Real World Anomalous Scene Detection and Classification using Multilayer Deep Neural Networks

Atif Jan[1]*, Gul Muhammad Khan[2]

[1] Department of Electrical Engineering, University of Engineering & Technology Peshawar (Pakistan)
[2] National Center of Artificial Intelligence, University of Engineering & Technology Peshawar (Pakistan)

Received 30 May 2021 | Accepted 1 September 2021 | Early Access 31 October 2021

Abstract

Surveillance videos record malicious events in a locality utilizing various machine learning algorithms for detection. Deep-learning algorithms being the most prominent AI algorithms are data-hungry as well as computationally expensive. These algorithms perform better when trained over a diverse and huge set of examples. These modern AI methods have a dire need of utilizing human intelligence to pamper the problem in such a way as to reduce the ultimate effort in terms of computational cost. In this research work, a novel methodology termed Bag of Focus (BoF) based training methodology has been proposed. BoF is based on the concept of selecting motion-intensive blocks in a long video, for training different deep neural networks (DNN's). The methodology reduced the computational overhead by 90% (ten times) in comparison to when full-length videos are entertained. It has been observed that training networks using BoF are equally effective in terms of performance for the same network trained over the full-length dataset. In this research work, firstly, a fine-grained annotated dataset including instance and activity information has been developed for real-world volume crimes. Secondly, a BoF-based methodology has been introduced for effective training of the state-of-the-art 3D, and 2D Convolutional Neural Networks (CNNs). Lastly, a comparison between the state-of-the-art networks have been presented for malicious event recognition in videos. It has been observed that 2D CNN even with lesser parameters achieved a promising classification accuracy of 98.7% and Area under the curve (AUC) of 99.7%.

Keywords

Volume Crime Classification, Volume Crime Detection, Malicious Activity Detection, Deep Learning.

DOI: 10.9781/ijimai.2021.10.010

I. Introduction

NOWADAYS surveillance cameras are installed at decisive locations across the city. The surveillance videos can record various malicious activities in its locality. To reduce the impact of crimes recorded by CCTV cameras, timely detection of the activity is needed for prompt actions by the concerned authorities. The network of CCTV cameras is monitored through a central control room operated round-the clock by human observers. However, firstly a large expert task force is required to monitor these hundreds of video streams. Secondly, the probability of detecting anomalous activities decreases with an increase in the number of video streams and the time of attention. According to [1] an operator may efficiently monitor a video stream for about 12 minutes continuously, after which he may miss up to 45% of screen activity. After 22 minutes this miss-rate may even elevate to 95%. Thus, to improve the efficiency of surveillance systems, various technological solutions have been adopted to continuously analyze the CCTV recordings. The solutions are based on various machine learning algorithms. Deep-learning being the most effective earning technique is data-hungry and required huge computation. Currently, deep neural networks are trained over a large set of videos for understanding motion information. Training on a large set of full-length videos is a computationally expensive problem. So, a human intelligence-based solution is needed to design data for algorithms in such a way as to avoid redundant information in the full-length video, to reduce the computational overhead for training DNN's. Considering the importance of human intelligence-based methodology for training DNN's, we propose BoF methodology. Comprising of a dataset for the identification of volume crimes in public places for training learning systems has been introduced.

Various approaches have been entertained to develop a system for automatic detection of abnormal behaviors in CCTV recording. The initial studies in abnormal event detection were focused on object tracking [2], [3], [4], where a moving object is considered abnormal if its trajectory doesn't follow the fitted model during the training period. Trajectory analysis can perform well in the case of an individual moving object in a scene but is less effective for complex and crowded scenes. Such efforts are less effective in tracking the motion of abnormal shapes. Handcrafted feature extraction techniques are also exploited for anomaly detection [5], [6]. The fundamental problem with the mentioned approach is that the selection of efficacious features, which was resolved through deep features by Gong et al [7]. They have used unsupervised deep learning-based features for addressing anomaly detection. Usually, deviation from the normal is considered as an anomaly in unsupervised learning; however, this may not be the case due to the existence of a _ne line between normal and

* Corresponding author.
E-mail address: aatifjan@uetpeshawar.edu.pk

Please cite this article in press as:
A. Jan, G. M. Khan. Real World Anomalous Scene Detection and Classification using Multilayer Deep Neural Networks, International Journal of Interactive Multimedia and Artificial Intelligence, (2021), http://dx.doi.org/10.9781/ijimai.2021.10.010

abnormal behavior, which results in a large number of false alarms. The strongest approach used so far is supervised deep learning algorithms. In supervised deep learning techniques, various labeled datasets are used for the detection of a particular group of activities. The latest approach used by Sultani et al. is Multiple Instance Learning (MIL) for real-world anomaly detection. They introduced an assorted dataset of 13 real-world malicious activities. They managed to achieve a classification accuracy of 28%. The developed dataset contains a very diverse set of classes. Nevertheless, class labels are assigned to whole videos while only a part of these videos contained the occurrence of the actual event. This causes MIL to perform poorly leading to low classification accuracy. We in this research, acquired a subset of the above mentioned dataset to address volume crimes consisting of four frequently occurring abnormalities (assault, fight, shooting, vandalism). Some of these examples are illustrated in Fig. 1.

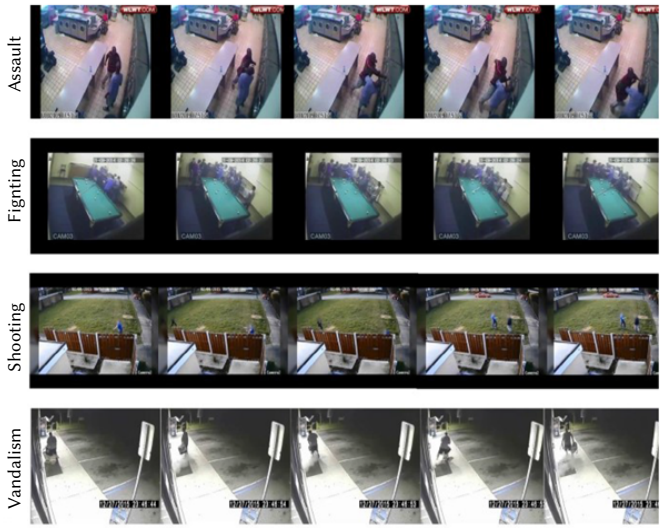


Fig. 1. Sequence of frames taken from video stream can related to different malicious events e.g. Assault, Fighting, Shooting, and Vandalism.

