance-based data. Similarly, *Chu et al. (2018)* have worked on anomaly detection problems by using results of sparse coding results of spatiotemporal features. *Wu et al. (2020)* have focused to improve the high feature extraction process and proposed a fast sparse coding network. Their method is based on the fusion of spatial–temporal features, and it achieves the highest performance at a maximum of 10,000 lower latency. *Chang et al. (2022)* have proposed an autoencoder architecture that is capable of learning spatial and temporal representations from videos. They have used the deep k-mean cluster to fuse the spatiotemporal features, which generate anomaly scores for anomalous events. *Li et al. (2022)* have worked on 3D integral images to perform anomaly detection and performed probability estimation using the Bayesian network. They have used a cube of a video as an event, which is represented as a histogram and then used the motion magnitudes and likelihood from the histogram template to estimate probabilities using prior knowledge.

Attention based modeling

Attention based strategies are also popular which involves extraction of high activity area to perform anomaly recognition. *Ma & Zhang (2022)* have proposed an attention-based framework using deep neural network-based architecture. The proposed framework includes an anomaly attention module (AAM) to score anomalies in a frame. *Muhammad et al. (2022)* have performed fine grained activity recognition to produce short video using video summarization. They have performed key frame extraction through deep CNN based attention modeling. *Ullah et al. (2021a)* have presented a timesaving deep feature-based intelligent anomaly detection framework for anomaly recognition. They have used CNN to extract spatial structure from a sequence of frame, which is then fed into a multi-layer bi-directional long short-term memory (BD-LSTM) model. Then, BD-LSTM

can classify ongoing anomalous events in smart city surveillance setup. *Duman & Erdem (2019)* has worked on anomaly detection tasks through dense optic flow. The optic flow is calculated for input video to extract velocity and direction-based data, which is then passed to a convolutional autoencoder and then to Conv-LSTM encoder to detect anomalies. *Ullah et al. (2021b)* proposed an attention residual LSTM-based lightweight anomaly recognition framework. The authors proposed using a lightweight CNN architecture for feature extraction. They used the MobileNet architecture for frame wise feature extraction, preceded by a sequential learning strategy. *Tang et al. (2020)* have contributed to overcome the limitations of both reconstruction and future prediction methods. They claimed that their proposed framework is robust to noise and outperformed both baseline approaches used for reconstruction and prediction tasks. *Zhong et al. (2022)* have performed anomaly detection using a cascaded reconstruction model along with an optic flow network and used the reconstruction error to choose an anomaly detection framework. *Dong, Zhang & Nie (2020)* have proposed a semi-supervised learning-based framework that can predict future normal frames for a given feed. Their proposed framework works through using two discriminators and a generator to target motion information from videos *i.e.,* optic flow.

Most of the existing work is focused on increasing accuracy of anomaly recognition that usually based on deep networks with high computational requirements. Recently, researchers are considering cost effective anomaly recognition systems, which still requires many improvements. In particular, online anomaly recognition process long untrimmed sequences, which is a challenging task. Our study tries to reduce the computational cost of anomaly recognition process without compromising on accuracy. It processes online streams and resolves the problem of long untrimmed video. We have performed anomaly recognition which identifies whether there is anomaly in data or it is a normal sequence. We used partial shift strategy to incorporate spatiotemporal learning in our framework which improves the accuracy of 2D CNN based architecture to outperform 3D CNN based methods.

TEMPORAL ANOMALY RECOGNIZER (TAR)

Overview

Deep learning-based solutions have outperformed many existing models, but the computational cost of such systems became another bottleneck. In real time scenarios, high resource systems are not useful which means cost of system is equally important as its efficiency. Anomaly recognition is a complex task and processes large amount of data such as 24/7 Real-time camera streams. We have considered both constraints: large amount of data and efficiency of system in terms of timely recognition. Therefore, this work attempts to provide an efficient anomaly recognition framework, which addresses the problem of high resource requirement while working with deep learning based methods. We have used 2D Convolution-based architectures due to their outstanding performance in a wide range of domains. 2D Convolutional networks are good for low-level feature extraction, but they are ineffective for capturing high-level information, such as temporal information (*Lin, Gan & Han, 2019*). As compared to 2D CNN models, 3D CNN models are capable of

extracting high dimensional data but at very high cost especially when we have to process the video stream. This study provides a method to perform 3D computations at cost of 2D model through using data shift operation. For example, videos are multidimensional array (batch size, number of channels, temporal, and spatial points) which requires expensive computations as compare to image. As each video is first converted into frames during prepossessing stage and these frames are representation of source video. 2D CNN avoids temporal dimension during learning process. In contrast, our temporal feature extractor (TFE) performs temporal modelling through shifting channels in both forward and backward direction so that information within frames can be exchanged. The process of shifting channels is explained Temporal feature extractor section. To perform anomaly recognition, we need a model which can extract spatial features of data too so we need a 2D CNN base model for example, MobileNetV2. We propose a temporal anomaly recognizer (TAR) in which we used a partial shift strategy in our method to transfer temporal learning across ConvNet layers. The process of using temporal learning and spatial data is explained in section Anomaly Recognition. So, Temporal Anomaly recognizer which is our proposed framework, is described in terms of of temporal feature extractor and a backbone 2D CNN model.

Figure 2 shows the multidimensional tensor of a video which needs to be processed during anomaly recognition process. Video activations can be termed as N Batch size, C number of channels, T temporal and H,W spatial points. TAR has three main components, which includes preprocessing of data to get normalized and augmented data for training. The other part is temporal feature extractor, which serves the purpose of reducing model size and number of parameters with high speed performance so that our framework can be used in practical scenarios. The third component performs anomaly classification using spatial features of the data and temporal learning of second component. Figure 3 depicts the temporal feature extractor and Fig. 4 the proposed framework (TAR), which is built on top of the residual connection of MobileNetV2. We have experimented with two different models MobileNetV2 and ResNet 50 to observe the performance variation to achieve a computationally cost-effective solution whereas MobileNetV2 is our final backbone model.

