Multimedia Tools and Applications (2024) 83:73363–73390
https://doi.org/10.1007/s11042-023-17770-z

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

A deep learning-assisted visual attention mechanism for anomaly detection in videos

Muhammad Shoaib[1,2] · Babar Shah[3] · Tariq Hussain[4,5] · Bailin Yang[4,5] · Asad Ullah[2] · Jahangir Khan[2] · Farman Ali[6]

Received: 11 April 2023 / Revised: 25 October 2023 / Accepted: 28 November 2023 /
Published online: 5 December 2023

Abstract
Ensuring public safety in urban areas is a crucial element in maintaining a good quality of life. The successful deployment of video surveillance systems depends heavily on the acquisition and processing of large volumes of urban data to derive meaningful insights. Manual monitoring and analysis of anomalous activities in the surveillance footage is both a time-consuming and error-prone process that is not scalable for urban environments with high levels of foot and vehicular traffic. Moreover, traditional surveillance systems are limited by their inability to process real-time data at scale, which can result in missed or delayed detection of potential security threats. This paper tackles this problem by proposing an automatic anomaly detection method via an attention mechanism. The attention area is identified using the background subtraction (BG) algorithm which identifies motion regions in the video frames. This information is then passed through a 3D convolutional neural network (3D CNN) to classify the normal and anomalous events. To evaluate the proposed method, experiments and analysis were conducted using the publicly available UCF crime dataset, demonstrating its effectiveness with an accuracy of 96.89% compared to the state-of-the-art methods. In case an anomaly is detected, an alert is sent to the nearest authorities to take immediate action to prevent further harm or damage.

Keywords Deep learning · Transfer learning · Attention mechanism · Background subtraction · Violence detection

1 Introduction

As the need for safety and security increases, monitoring systems are being installed in public facilities and places to evaluate existing aggression. By utilizing surveillance cameras, unusual occurrences like traffic accidents, robberies, and other criminal actions can be identified before they happen [1]. However, most existing surveillance systems require manual operation and inspection, making them vulnerable to interference and causing fatigue to the operator. Thus, intelligent computer vision algorithms

Muhammad Shoaib and Farman Ali these authors contributed equally to this work and co-first authors.

Extended author information available on the last page of the article for automated anomaly detection and video violence detection have become increasingly critical. To tackle anomaly detection, various algorithms have been developed to detect specific types of anomalies, such as violence, battles, and road traffic accidents [2, 3]. These algorithms rely on a combination of machine learning, computer vision, and deep learning techniques to extract relevant features from the video data and identify abnormal events. Although these algorithms signify a substantial advancement in anomaly detection, their capacity to discern intricate and contextually dependent anomalies remains restricted. Additionally, their accuracy can be compromised by variables including lighting conditions, camera positioning, and occlusions. Consequently, current research endeavors in this domain are focused on the creation of more advanced algorithms designed to overcome these constraints and deliver heightened precision and dependability in the detection of anomalies within video surveillance data.

A simple and effective microscopic traffic model called the optimal velocity model (OVM) is proposed to capture many properties of real traffic flows [4, 5]. This model has garnered interest in recent years due to the robustness of convolutional neural networks (CNNs) for video action recognition. A large dataset and multiple instance learning (MIL) based solution is presented to address the challenge of monitoring surveillance camera footage with limited human resources [6–8]. Meanwhile, a local region-of-interest (RoI) detection is used to analyze video frames [9, 10]. The authors proposed a model for identifying intrinsic anomaly regions and determining the need for spatial and temporal information to aid in anomaly detection [11]. However, these approaches rely on manual annotation, which can be time-consuming and inefficient. In contrast, a different approach is employed by extracting optimal frames based on motion data and using the activation map to automatically identify potential attention areas where combat actions might occur [10]. They then analyzed the ratio of violent and normal activity at each location and used the spatial relationships between human offers and activation boxes to identify all relevant local requests within the extracted region of attention. Although current algorithms for anomaly detection in video surveillance data have demonstrated efficacy in identifying specific events like fights, violence, and traffic accidents, it is imperative to recognize that these methods may not be suitable for a comprehensive anomaly detection system. This limitation arises because anomalies can manifest in diverse forms, necessitating a more intricate and contextually attuned approach. Managing unlabeled public video clips presents inherent complexities, including challenges related to masking, light fluctuations, blurriness, and other alterations. Moreover, violence detection in video data introduces additional complexities such as viewpoint variations, cluttered backgrounds, and the absence of a global representation for violent actions due to inconsistencies in individual performance, as well as disparities in lighting and scale.