State-of-the-art models implemented for understanding motion information in videos adopt the methodology of training network on full-length videos. The conventional methodology trains the networks on redundant data having almost similar motion information and thus creates a computational overhead during the training process. The work presented here provides a mechanism where a properly labeled dataset is developed in which each video instance is labeled based on the type of activity present in it to make it suitable for supervised learning methodology. Furthermore, a BoF-based methodology for training networks has been introduced. The proposed methodology reduces the computational overhead for training a deep-learning network for understanding motion information in videos, without affecting the overall accuracy of the network. In addition to properly arranging the training datasets, the most effective concept of intermediate frame fusion and early frame fusion for Spatio-temporal analysis of videos has been adopted and deployed using state-of-the-art 2D, and 3D CNN. This research work has three tier contributions including:

1. Development of dataset for anomalous activity detection.
2. Development of BoF based methodology for training networks.
3. Validation of the proposed methodology for training deep-learning networks over state-of-the-art 3D-CNN, and 2D-CNN based networks developed for understanding motion information in videos.

The remainder of this paper is organized as follows: Section II reviews the related work in anomalous activity detection. Section III presents the architectures of the proposed models including problem formulation and its implementation details. Section IV explains the experimental setup including dataset while section V provides the results and analysis. Section VI concludes the paper.

II. Related Work

In this section, we will be discussing various approaches entertained in past for malicious activity detection using Spatio-temporal CNNs. Generally, malicious activities are anomalous parts of the video sequence; however, it could be observed in the literature that maliciousness of the activity in the given video sequence has been contextually specified according to the target application. According to [4], the definition of malicious activities changes with the scenario. It is a complicated task to differentiate the malicious event from the rest of the video recording at the given instance. Some of the malicious events identified in CCTV footages have been shown in Fig. 1. The following subsections present the overview of various approaches used for the detection of context-sensitive anomalies and established anomalies.

A. Context Sensitive Anomalies

Object tracking as abnormal motion in restricted areas, running, and loitering are often considered as an anomaly in a sensitive environment. Such type of movements in a video sequence are detected by various trajectory analysis techniques [8], [4]. Calderara et al. considered inter node transition pattern in a graph as trajectory for abnormal motion detection [2] while Morris et al. found the interesting node using Gaussian Mixture Model and then Hidden Markov Model for the same purpose [9]. Moreover, techniques based on low level local features have been used for detection of abnormal motion [10], [11], [6]. Ermis et al. constructed a probabilistic model for abnormal motion detection by generating behavior cluster derived from behavior profile [12]. Reddy et al. exploited ground truth segmentation in combination with the motion and size feature modeled by kernel density estimation [11]. This approach claims its effectiveness in detecting abnormal object in crowded scenes. Xiao et al. used hybrid combination of sparse semi non-generative matrix factorization (SSMF) and histogram of non-negative coefficient (HNC) for anomaly detection in surveillance videos [6]. In their work only normal data is used for parameter tuning and deviation from normal motion is considered as an anomaly. Li et al. [13] proposed a Spatio-temporal model for anomaly detection in complex and crowded scene. Dynamic texture model in combination is used for considering both dynamics and appearance information. In proposed model spatial saliency score is computed using a center-surround discriminant. Whereas, temporal saliency score is produced using a model of normal behavior learned from data. Although, all of these techniques performed well for abnormal motion detection such as running in a scene and walking in the wrong direction, however, they are specifically designed for tracking objects in image sequence.

B. Established Anomalies

Anomaly detection remained in focus for the last decade to detect the abnormal human behavior in surveillance videos. A group of researchers working in the area of ensuring safety of pedestrian walkways focused on the detection of non-pedestrian entities in public walkways [13], [19], [22]. Li et al. [13] proposed the Real Time Volume Crime Detection and Classification using Deep Learning. [14] proposed the use of Spatio-temporal information Center-surround discriminant saliency detector and normal behavior model for extracting spatial and temporal saliency score respectively, for the categorization of pedestrian abnormalities. Tahboub et al. [14] used local binary pattern in combination with random forest for detecting pedestrian anomalies. Ravan et al. [3] used generative adversarial

TABLE I. Existing Datasets for Malicious Activity Detection

Name of Dataset	Total Videos	Environmental Conditions	Identified Anomalous Activities
UCSD PED1	170	Outdoor	Non-pedestrian entities, and walking across walkways
UCSD PED2	28	Outdoor	Non-pedestrian entities, and walking across walkways
Subway Entrance	1	Indoor	Wrong direction, no payment, and loitering
Subway Exit	1	Indoor	Wrong direction, no payment, and loitering
Avenue	37	Outdoor	Running, and Throwing object
UMN	5	Outdoor, Indoor	Running
Hockey Fight	1000	Playground	Fighting
The Movies	200	Outdoor, Indoor	Fighting
UCF Crimes	1900	Outdoor, Indoor	Abuse, Accident, Arrest, Burglary Explosion, and Fighting

network for learning normal pattern of public walkways and deviation from the learned pattern is considered as an abnormality. Ameer et al. [15] proposed combination of connected component analysis, histogram of oriented gradient and Gaussian mixture model for non-pedestrian object detection. Khan et al. [7] used Gaussian discriminant model in combination with k-means clustering for classification of events recorded in surveillance videos installed in pedestrian walkways. Various applications of the anomaly detection have been previously validated with certain benchmark dataset listed in Table I. These include some of the prominent datasets of UCSD (PED1, PED2), Avenue, Subway Entrance, Subway Exit, and UMN. Most of these datasets have been developed for identification of a particular set of anomalies in specific environments e.g., UCSD is used to identify non-pedestrian entities and walking across walkways, Subway dataset is for identifying walking in wrong direction, non-payments, and loitering, while Avenue, UMN, Hockey Fight, and The Movies datasets have been used for identification of single activity mostly fighting. Moreover, these datasets have been developed with the help of actors and does not provide any real-life situation in any of their videos [16]. We can conclude that the approaches discussed in the previous section mainly targeted contextual anomalies tested over fabricated datasets in which anomalous scenes are acted by the actors. Considering the importance of detecting real-world anomalies recorded in real-time CCTV footage to assist security agencies. Sultani [16] introduced UCF crimes dataset incorporating 13 real-world anomalies and proposed multiple instance learning (MIL) for the classification of abnormal activities. The dataset is developed by downloading videos of CCTV recordings from live leaks and youtube. Each video is labeled according to the type of anomaly recorded in that video.