Preprocessing

The first step is to divide the video into frames, and data values are scaled by mapping data pixels between 0 and 1. The next step is to resize data frames, and we choose 170x170 after considering other higher resolution scales. We chose this as the final value because there is no significant change in system performance on higher scales. Prior to applying the augmentation technique, the frames are normalized. Data augmentation is a popular approach that is used to increase data examples of a scene without exploiting its meaning. As videos are represented by the sequence of frames which are like image domain problems in various characteristics. Image based methods work within the spatial domain since there is no temporal dimension so linear transformations are the common method to perform data augmentation in such scenarios. Whereas video data augmentation is complex and prone to noise. Such as without considering the temporal dimension, any transformation within

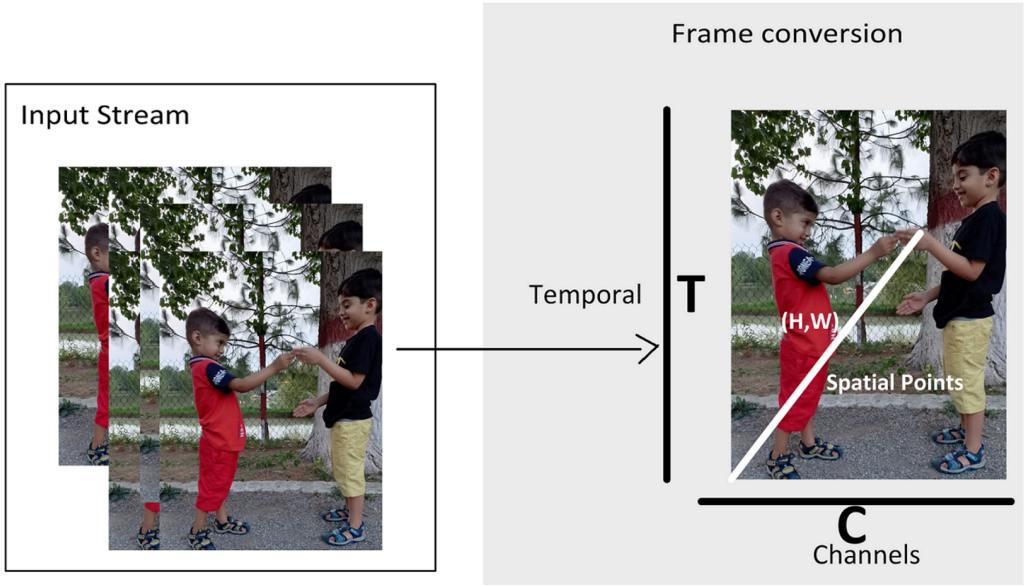

Figure 2 **Video representation in terms of multidimensional array.**
Full-size DOI: 10.7717/peerjcs.1117/fig-2

data can change the entire meaning of the video. The task of video data augmentation can be simplified if video annotations are available which is another open challenge (*Um et al., 2017*; *Cui, Goel & Kingsbury, 2015*). In this study, we used frame-level annotations of data which reduces the complexity of the task and we applied spatial augmentation to our data to increase data example. Our data is first converted into a fixed number of frames, which were then horizontally and vertically flipped. Such scenarios where data is already prone to noise (*i.e.,* anomalies) both horizontal and vertical flips are preferable for data augmentation (*Maqsood et al., 2021*). After performing spatial augmentation, frames are converted to create a video after performing vertical and horizontal flips.

Temporal feature extractor

A video is multidimensional array and expensive to process as compared to 2D images. For 2D image problems, we have plenty of algorithms such as VGG 16, ResNet 101 and CNN. However, this study is based on anomaly recognition from videos to provide security surveillance which requires both spatial and temporal modelling. After preprocessing of data, next step is feature extraction and temporal feature extractor is part of TAR. The intuition behind the method is to use 2D CNN along with a partial shift strategy, which can learn and pass temporal information over the frames. In Fig. 2 T is giving frames and C Channels which are being processed. Convolutional neural network models are built on ConvNet layers and mathematical expression of 2D Convolution is shown in (1):

$$y[a,b] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h[m,n].x[a-m,b-n]. \tag{1}$$

Here h is the kernel matrix with m and n indices whereas is the input image matrix with a and b indices. For input vector A and weight vector W, convolution ($X = Conv(W,A)$)

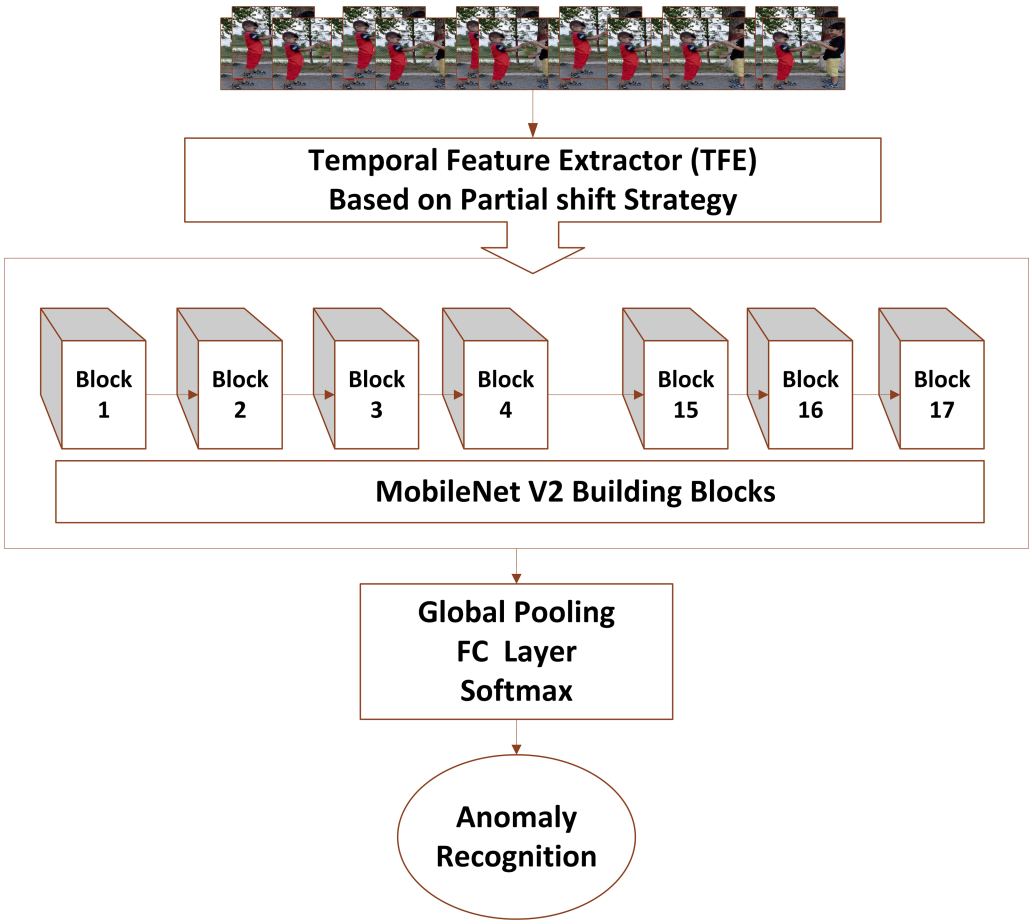

Figure 3 **Video input is forwarded to temporal feature extractor which performs temporal and spatial modeling with the help of MobileNetV2 whereas fully connected layers perform anomaly recognition *via* spatiotemporal modelling.**