To address these challenges, this paper proposes a method for automatic and effective attention zone positioning based on background subtraction. The method involves using a robust background subtraction technique to locate the attention zone or movement, which is then fed into a 3D CNN motion recognition model. Notably, the proposed system only uses the focus regions identified in each frame. Detection is treated as a prediction problem, with a visual encoder and a highly flexible prediction structure used to predict violent actions. The following are the contribution of this research work:

- Proposes a novel approach leveraging deep learning and an attention mechanism to automate surveillance systems for the detection and recognition of violent subtypes in video data.

- Presents a comprehensive framework capable of detecting and recognizing various types of violent subtypes in video data, which can enhance public safety and security in smart cities.
- Develops a texture-based bilateral filter that uses the foreground and background region in a video frame to detect the attention region, which is fed into the C3D model for violence prediction.
- Uses a C3D model to extract learnable features from the video frame and a genetic algorithm-based feature selection method to reduce the size of 3D CNN features by selecting optimal feature attributes and discarding low-ranked ones.
- Trains a CNN softmax classifier over the GA-based CNN features to classify videos as violent or normal. If a frame is classified as violent, the system further recognizes the violence subtype and estimates the violence score based on the severity of the violent scene.

The paper's structure can be outlined as follows. Section 2 provides an overview of the relevant work that the paper builds upon. In Section 3, the proposed methods are presented in detail, which includes the process of extracting attention areas from spatiotemporal data and detecting local anomalies. Section 4 evaluates the strategy by analyzing the results. Section 5 concludes with observations and conclusions.

2 Related works

Detecting abnormalities remains the most persistent and challenging problem in computer vision, as referenced in several sources [12–17]. Public areas such as runways, courtyards, metro stations, and transportation hubs have installed numerous cameras in response to the growing need for public safety and monitoring. However, the vast amount of video data generated by these cameras makes it difficult and time-consuming for operators to search for unusual or suspicious events. To increase productivity while preserving efficiency, automated equipment is necessary. Therefore, significant advancements have been made in the field of intelligent filmed investigation, and various techniques have been proposed to enhance video anomaly detection [18, 19].

Recently, numerous methods for noticing irregular behavior have been advanced. To identify violence, use video and audio data from a filmed investigation [20, 21]. To distinguish between violent and peaceful videos, Mohammadi et al. developed a categorization method. A new heuristic approach based on behavior is proposed. In contrast to previous research, the authors recommend using tracking as an exception to simulate normal motion. Many systems employ graph methods [22, 23], interpersonal force models, hybrids of impact assessment models, hidden Markov models (HMMs), thematic modeling, mobility modes, and situational techniques to avoid tracking and interpreting global motion patterns. Melan and his colleagues, the difference between the expected and actual velocity obtained using particle starvation, and a social force model is used to calculate the scene interaction force [24, 25].

Numerous methods have recently been developed to detect irregular behavior. One method involves using video and audio data from a recorded investigation to identify violent behavior [20]. Mohammadi et al. developed a categorization method to distinguish between violent and peaceful videos [21]. Another approach proposed a heuristic method based on behavior, which recommends using tracking as an exception to simulate normal motion, unlike previous research. Many systems use various methods such as graph methods, interpersonal force models, hybrids of impact assessment models, hidden Markov models, thematic modeling, mobility modes, and situational techniques to avoid tracking and interpret global motion patterns [22, 23]. Melan and his colleagues used a difference algorithm to calculate the scene interaction force, which identifies low-likelihood patterns as anomalies based on a training film for specific behavior [24, 25]. A deep learning-based approach has also been proposed, where handcrafted features are not used to learn the video features. Xu et al. use machine learning infrastructure and singular SVM models to score each input's exception level, and the results are combined for final anomaly detection [10]. Hassan and his colleagues use the Conv-AE framework for scene reconstruction to calculate the reconstruction cost of anomaly detection [26]. Another approach, proposed by Ionescu and colleagues, uses deep neural networks and multiple learning instances to classify real-world anomalies like accidents, explosions, battles, abuse, and arson [27]. Like other approaches, this method considers both normal and abnormal behavior when detecting anomalies. Finally, the research effort introduces the concept of graphic courtesy to aid in identifying areas of interest [28, 29].