According to Sultani the developed dataset is weekly annotated. The proposed algorithm achieved a classification accuracy of 28%, thus, a state-of-the-art solution for detecting real-world abnormality in CCTV recording is still a dream.

C. Spatio-temporal Analysis

Spatio-temporal analysis is usually desired for identification of a function between spatial and temporal data to affect the performance of any process. While it defines a relation between GPS coordinates and its time instance for activity recognition in [17] and a relation between location and time of the day for prediction of criminal activity in [18], it links the spatial information of each frame of a video sequence to its temporal distribution in [19]. Extraction of useful information from a video sequence relies not only on the visual information spread spatially in each static frame but also on the complex motion information distributed along the continuous sequence of frames. Previously, hand-crafted features are used for obtaining appearance and motion information from video streams [20], [14]. However, these hand-crafted features contain less discriminant information and recent deep supervised and unsupervised features are employed for different applications [16]. More advanced, Spatio-temporal convolutional neural networks (CNN) [21] are introduced to learn appearance as well as motion information in video stream [19] which previously gained popularity in area of action recognition [22], [23] and hand gesture recognition [24]. Inspired from the popularity of Spatio-temporal CNN's, we in this research propose the use of a similar CNN modified for malicious activity detection.

In this research, we have developed a dataset specifically to identify malicious events in a video stream. For this purpose, videos are taken from UCF crimes dataset and are annotated for instance recognition task. Followed by a BoF based training mechanism to reduce the computational overhead. The proposed methodology is then verified by training state of the art Spatio-temporal CNN's developed for understanding motion information in videos.

III. Proposed Framework

Instance recognition in videos demands analysis of the information spread across spatial and temporal domains of video sequence. We can acquire semantic information of the scene from spatially distributed objects in a single frame, while sequence of such consecutive frames provides the positional changes of objects, hence enabling us to understand the overall activity in the video stream. To perform this task through CNNs, we have designed a framework that takes in samples of the video stream and outputs the activity performed in each sample. The overall architecture has been divided into the modular phases of video processor, feature extraction, and instance recognition. The overall architecture is presented in Fig. 2.

A. Video Pre-Processing

Videos consist of a sequence of stationary image frames. For the interpretation of useful information through convolutional neural network from these videos, all video frames are processed for understanding motion information in full length video. So, training data on full length video creates a huge computational overhead. To train CNN's for understanding motion information in videos, we have designed a framework for training CNN's over a defined set of frames i.e. BoF containing key information required for understanding motion in particular videos. The overall proposed framework is divided into modular phase of BoF extraction, block formation, block selection, and down framing.

BoF Extraction The set of frames containing the activity in full length video recording has been termed as BoF. Initially BoF has been determined in each video. The process of BoF extraction is given in Fig. 3. The figure shows that only a portion of the full-length video labeled as assault contains the actual activity. So, 128 out of 294 frames have been considered as a BoF. The same process has been repeated for all videos in the dataset and thus reduced the training data by removing the unwanted information for understanding. A comparison of total frames in dataset and the number of frames in the set of BoF's has been mentioned in Table II.

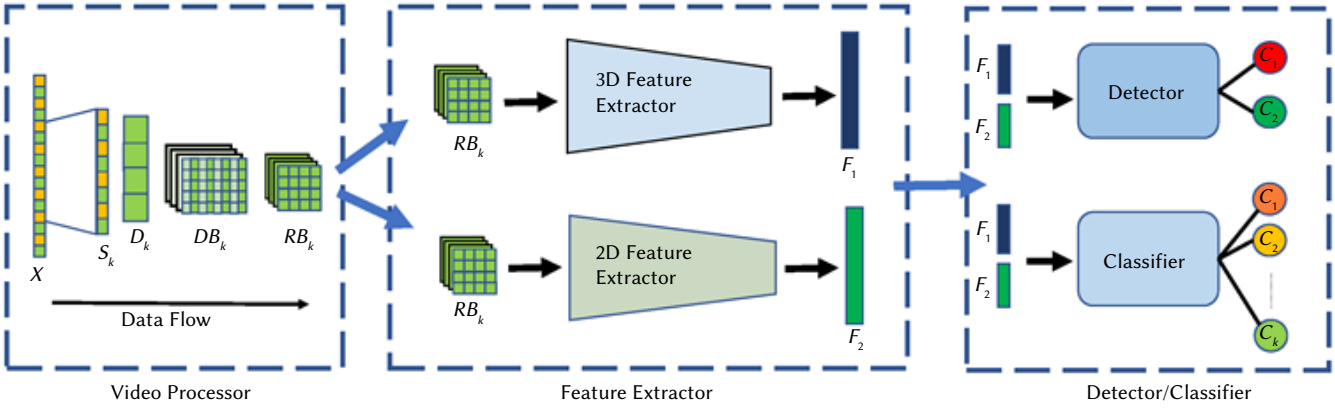


Fig. 2. Overall architectural diagram of the framework developed for malicious instance recognition.

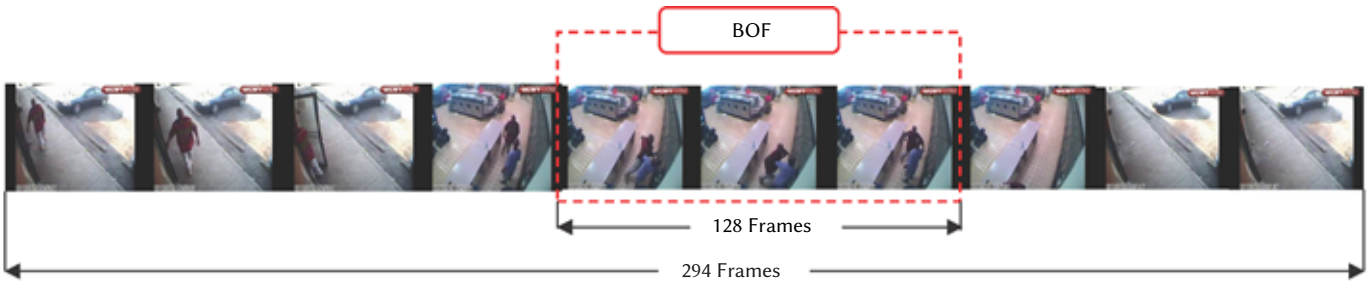


Fig. 3. BOF Extraction.