Full-size DOI: 10.7717/peerjcs.1117/fig-3

requires weights $(w1, w2, w3)$ which is convolved with input vector A. Hence, Convolution value is as: $X_i = wA_{i-1} + wA_i + wA_{i+1}$. Rather than applying convolution by shifting and multiply-accumulate operations, the input channel has shifted as -1 and $+1$. The shift operation will transfer the information with its neighbor frames (*i.e.*, (A_{i-1}, A_i, A_{i+1}). The shift operation is shown in Fig. 3 and Eqs. (2), (3) and (4):

$$A_i^{-1} = A_{i-1} \tag{2}$$

$$A_i^{0} = A_i \tag{3}$$

$$A_i^{+1} = A_{i-1}. \tag{4}$$

Hence, we can use shift and multiply-accumulate operations in such a way that it will reduce computation cost. Computation cost of a model is basically number of parameters

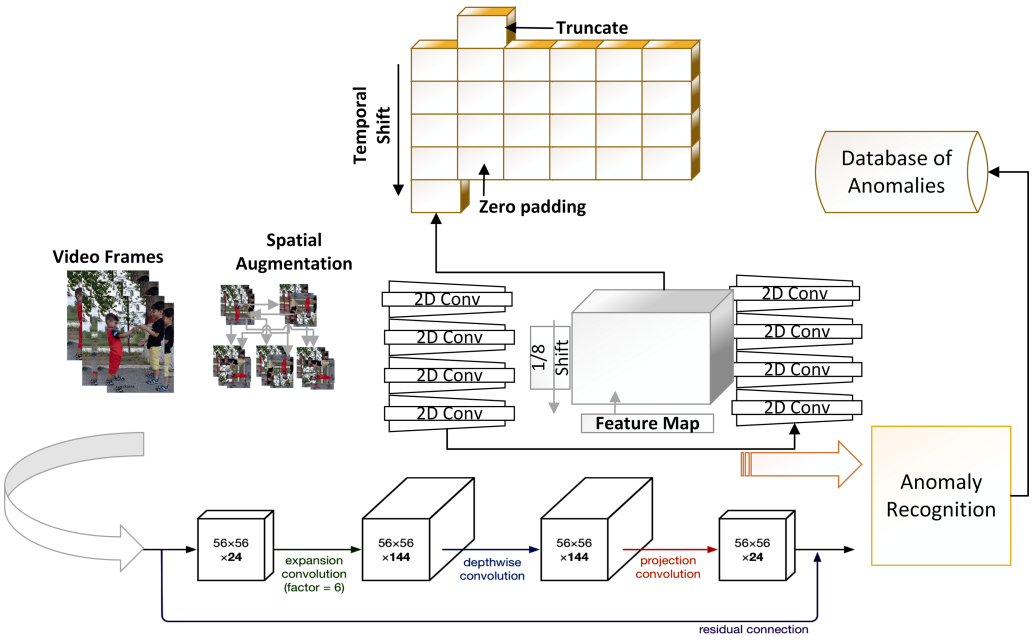

Figure 4 **Proposed temporal based anomaly recognizer (TAR) framework with 2D MobileNetV2 baseline architecture.**

Full-size DOI: 10.7717/peerjcs.1117/fig-4

of a model which need to be processed to perform a task such as anomaly recognition. By considering a 16 frame input sequence, we have a lot of features which needs to be extracted and so multiple channels over temporal dimension. If we do not shift any channels within frames like 2D CNN then we cannot extract temporal behaviour of data. As shifting of channels cause the exchange of temporal learning which is our ultimately goal to extract from data. One solution is to shift all the channels so that we can have maximum temporal learning and eventually our model can achieve a better performance. Shifting channels across the temporal dimension may involves huge data movement which means high cost model and drawback of a system. However, if we partially shift our channels (*i.e.,* 1/8) then it will provide temporal learning as well as reduced cost.

Hence temporal feature extractor shifts channels in both direction +1 in one direction and −1 in other to exchange data in neighbourhood frames. This allows us to extract high-dimensional data features (temporal data), which is impossible with 2D CNN. Along with temporal features, spatial features are also important because temporal features describe changes over time, whereas spatial features extract interactions within a frame. Shifting channels can result in data loss because the information from the current frame is no longer available, resulting in reduced spatial features. The shift operation is associated with the convNet layers of the 2D CNN model, and it is necessary to understand how the temporal feature extractor (TFE) collaborates with 2D CNN to perform anomaly detection. Consider the convNet layers of MobileNetV2 which means shifting can be done before convNet layers (outside of residual connections) or within residual connections. Using channel shift outside of the residual branch may result in spatial data loss because we

cannot access the original data of the frame. However, if we apply shift within the residual branch, we can still access the original data of a frame, indicating that spatial data is not jeopardized. The 2D CNN model aids in the extraction of spatial features, while partial shift strategy aids in the extraction of temporal features.

Anomaly recognition

Temporal and spatial modeling provides temporal and spatial behaviour of data which is then forwarded to perform recognition using a classification algorithm. The method used to perform anomaly recognition is based on temporal feature extractor module and 2D convolutional architecture *e.g.*, ResNet-50 and MobileNetV2. The framework takes video as input that is converted into frames and then data augmentation modules perform spatial augmentation of data. Spatial augmentation produces new training examples to increase the amount of labeled data. Then, the temporal feature extractor learns and transfers the temporal information of each frame to its neighborhood frames. The temporal feature extractor used the strategy as discussed under temporal feature extractor heading. As shown in Fig. 4, after performing preprocessing and data augmentation, our model performs feature extraction to extract useful data from anomalous videos. In Fig. 4, a tensor is shown which shows channel shift in both direction along temporal dimension. This represents Deep features are extracted through a combination of MobileNetV2 model and partial shift strategy.

Within the architecture of MobileNetV2, there are two type of blocks: residual blocks and downsizing blocks. The stride of a residual block is one, but the stride of a downsizing block is two. Each block comprises three layers: a convolutional layer with a non-linear cost function, a depth wise convolution layer, and a convolutional layer without a non-linear cost function. We proposed using a partial shift function within MobileNetV2′s residual blocks to extract both spatial and temporal data. To lower the model's latency overhead, the model stores only 1/8 of the feature set for an incoming frame to reduce the computation cost of the framework. The partial shift function changes the information in the current frame along temporal dimensions and exchanges the learned parameters with associated or neighboring frames. The next component is anomaly identification, which uses features to divide data into anomaly and normal classes. The final feature map, which is used by the classification layer, is improved further by applying a variance-based filter, which analyzes the variance of extracted features and highlights those with high variations. This method will improve the final feature map, which will be utilised to identify anomalies during the classification stage. Anomaly events are saved in an anomaly database after they have been identified based on recognition results. The identified anomalous events are saved using hardware based implementation as we have tested our method by designing a surveillance application using Jetson Nano. However, our proposed framework is limited to perform anomaly recognition task only.

MobileNetV2 architecture

For anomaly recognition, we propose using both temporal and spatial features. As a result, MobileNetV2 is used for spatial feature extraction and anomaly classification (*i.e.,* anomaly,

normal). MobileNetV2 is a popular convolutional-based architecture that is specifically designed to operate on low-resource systems. The MobileNetV2 has the same residual block as the V1, but it also has an expansion layer, projection layer, and depth wise convolution layer. The bottlenecks residual block is a MobileNetV2 building block that shrinks the input to reduce computation. Our framework is built on top of MobileNetV2, which provides lower computation costs. Temporal feature extractors are useful for transferring temporal learning, but spatial learning of video frames is also important and should be saved for prediction. The bottleneck residual block's function is to extract spatial features from video frames. As a result, the model can extract both spatial and temporal features from video data while maintaining a reasonable recognition accuracy.

EXPERIMENT

Dataset description

The objective of anomaly detection from surveillance videos is to find data points or elements that do not meet the criteria of a normal class. We require data that can depict the problem at hand during the model evaluation phase. There are many anomaly-based datasets available; however, the most of them are actor-based, which are valuable for experiments but limited in their ability to capture real-time variations of a scene. For this study, we have used UCF Crime dataset (*Sultani, Chen & Shah, 2018*) and UCFCrime2local dataset (*Landi, Snoek & Cucchiara, 2019*), which are based on real-time surveillance videos.

UCF Crime (*Sultani, Chen & Shah, 2018*) is a large scale video dataset of untrimmed videos, which means an anomalous video may have few frames which are part of anomaly. It is based on 13 real-time anomalies, which are arson, assault, fight, robbery, burglary, accident, vandalism, explosion, abuse, arrest, shoplifting, shooting, and stealing. The last and 14th class contains normal events only. The dataset offers few limitations due to which the performance of a system can be compromised, such as anomalous video may have more normal frames than the anomaly frames. This issue can cause model biases over normal events. The other issue is an intra-class variation which may cause over fitting of the model (*Maqsood et al., 2021*). UCF Crime2Local (*Landi, Snoek & Cucchiara, 2019*) is the subpart of UCF Crime dataset and contains six anomalous classes which are burglary, robbery, assault, vandalism, arrest, and stealing. The UCF Crime2Local provides the temporal annotations for anomalies that are missing in UCF Crime.

Data processing

UCF Crime only provides annotations at the video frame level, which means that each video frame has the same label. We have spatially annotated video data, which means that each frame has relevant annotation. This process also improves training by providing enough labelled data and a training set based on anomalous frames. The dataset necessitates frame-level annotations and has a class imbalance issue. To perform data augmentation and resolve class imbalance issues, this study employs a preprocessing strategy as discussed in previous Preprocessing section. We require a large amount of data for model training, which has been generated using data augmentation strategies. Data augmentation generates new data samples with correct labels as well as view invariants of the same frame (*i.e.,* image).

This study employs spatial augmentation techniques to perform data augmentation, as well as frame level data annotation. Each video is divided into frames, which are then used to perform vertical and horizontal flips for data augmentation.

Training and testing

The training split of the UCF Crime dataset has been expanded to include more training examples, while the testing split remains the same as provided by CRCV. The testing split consists of 168 temporally annotated anomalous and normal videos, whereas our augmented training split consists of 1120 (560 normal, 560 abnormal) videos. On Colab, we trained the deep neural network with a Tesla T4/K80 GPU. The temporal feature extractor operates in conjunction with two different 2D CNN baseline models, Resnet50 and MobileNetV2. The training strategy included disk-resident frame extraction with a python-based directory/file parser and OpenCV for automated dataset labelling (normal *vs.* abnormal).

To achieve the desired output, the training phase of the anomaly recognition framework includes data handling and hyperparameter tuning as given in Table 1. The video frames from the UCF Crime dataset are used, with training epochs set to 100 and batch size set to 16. Batch normalization is carried out using stochastic gradient descent, with the model's learning rate set to 0.01, and dropout set to 0.5 with a decay of 0.1 after 50 epochs.To improve the feature extraction process, we used a MobileNetV2 model that had been pretrained on the imageNet dataset. The same setup is used to experiment with ResNet50 on the UCF Crime dataset. Experiments are carried out on PyTorch platforms and data modeling libraries.

RESULTS

Comparison with state-of-the-art methods

In this article, we repurposed a 2D CNN model to extract high-dimensional data such as temporal information, which is important in video-based anomaly recognition. In the preceding section, we thoroughly discussed the proposed TAR and our experimental setup. We have compared its performance to other state-of-the-art methods, such as *Ullah et al. (2021b)* have used an attention residual LSTM-based lightweight anomaly recognition framework. They proposed using a lightweight CNN architecture for feature extraction. The extracted features have fed into a residual attention LSTM, which has trained to learn contextual information in abnormal activities. Moreover, the authors claimed that residual attention LSTM is 10% more efficient in terms of learnable parameters than conventional LSTM. *Ullah et al. (2021a)* have worked on a time saving deep feature-based intelligent anomaly detection framework for anomaly recognition. They have extracted spatial structure from a sequence of frames using a pre-trained convolutional neural network (CNN) model. Then, BD-LSTM can classify ongoing anomalous events in smart city surveillance setup. The authors have experimented with their proposed framework on UCF Crime and the UCF Crime2Local dataset. *Biradar, Dube & Vipparthi (2018)* have proposed DEARESt for aberrant behavior detection in surveillance videos. DEARESt employs a two-stream network to extract appearance and motion flow features separately,

Table 1 Hyperparameters settings of experiment.