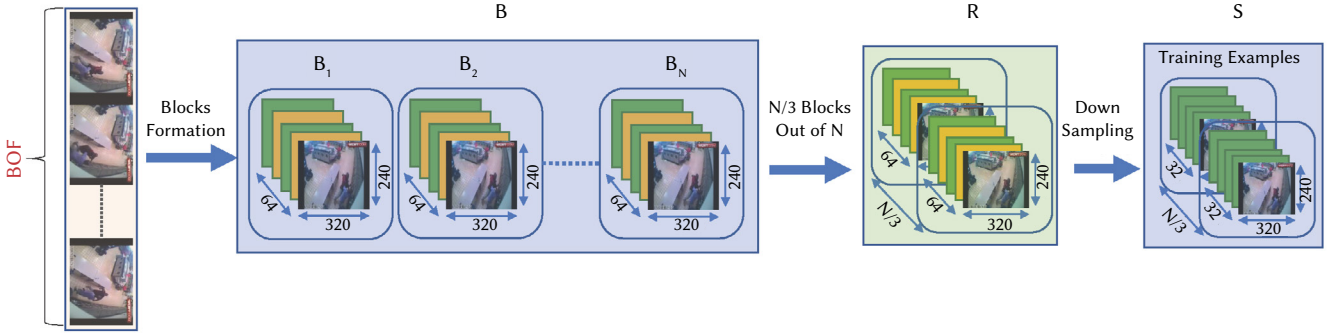


Fig. 4. The Process of block formation, block selection, and down sampling is illustrated in the figure.

TABLE II. Comparison of Frames in the Original Dataset and in the Reduced Dataset

Activity	Total Frames	Frames in BOF	Frames in Training set
Assault	129,284	54,912	9,152
Fighting	258,144	124,608	20,768
Shooting	146,624	56,768	9,472
Vandalism	146,400	84,352	14,059
Total	564,152	265,728	53,451

Block Formation The BoF is then divided into blocks of 64 frames each. The block of 64 frames covers a video length of almost 2sec. the process of block formation is mentioned in the first section of Fig. 4. The whole process of block formation is expressed mathematically in Equation (1). Considering the sequence of discrete frames from BoF expressed as X[n] we can present the m*th* block B*m* as:

$$Bm = X[n](u(n - 64m) - u(n - 64(m + 1))) \quad (1)$$

where u(n) represents the unit step function.

Block Selection and Down Sampling The set of blocks B obtained in previous section consists of total N blocks. 1/3rd of the total blocks is randomly selected for training the CNN's. Let the set of randomly selected blocks is represented by R this step is based on the fact that consecutive blocks will contain almost same motion information. Each block of set R is then down sampled to block *s* by removing alternating frame to avoid redundant information in consecutive frames. The process of down framing has been presented in Fig. 4. The overall process of block formation could be expressed mathematically by Equation (2) whereas $\delta(.)$ represents the unit impulse function. Table II shows that only 10% of the total frames have been selected for training the networks. Thus 90% of the redundant data has been avoided during training process.

$$S[n] = \sum_{k=0}^{31} R[k]\delta[n - 2k] \quad (2)$$

B. Feature Extraction

Deep CNNs have been extensively used for feature extraction in various image domains. They proved to extract much more representative features from images compare to previous hand-crafted approaches that relied mostly on local features in images. To validate

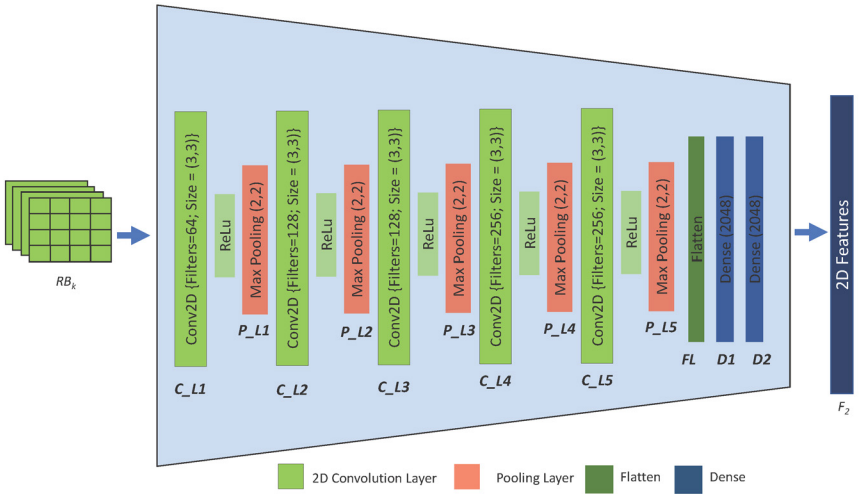


Fig. 5. 2D Convolutional Neural Network.

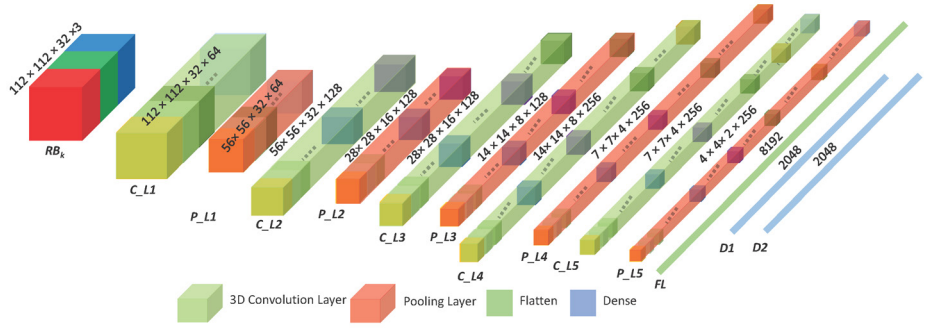


Fig. 6. 3D Convolutional Neural Network.

the concept of BoF based training mechanism, we have explored two types of deep learning architectures for extracting features from the given block of video stream. Both of these networks have been developed in such a way that take the block of video and extract a single dimensional feature.

3D CNN based Feature Extractor We believe motion information of the objects to be equally important for instance recognition in addition to the spatial distribution of objects in a given frame. For this purpose, we developed a CNN network comprising of 3D convolution layers that could learn spatial as well as temporal features from the given block of the video sequence. The proposed model is obtained by removing a few convolution layers from the standard C3D network to reduce the network complexity. Our 3D CNN feature extractor uses 5-tiered 3D convolution layers followed by 2 fully connected layers to learn a single-dimensional feature vector. Each 3D convolution layer is followed by a Max-pooling layer with stride 2×2×2 to transform the object and motion information from spatial and temporal dimension to depth. This transformation leaves us with a frame size of 4×4×2×256. Recently, 3D-CNN gained popularity in the area of action recognition [6], [23], [17]. Inspired by the performance of 3D-CNN in the field of action recognition, we developed the model in Fig. 5.

$$Y[x,y,z] = \sum_{i=0}^{L-1}\sum_{j=0}^{M-1}\sum_{k=0}^{N-1} h[i,j,k]S[x-i,y-j,z-k] \quad (3)$$

2D CNN based Feature Extractor Network parameters of proposed system are 24,109,437. In order to reduce the network parameter, the dimensions of input block are reduced from (112×1×12×3×32) to (112×112×32) by converting each frame to grayscale. Conversion to grayscale does not affect the system performance in case of activity detection due to the fact that activity detection procedure is not sensitive to the color tone in video frames. The gray-scale block is then fed to 2D CNN of same number of convolution layers and dense layers. Thus, reduces the network parameters to 13,722,437. Block diagram of the proposed system based on 2D-CNN is given in Fig. 6. In the figure, the last layer represents the feature vector. The feature vector is then fed to two fully connected layers and an output layer. The number of neurons in fully connected layers and dense layer are same as that in 3D-CNN. For instance, S is the block of 32 gray-scale frames of a video, S_k is the kth frame of block S, h is the 2D filter of dimension L _M. The mathematical process for considering temporal as well as special information is shown in Equation (4).

$$Y[x,y] = \sum_{i=0}^{L-1}\sum_{j=0}^{M-1}\sum_{k=0}^{N-1} h[i,j]S_k[x-i,y-j] \quad (4)$$

C. Instance Recognition

The objective of the research is to detect the combination of frames in a video stream with one of the categories of the volume crimes mentioned earlier on. For this purpose, features of the block acquired in section III.B are classified through various classification algorithms including Gaussian Naive Bays, Decision Tree, Support Vector Machine (SVM), k-Nearest Neighbor, and Softmax. All these classifiers are developed in such a way to address both instance detection (binary classification) and instance classification (multi-class classification). Among all these, softmax classifier has been designed with softmax activation optimizing the binary cross entropy loss and categorical cross entropy loss with detection and classification, respectively. Both of these losses are mathematically presented as:

$$L_b = -\frac{1}{N}\sum_i^N t_i\, log(p_i) + (1-t_i)\, log(1-p_i) \quad (5)$$

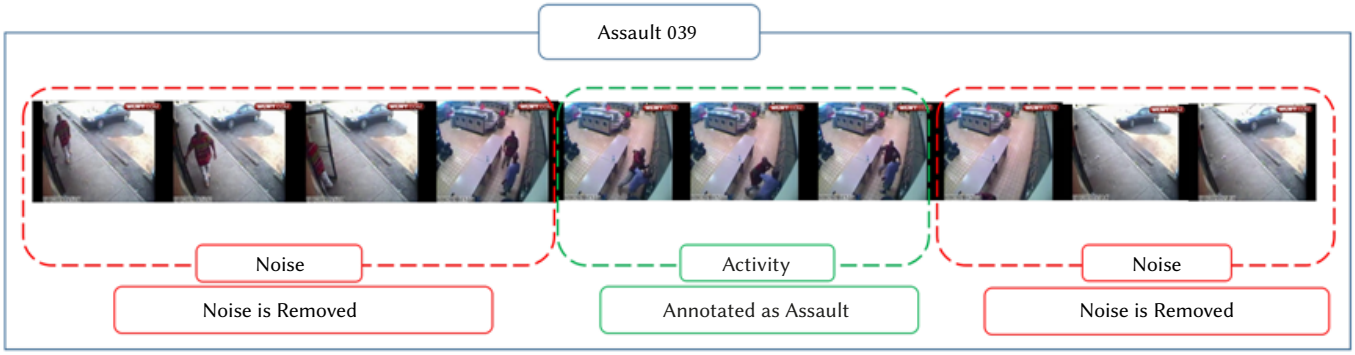


Fig. 7. Annotation of video sample labeled as Assault in original dataset.

$$L_C = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i}^{C} t_{ij}\, log(p_{ij}) \quad (6)$$

where N represent total number of training examples (blocks), t is target value, p is predicted value, and C denotes the number of classes for multi-class-classification.

IV. Experimental Setup

We have developed a unified framework for the tasks of detection and classification. For the case of instance detection, the block is identified as a normal or anomalous event. This is like the concept introduced in [28] which considered everything that doesn't look normal as anomaly. In classification, on the other hand the specific type of activity associated with each of volume crime is identified. Technically, detection performs a binary classification task (0; 1) while in classification we perform a multi-class classification task (0; 1; 2; 3; 4) with the same framework given in Fig. 2. Both tasks have been validated through the dataset developed specially for malicious instance recognition.

A. Dataset

Apart from the methodology for effective training of CNN's, this research work is also focused on the anomaly detection in safe-city environment, this is why the subset of a very recent dataset (UCF crimes) developed for real-world anomaly detection in surveillance videos has been considered. The dataset consists of CCTV footages of real-world anomalies from Liveleaks, and Youtube including 13 real-world anomalies containing 1900 videos spanned over 128 hours. For this research, a subset of four most crucial anomalies (shooting, assault, fighting, and vandalism) are annotated for in-video event recognition. This is carried out by specifically separating normal frames from the ones that contain anomalous activity. The process has been demonstrated in Fig. 7. Previously, it was very difficult to use the video labeled as assault, it was observed that videos labeled as (Assault 039) contain (43:8%) frames belonging to normal activity and the rest belonging to the assault. Each video in our dataset consists of frame-level labels for its class annotated by three skilled annotators through visually inspecting each video stream.

B. Experimentation

We have conducted experiments for detection and classification using 3D-CNN and 2D-CNN features with various classifiers explained in section III.3. Hyper parameters setting for 2D and 3D CNNs are listed in Table III. We have performed our experiments on Intel Core-i5 with 8Gb RAM and Nvidia GTX 1050Ti Graphical Processing Unit. In each experiment Stochastic Gradient Descent optimizer is used for learning weights.