Hyperparameter	Hyperparameter value
Batch size	16
Epochs	100
Learning rate	0.01 (decays by 0.1 at epoch 40 80)
Dropout	0.5

from a video stream. Then, these features are concatenated to form a single feature vector that is further used to classify a video.

Convolutional neural network based models are getting popular in computer vision applications. Most of the systems are based on deep networks, which made them resource dependent. This study proposed to use resource efficient strategy without compromising on accuracy of the system. MobileNetV2 and ResNet-50 are used as the experiments' backbone 2D convolutional architecture. However, due to its resource efficiency, the MobileNetV2 architecture is considered the final model for the proposed study. MobileNetV2 is used for spatial feature extraction and in conjunction with a temporal feature extractor (TFE) for temporal modeling. We conducted experiments on two benchmark datasets, UCF Crime and UCF Crime2Local. The study compared the performance of UCF Crime's testing split to that of existing methods. During experiments, we also used ResNet-50 to investigate differences in accuracy and resource utilization.

As shown in Table 2, MobileNetV2 outperforms in terms of resource utilization with compromising only a minor amount of accuracy. Table 2 includes different measures to compare the performance of our model with other methods, such as accuracy, precision, recall, model parameters, model size, and FLOPs of the model. We trained the proposed framework with two different architectures *i.e.,* ResNet50 and MobileNetV2. The MobileNetV2 provides accuracy of 88% with a loss of 0.201 and ResNet provides accuracy of 93.4% with a loss of 1.2101. We have also drawn a comparison of high resource system and low resource system based performance. In Figs. 5 and 6, we have presented confusion matrix of proposed framework with both ResNet50 and MobileNetV2 models. Confusion matrix describes classification behaviour of a model such as in terms of true positive (TPR) and false positive rate (FPR). As shown in the figures, ResNet50 achieves better performance in terms of recognition rate as compared to MobileNetV2 but we selected MobileNetV2 as final baseline architecture as our agenda is to provide resource efficient framework. Figs. 7 and 8 shows accuracy and loss curve of proposed anomaly recognition framework on UCF crime dataset with MobileNetV2 as baseline model. We have implemented our framework using NVIDIA Jetson Nano to compare performance for edge devices with CPU and GPU. Moreover, our method provides fast computation speed with time complexity of 0.198 s. While discussing performance of our method, its resource efficiency is also notable. There is a difference in the number of frames processed per second, so GPU based systems are fast in prediction. CPU based systems usually have slow processing speed, but it does not affect the recognition rate. The prevalence of CPU-based systems in real-time scenarios is the primary motivation for proposing a resource-efficient

Table 2 Comparison of the results achieved using ResNet50 and MobileNetV2 architecture.

Method	Accuracy	Precision	Recall	Model size (MBs)	Parameters	Mega FLOPs
Attention Residual LSTM (*Ullah et al., 2021b*)	78.43%	87%	78%	12.8	3.3M	618.3
DEARESt (*Biradar, Dube & Vipparthi, 2018*)	76.786%	–	–	1187.5	305M	–
ResNet50+multi-layer BD-LSTM (*Ullah et al., 2021a*)	85.53%	–	–	143	25M	–
TAR(Baseline ResNet50)	93.4%	97.8%	89%	91.2	23.5M	6768
TAR(Baseline MobileNetV2)	88%	92.2%	83%	8.61	2.2M	564

Figure 5 Confusion matrix of proposed framework with ResNet50 as the 2D CNN baseline model.
Full-size DOI: 10.7717/peerjcs.1117/fig-5

system. Our model work efficiently on both CPU and GPU and produces low latency of 42.1 ms and 12.01 ms, respectively. That is why, we have claimed that our proposed framework (TAR) is resource efficient.

The proposed system is also extended for a desktop application to serve the purpose of security surveillance as presented in Fig. 9. It has functionality to store the online streams as a database, which stores only anomalous frames from camera streams as shown in Fig. 10. The purpose of managing a database of anomalies is twofold, *i.e.,* it can be used for the forensic purpose to serve as proof of an event, and it can be used as training data to improve the anomaly recognition rate.

Multi-Class problem

We have performed a multi-class anomaly recognition task on UCF Crime2Local dataset, which has six types of anomalies *e.g.*, Arrest, Assault, Burglary, Robbery, Stealing, and Vandalism. Figure 11 provides the confusion matrix for different classes of UCF Crime2Local, which shows values are comparatively low for classes arrest and burglary. It

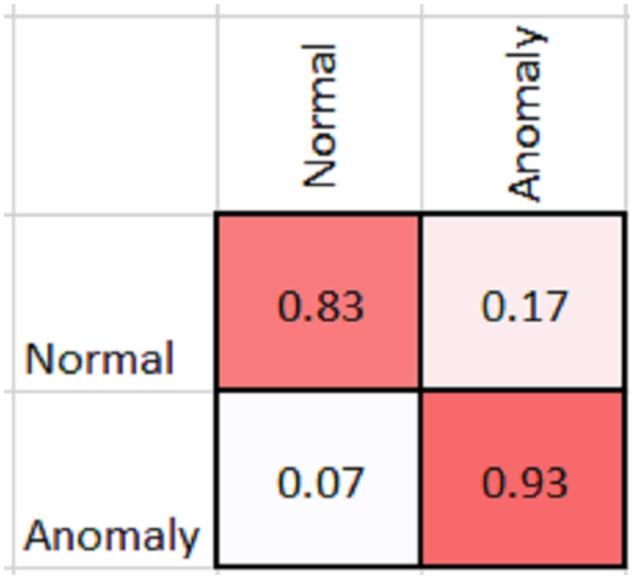

Figure 6 **Confusion matrix of proposed framework with MobileNetV2 as 2D CNN baseline model.**
Full-size DOI: 10.7717/peerjcs.1117/fig-6