TABLE III. Hyper Parameter Choice for 2D-CNN and 3D-CNN

Parameters	2D-CNN	3D-CNN
No. of Epochs	140	70
Initial Learning rate	0.001	0.001
Momentum	0.9	0.9
Kernel Size	(3,3)	(3,3,3)
Pooling window	(2,2)	(2,2,2)

V. Results and Analysis

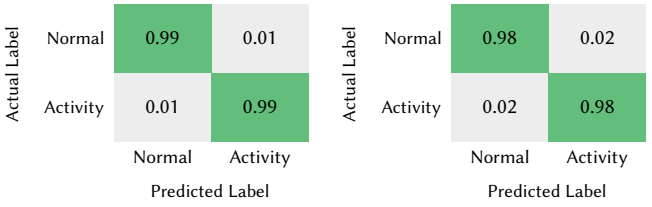
We have conducted numerous experiments for detection and classification using features extracted from 2D and 3D-CNNs. The performance of the mentioned model is evaluated based on the performance metric like AUC, accuracy, and false-positive rate. This section provides a performance comparison of our proposed system. Table IV summarizes the results.

A. Anomalous Event Detection

From the results on event detection using different features, it was observed that 2D CNN outperforms 3D CNN achieving the overall detection accuracy of 99:0%. Although, all the classifiers equally performed well for both type of features, however, Softmax classifier outperforms the rest in detecting malicious video events as shown in Table IV. Similar patterns have been observed in confusion matrices and t-SNE plots of the detection process as shown in Fig. 8 and 11 respectively.

TABLE IV. Comparison of the Features From 2D CNN and 3D CNN With Other Classifiers

Classifier	2D CNN Detection	3D CNN Detection	2D CNN Classification	3D CNN Classification
Gaussian Naïve Bays	98.80	98.67	88.83	88.72
Decision Tree	98.88	98.65	74.41	74.41
Support Vector Machine	97.80	98.57	74.41	74.41
K Nearest Neighbor	98.86	**98.69**	82.15	81.80
Softmax	**99.00**	98.10	**98.80**	**97.80**



(a) 2D-CNN (b) 3D-CNN

Fig. 8. Confusion Matrix for 2D features and 3D features using Softmax classifier.

B. Anomalous Event Classification

For classification of the event among one of assault, fighting, shooting, vandalism, and normal, the features from 2D and 3D CNN have been extracted in similar manner and evaluated with various classifiers. It is observed that the performance of Softmax classifier combined with 2D features is much better in comparison to the rest of classifiers. Overall classification accuracy of 98:89% has been achieved for this specific task. Even though, 3D features also performed well in classification; however, the number of network parameters in 2D CNN are much less as compared to 3D-CNN.

C. Extended Analysis on Anomalous Event Classification

We also presented the results in terms of detail performance metrics including Precision, Recall, and F1-Score for each class of anomalous events. Mathematical expressions for Precision, Recall, and F1-Score are given in equations (7-9), respectively. Upon observation of table VI, we concluded that 2D-CNN performs better for each class in comparison to the 3D-CNN. A similar phenomenon could be observed in the recall score for the shooting as 0:846 and 0:916 for 3D-CNN and 2D-CNN, respectively. Fig. 10 and 9 show the confusion matrices and t-SNE plots for event classification using 2D and 3D features with Softmax classifier. It should be noted that visibly separable clusters could be seen in t-SNE plots which validates the accuracy achieved for the given task of classification.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{7}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{8}$$

$$F1 = \frac{2 \times Precission \times Recall}{Precision + Recall} \tag{9}$$

D. Comparison With the State-of-the-art

We have also compared our approaches with the state-of-the-art techniques using 14 UCF crimes dataset. It is observed from Table V

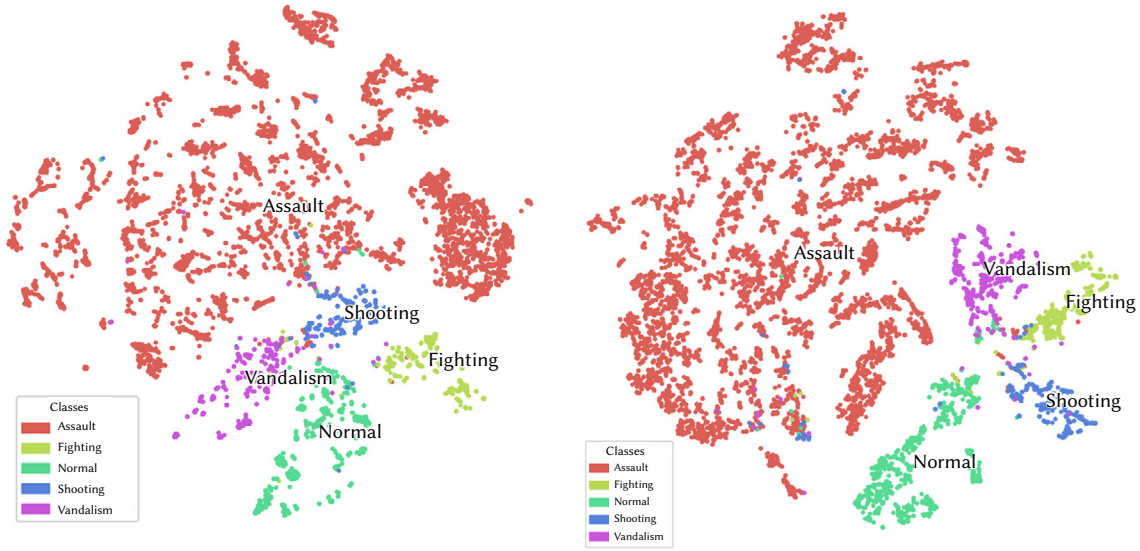
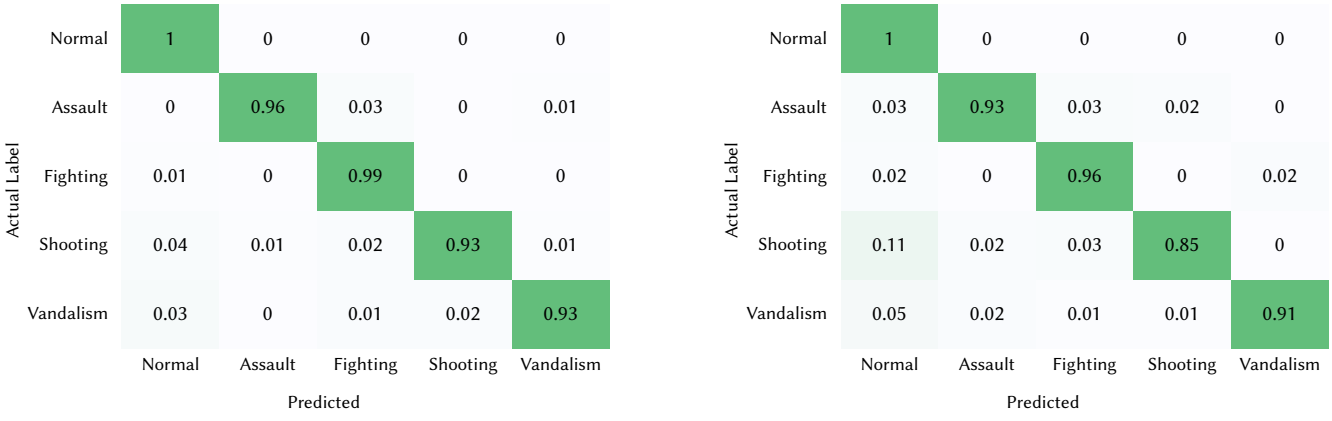