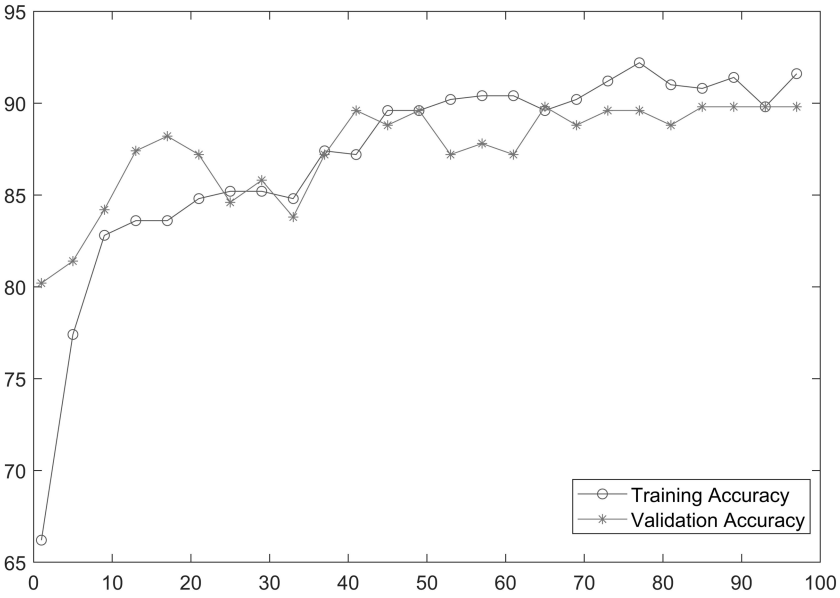

Figure 7 **Accuracy of proposed framework.**
Full-size DOI: 10.7717/peerjcs.1117/fig-7

is because arrest videos usually involve confused events, as arrest act is not clear. Similarly, Burglary act involves clear visibility of event to represent illegal entry. Robbery has also low value because it is sometimes learnt as stealing or assault class. Table 3 provides the comparison of multi-class problem with some baseline methods and our proposed method achieve an average of 52.7% accuracy. The achieved performance is good as it provides a

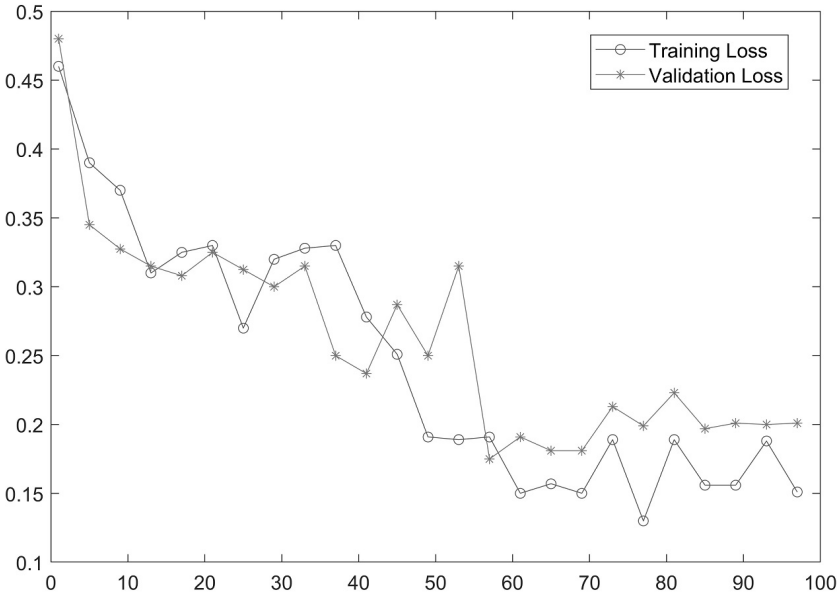

Figure 8 Loss curve of proposed framework.
Full-size DOI: 10.7717/peerjcs.1117/fig-8

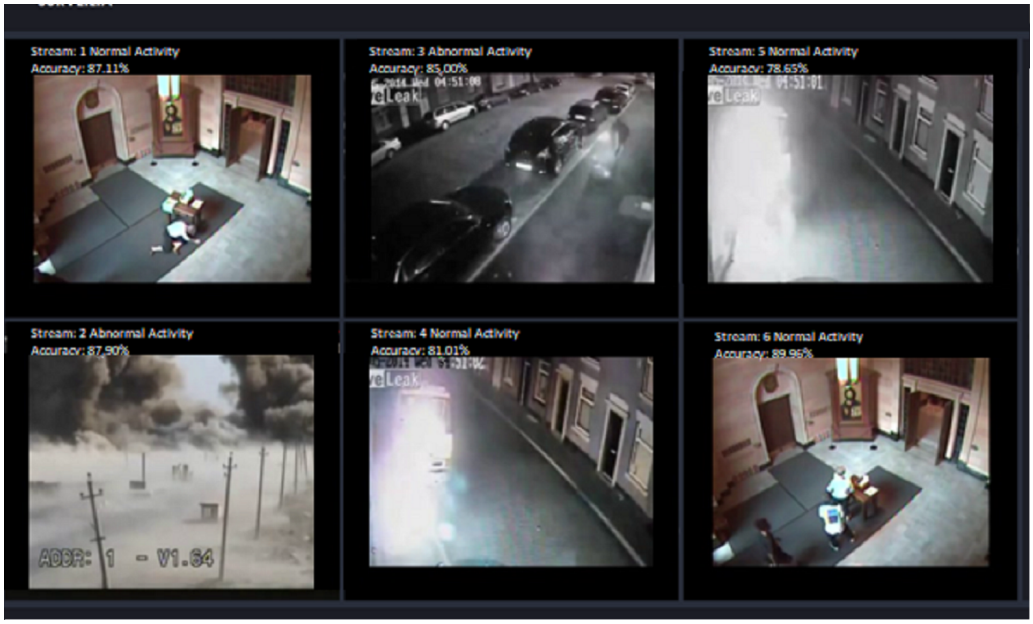

Figure 9 Performance of our desktop application based on proposed framework (TAR).
Full-size DOI: 10.7717/peerjcs.1117/fig-9