Fig. 9. t-SNE plots for 2D features and 3D features using Softmax classifier.



(a) 2D-CNN (b) 3D-CNN

Fig. 10. Confusion Matrix for Classification Task for 2D features and 3D features using Softmax classifier.

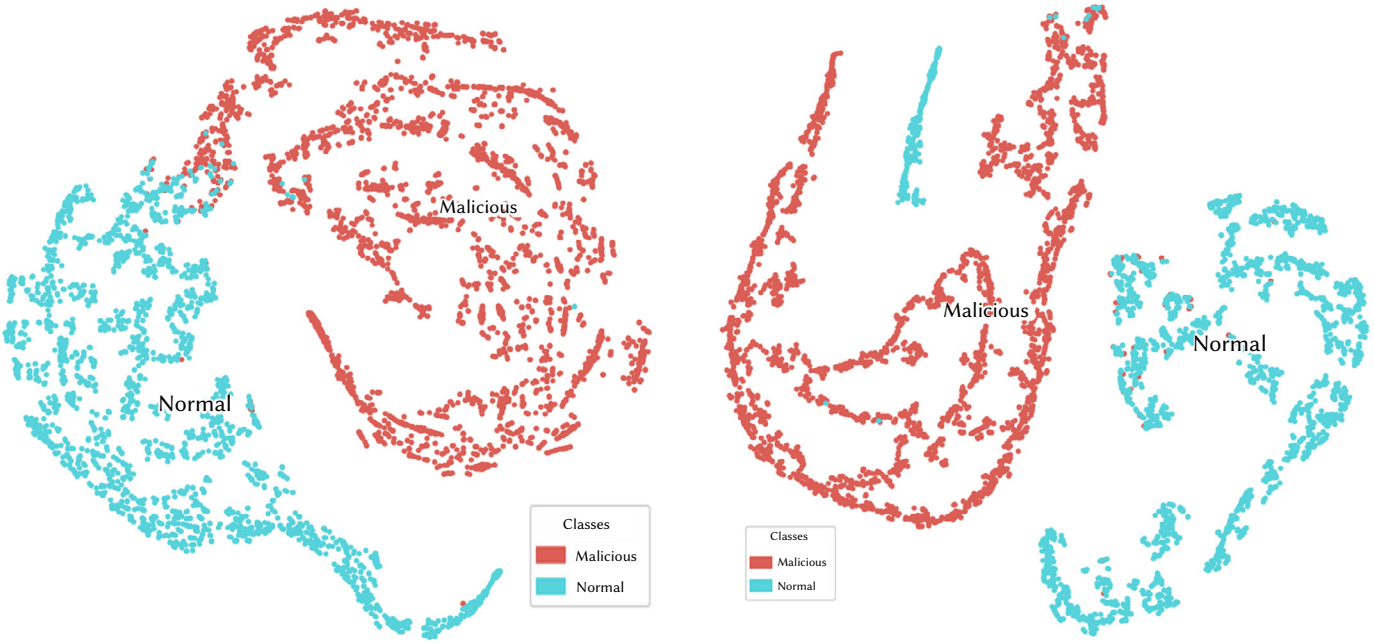


Fig. 11. t-SNE plots for Classification Task for 2D features and 3D features using Softmax classifier.

TABLE V. Comparison of Our Approach to the State-of-the-art Techniques Using the Dataset Proposed By [21]

Authors	Technique	Accuracy	AUC	False Positive
Sultani *et al.* [16]	MIL	28.4	75.41	1.9
Khan *et al* [7]	GDA	-	64.30	-
Ours	2D CNN	**98.8**	**99.7**	1.2
Ours	3D CNN	97.4	99.5	**0.7**

TABLE VI. Extended Analysis on Anomalous Event Classification

Technique	Activity	Precision	Recall	F1 Score
ResNet3D [29]	Normal	0.997	0.990	0.993
	Assault	0.993	0.944	0.968
	Fighting	0.955	0.998	0.976
	Shooting	0.941	0.971	0.955
	Vandalism	0.994	0.96	0.976
(2+1) D [26]	Normal	0.996	0.995	0.995
	Assault	0.996	0.944	0.970
	Fighting	0.984	0.994	0.989
	Shooting	0.911	0.988	0.948
	Vandalism	0.994	0.965	0.979
P3D [30]	Normal	0.993	0.996	0.995
	Assault	0.967	0.965	0.966
	Fighting	0.994	0.986	0.990
	Shooting	0.953	0.956	0.955
	Vandalism	0.990	0.986	0.988
3D-CNN	Normal	0.985	0.996	0.991
	Assault	0.933	0.929	0.931
	Fighting	0.968	0.963	0.965
	Shooting	0.962	0.846	0.900
	Vandalism	0.966	0.912	0.938
2D-CNN	Normal	0.993	0.997	0.995
	Assault	0.986	0.965	0.975
	Fighting	0.985	0.990	0.988
	Shooting	0.955	0.916	0.935
	Vandalism	0.952	0.943	0.948

that our approach is performing far better on 5 classes as compared to the rest of the techniques using MIL and GDA techniques. Moreover, our approach achieves the least false positive rate of 0:7% for our 3D CNN. It is observed that the proposed 2D model outperforms the performance of 3D model in case of the activity detection and manage to achieve false positive rate of 0.1. Furthermore, the model is suitable for real time application due to its low false positive rate and high frame processing rate that is 1000 frames/sec. Also, the proposed methodology for training CNN's is further verified by training three (ResNet3D [29], (2+1)D [26], and P3D [30]) well known algorithms for video understanding. Validation results of the trained model on the data samples removed from the dataset following the proposed methodology discussed in section III.A are mentioned in Table VI. It has been observed that training the state-of-state-of-the-art algorithm by following the methodology mentioned in section III.A performed in its order of popularity. Hence, it is verified that even training on 10% of video frames are enough for understanding motion information in video, and thus reduces the computation overhead during training process by 90% irrespective of the network used.

VI. Conclusion and Future Work

Lack of implementable software solutions for the identification of real-world malicious activities from video streams in a safe city environment requires a blend of computer vision and machine learning algorithms. In this regard, a training mechanism has been introduced to reduce the computation required for training the learning algorithms. The proposed training methodology achieved a promising accuracy even reducing the computational overhead by 90%. Also, an optimal solution for the analysis of temporal frames extracted from CCTV recordings is proposed. Our proposed models managed to achieve high accuracy for not only the identification of malicious events but also classification of real-world volume crimes including assault, fighting, shooting, and vandalism in a video sequence. Furthermore, our models are also suitable for real-time applications due to their high frame processing rate and low false alarm rate, with high classification accuracy of 98.7% and AUC of 99.7% on four classes.

The system can further be modified for other classes of crimes including but not limited to burglary, riots, attempted murder, arson, explosion, robbery, theft, and arrest etc. In order to get a unified framework for the detection of multiple malicious activities recorded by a CCTV camera, we need to train the same system with the data for the above-mentioned events.

Acknowledgment

Indeed, it was a tough journey to compile all this work. At many instances, I felt that I would not be able to do this. But thanks to my team at National Center of Artificial Intelligence who helped, supported, and motivated me on every step.

References

[1] T. Ainsworth, "Buyer beware," *Security Oz*, vol. 19, pp. 18–26, 2002.

[2] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, "Detecting anomalies in people's trajectories using spectral graph analysis," *Computer Vision and Image Understanding.*, vol. 115, no. 8, pp. 1099–1111, 2011.

[3] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in 2017 *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1577–1581.

[4] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2054–2060.

[5] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, 2014.

[6] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, 2015.

[7] M. U. K. Khan, H.-S. Park, and C.-M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541–556, 2018.

[8] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[9] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011.

[10] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International journal of computer vision*, vol. 74, no. 1, pp. 17–31, 2007.

[11] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *CVPR 2011 WORKSHOPS*, 2011, pp. 55–61.

[12] E. B. Ermis, V. Saligrama, P.-M. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," in *2008 15th IEEE International Conference on Image Processing*, 2008, pp. 769–772.

[13] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[14] K. Tahboub, A. R. Reibman, and E. J. Delp, "Accuracy prediction for pedestrian detection," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4192–4196.

[15] S. Amraee, A. Vafaei, K. Jamshidi, and P. Adibi, "Anomaly detection and localization in crowded scenes using connected component analysis," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14767–14782, 2018.

[16] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[17] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu, "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4006–4015.

[18] L. Duan, T. Hu, E. Cheng, J. Zhu, and C. Gao, "Deep convolutional neural networks for spatiotemporal crime prediction," in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, 2017, pp. 61–67.

[19] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial--temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.

[20] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE transactions on image processing*, vol. 24, no. 7, pp. 2153–2166, 2015.

[21] M. Khari, A. K. Garg, R. Gonzalez-Crespo, and E. Verdú, "Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, p. 22, 2019.

[22] T. Lima, B. Fernandes, and P. Barros, "Human action recognition with 3D convolutional neural network," in *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2017, pp. 1–6.

[23] J. D.Pujari, R. Yakkundimath, and A. S. Byadgi, "SVM and ANN Based Classification of Plant Diseases Using Feature Reduction Technique," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 7, p. 6, 2016.

[24] N. L. Hakim *et al.*, "Dynamic Hand Gesture Recognition Using 3DCNN and LSTM with FSM Context-Aware Model," *Sensors*, vol. 19, no. 24, p. 5429, 2019.

[25] F. Cronje, "Human action recognition with 3D convolutional neural

networks," University of Cape Town, 2015.
[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450-6459
[27] Z. Liu, C. Zhang, and Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016.
[28] D. Gong *et al.*, "Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection," *arXiv Prepr. arXiv1904.02639*, 2019.
[29] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-Temporal features with 3D residual networks for action recognition," *Proc. - 2017 IEEE International Conference on Computer Vision Workshops. ICCVW 2017*, vol. 2018-January, pp. 3154–3160, 2017.
[30] J. Chen, J. Hsiao, and C. M. Ho, "Residual Frames with Efficient Pseudo-3D CNN for Human Action Recognition," *arXiv preprint* pp. 5–9, 2020.

Atif Jan

Atif Jan obtained his B.Sc. degree in Electrical Engineering from University of Engineering and Technology (UET), Peshawar in 2011 and his master's degree in Electrical Engineering from UET, Peshawar in 2015. He is also pursuing his Ph.D. Currently; he is working as a Lecturer at Department of Electrical Engineering, UET Peshawar. His research interests include image processing, computer vision, machine learning and deep learning.

Dr. Gul Muhammad Khan

Dr. Gul Muhammad Khan is blessed with the capability to do things and face challenges. He graduated from UET Peshawar with distinction. He completed his PhD in intelligent System design from University of York, UK. During his PhD, he devised a brain inspired "learning to learn "system that has the capability of learning for itself at run-time. He is the pioneer of Cartesian Genetic Programming evolved Developmental network (CGPDN) providing an indirect method of decoding for neural programs. He is amongst the top neuro-developmental scientists around the world. He has also introduced a new concept for neuro-evolution and presented new algorithms for Automatic generation of feedforward (CGPANN)/ feedback (CGPRNN), markovian and non-markovian non-linear systems. He has introduced the concept of plastic networks and presented algorithms for both feed forward and feedback system, termed as Plastic CGPANN and Plastic RCGPANN. He joined UET as Assistant Professor on Tenure Track System (TTS) in Aug 2008 after completion of his PhD. In March 2012, he established a Research facility, Centre for Intelligent Systems and Networks Research (CISNR) at UET Peshawar and was appointed as Director of the Centre, still serving. In a short period of three years, he obtained research funding of more than 50 million, completed five research projects, with three in progress, supervising over 25 PhD/MPhil students and Researchers. He won the Research Productivity award for the year 2017. He is the sole author of a book published by springer titled: Evolution of Artificial Neural Development.