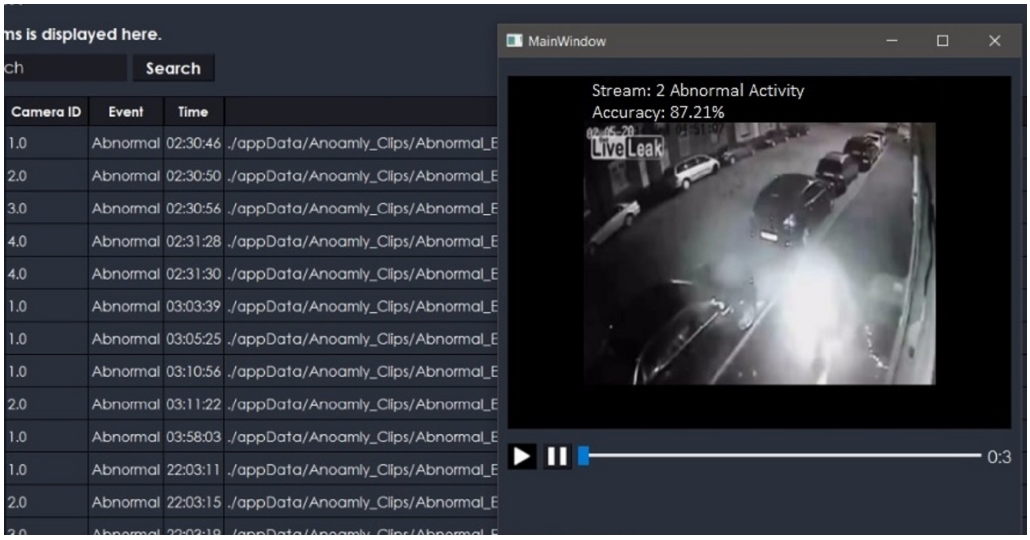

Figure 10 **Desktop application stores anomaly clips.**
Full-size DOI: 10.7717/peerjcs.1117/fig-10

Table 3 **Performance of proposed framework (TAR) for multiclass anomaly recognition.**

Method	Accuracy
ResNet50	45.20%
MobileNetV2	41.9%
Proposed method	52.70%

reasonable accuracy and do not require high computational resources. This makes it useful for real-time application, although our focus is to perform two class recognition to identify whether it is an anomaly or not. Overall proposed framework provides suitable accuracy with reduced computational cost.

CONCLUSION

Automated surveillance is popular area which is continuously improving over the time. Such systems are designed to process huge amount of data which requires a lot of resources which is a challenging requirement. In practical scenarios, time, and cost both are critical to handle and hence a system needs a lot of computation time and resources to serve the purpose. Therefore, this research aimed at addressing these problems through providing a resource-efficient, high-performing system for anomaly recognition. Increased numbers of CCTV generates vast amount of unlabelled video data and its labeling is a difficult task. Moreover, large amount of data requires substantial computational resources to process it. This study provides a lightweight and cost-effective approach for anomaly recognition in terms of memory consumption, processing parameters, and computation time that is essential for a low-resource systems, such as CPU-based systems. The proposed framework is based on 2D convolutional architecture (2D CNN), with a spatiotemporal feature extractor which functions as a partial shift that learns and distributes temporal information

Classes	Arrest	Assault	Burglary	Normal	Robbery	Stealing	Vandalism
Arrest	**0.214**	0.047	0.034	0.0102	0.01	0	0
Assault	0	**0.57**	0	0.026	0	0	0
Burglary	0.36	0	**0.268**	0	0.33	0.311	0.0791
Normal	0	0	0	**0.86**	0	0	0.33
Robbery	0	0	0	**0**	**0.48**	0	0.008
Stealing	0	0.01	0	0	0.24	**0.631**	0
Vandalism	0	0	0	0	0	0.0607	**0.67**

Figure 11 **Multiclass confusion matrix of temporal based anomaly recognizer (TAR).**
Full-size DOI: 10.7717/peerjcs.1117/fig-11

among its neighborhood frames. MobileNetV2 baseline performs spatial feature extraction, which is then combined with temporal learning to perform anomaly recognition. Our proposed framework works with low latency rate of 12.01 ms which makes it effective for performing online video recognition and handle up to six streams at once. On the UCF Crime dataset, the proposed framework achieves an accuracy of 88% for binary anomaly recognition problem with time complexity of 0.198 s. On the UCF Crime2Local dataset, the proposed framework achieves accuracy of 52.7 percent for a multi-class problem. The model outperforms previous models in terms of computational parameters requiring 2.2 M parameters and 0.564 GFLOPs with MobileNetV2 as the baseline architecture. Moreover, our proposed framework has achieved an increased accuracy of 2.47% on UCF Crime dataset with reduced computational requirement. Overall, it performs well in a lot of aspects and can be used for realtime recognition but it can be further improved. Some limitations of this study are highlighted in below section to consider in future.

LIMITATIONS AND FUTURE DIRECTIONS

We have proposed a resource-efficient anomaly recognition system that effectively performs recognition tasks, but it is not evaluated for object level detection and tracking. Object detection and tracking can significantly improve security surveillance. We believe that with

minor modifications to the current model, it could be useful for detection and tracking as well. Our model performs recognition with adequate speed efficiency, but its early response efficiency can be investigated further. It requires validating a model's ability to perform recognition on a small sample of incoming frames as soon as possible to improve system's response time. Surveillance systems are designed to perform anomaly recognition as precisely as possible, but there is an additional issue posed by the false positive rate, which can compromise system reliability. To increase the usefulness of our model, we will strive to reduce the false positive rate as much as possible.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Gulshan Saleem, Usama Ijaz Bajwa, and Rana Hammad Raza received support for this study from the National Center of Big Data and Cloud Computing (NCBC) and the HEC of Pakistan. Fayez Hussain Alqahtani and Amr Tolba received funding for this work from the Researchers Supporting Project No. (RSP2022R509) at King Saud University, Riyadh, Saudi Arabia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The National Center of Big Data and Cloud Computing (NCBC) and the HEC of Pakistan.
The Researchers Supporting at King Saud University, Riyadh, Saudi Arabia: RSP2022R509.

Competing Interests

Feng Xia is an Academic Editor for PeerJ.

Author Contributions

- Gulshan Saleem conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Usama Ijaz Bajwa conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Rana Hammad Raza conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Fayez Hussain Alqahtani conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Amr Tolba conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Feng Xia conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub: https://github.com/GulshanSaleem/Surveilia.

The data is available at Zenodo: Saleem, Gulshan, Bajwa, Usama Ijaz, Raza, Rana Hammad, Alqahtani, Fayez, Tolba, Amr, & Xia, Feng. (2022). Efficient Anomaly Recognition Using Surveillance Videos (V1.1). Zenodo. https://doi.org/10.5281/zenodo.7047216.

The UCF Crime dataset was used for training and testing. The video from the UCF Crime dataset is available at: https://www.crcv.ucf.edu/projects/real-world/.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1117#supplemental-information.

REFERENCES

Azizjon M, Jumabek A, Kim W. 2020. 1D CNN based network intrusion detection with normalization on imbalanced data. In: *2020 international conference on artificial intelligence in information and communication (ICAIIC)*. Piscataway: IEEE, 218–224.

Biradar K, Dube S, Vipparthi SK. 2018. DEARESt: deep Convolutional aberrant behavior detection in real-world scenarios. In: *2018 IEEE 13th international conference on industrial and information systems (ICIIS)*. Piscataway: IEEE, 163–167.

Canizo M, Triguero I, Conde A, Onieva E. 2019. Multi-head CNN–RNN for multi-time series anomaly detection: an industrial case study. *Neurocomputing* **363**:246–260 DOI 10.1016/j.neucom.2019.07.034.

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*. Piscataway: IEEE, 6299–6308.

Chang Y, Tu Z, Xie W, Luo B, Zhang S, Sui H, Yuan J. 2022. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition* **122**:108213 DOI 10.1016/j.patcog.2021.108213.

Chu W, Xue H, Yao C, Cai D. 2018. Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos. *IEEE Transactions on Multimedia* **21(1)**:246–255.

Cui X, Goel V, Kingsbury B. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23(9)**:1469–1477 DOI 10.1109/TASLP.2015.2438544.

Dong F, Zhang Y, Nie X. 2020. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access* **8**:88170–88176 DOI 10.1109/ACCESS.2020.2993373.

Duman E, Erdem OA. 2019. Anomaly detection in videos using optical flow and convolutional autoencoder. *IEEE Access* **7**:183914–183923 DOI 10.1109/ACCESS.2019.2960654.

Landi F, Snoek CG, Cucchiara R. 2019. Anomaly locality in video surveillance. ArXiv preprint. arXiv:1901.10364.

Li S, Cheng Y, Liu Y, Yang Y. 2022. Fast anomaly detection based on 3D integral images. *Neural Processing Letters* **54(2)**:1465–1479.

Lin J, Gan C, Han S. 2019. Tsm: temporal shift module for efficient video understanding. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Piscataway: IEEE, 7083–7093.

Liu K, Liu W, Gan C, Tan M, Ma H. 2018a. T-C3D: temporal convolutional 3D network for real-time action recognition. In: *Proceedings of the AAAI conference on artificial intelligence, volume 32*.

Liu K, Liu W, Ma H, Tan M, Gan C. 2020. A real-time action representation with temporal encoding and deep compression. *IEEE Transactions on Circuits and Systems for Video Technology* **31(2)**:647–660.

Liu W, Luo W, Lian D, Gao S. 2018b. Future frame prediction for anomaly detection–a new baseline. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 6536–6545.

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*. Piscataway: IEEE, 2720–2727.

Ma H, Zhang L. 2022. Attention-based framework for weakly supervised video anomaly detection. *The Journal of Supercomputing* **78(6)**:8409–8429.

Mansour RF, Escorcia-Gutierrez J, Gamarra M, Villanueva JA, Leal N. 2021. Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model. *Image and Vision Computing* **112**:104229 DOI 10.1016/j.imavis.2021.104229.

Maqsood R, Bajwa UI, Saleem G, Raza RH, Anwar MW. 2021. Anomaly recognition from surveillance videos using 3D convolution neural network. *Multimedia Tools and Applications* **80(12)**:18693–18716 DOI 10.1007/s11042-021-10570-3.

Mehran R, Oyama A, Shah M. 2009. Abnormal crowd behavior detection using social force model. In: *2009 IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 935–942.

Muhammad W, Ahmed I, Ahmad J, Nawaz M, Alabdulkreem E, Ghadi Y. 2022. A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system. *PeerJ Computer Science* **8**:e911 DOI 10.7717/peerj-cs.911.

Nawaratne R, Alahakoon D, De Silva D, Yu X. 2019. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics* **16(1)**:393–402.

Piza EL, Welsh BC, Farrington DP, Thomas AL. 2019. CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology & Public Policy* **18(1)**:135–159 DOI 10.1111/1745-9133.12419.

Ren J, Xia F, Liu Y, Lee I.. 2021. Deep Video Anomaly Detection: opportunities and Challenges. In: *2021 international conference on data mining workshops (ICDMW)*. IEEE, 959–966.

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*. Piscataway: IEEE, 6479–6488.

Sun J, Wang X, Xiong N, Shao J. 2018. Learning sparse representation with variational auto-encoder for anomaly detection. *IEEE Access* **6**:33353–33361 DOI 10.1109/ACCESS.2018.2848210.

Tang Y, Zhao L, Zhang S, Gong C, Li G, Yang J. 2020. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters* **129**:123–130 DOI 10.1016/j.patrec.2019.11.024.

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*. Piscataway: IEEE, 4489–4497.

Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW. 2019. Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors* **19(11)**:2472 DOI 10.3390/s19112472.

Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW. 2021a. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimedia Tools and Applications* **80(11)**:16979–16995 DOI 10.1007/s11042-020-09406-3.

Ullah W, Ullah A, Hussain T, Khan ZA, Baik SW. 2021b. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors* **21(8)**:2811 DOI 10.3390/s21082811.

Um TT, Pfister FM, Pichler D, Endo S, Lang M, Hirche S, Fietzek U, Kulić D. 2017. Data augmentation of wearable sensor data for parkinsons disease monitoring using convolutional neural networks. In: *Proceedings of the 19th ACM international conference on multimodal interaction*. New York: ACM, 216–220.

Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Gool LV. 2016. Temporal segment networks: towards good practices for deep action recognition. In: *European conference on computer vision*. Cham: Springer, 20–36.

Wu P, Liu J, Li M, Sun Y, Shen F. 2020. Fast sparse coding networks for anomaly detection in videos. *Pattern Recognition* **107**:107515 DOI 10.1016/j.patcog.2020.107515.

Yao L, Qian Y. 2018. Dt-3dresnet-lstm: an architecture for temporal activity recognition in videos. In: *Pacific rim conference on multimedia*. Cham: Springer, 622–632.

Zhong Y, Chen X, Jiang J, Ren F. 2022. A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos. *Pattern Recognition* **122**:108336 DOI 10.1016/j.patcog.2021.108336.