A framework for online anomaly detection is proposed in this article. In the work presented here, the support area is managed in response to the dynamic changes in the scene [30]. This method identifies features that enhance online performance by concentrating on features within a restricted processing area. This assists in narrowing the search for features that enhance online performance. Encoding motion dynamics necessitate using low-level characteristics, such as information about the optical flow. In the final step, bags of words and Gaussian mixed models are employed to identify anomalous events. Local Spatio-temporal features provide an abstraction of a scene's behavior by directly analyzing shape, trajectory, and size information at the object level. This allows the features to provide an abstraction of behavior. Trajectory analysis as a technique for describing video anomalies is an effective method. To identify visual anomalies, the trajectory-based method compares the degree of similarity or distance between clusters that have been generated. Another study proposed a semi-supervised learning technique for detecting violent behavior [31].

This method simultaneously trains a singular dictionary and a linear classifier. Combining the reconstruction loss and representation constraints of expensive labelled and inexpensive unlabeled data defines the objective function of dictionary learning. This is done to increase the dictionary's discriminatory power. To circumvent the constraints imposed by k-recent classification, the authors propose using a group of prototype objects as representations to employ weighted combinations of different types of similarity [32]. The authors of the research paper propose a multi-dictionary-based method for hyperspectral anomaly detection to circumvent the current limitations of hyperspectral anomaly detection. These limitations include dealing with large spectral dimensions and obtaining spectral correlations with difficulty [33].

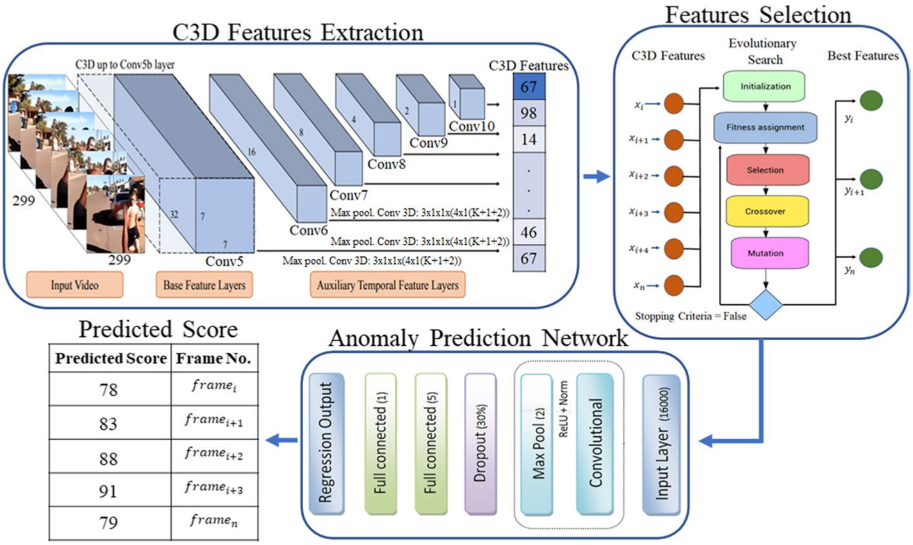
By training sparse representations based on multiple dictionaries and applying this training to different background scenarios for remote sensing, it is possible to acquire discriminative features for anomaly detection. The authors used a technique called maximum pooling in conjunction with sparse encoding to extract the distinguishing characteristics from the video [34]. A new Motion Weber local descriptor was proposed as a possible solution for identifying abnormal motion in video sequences [35]. Low-level appearance-based features and kinematic dynamics-defining components are added to train the Weber Local Descriptor (WLD). Consequently, identifying violent video content using manual feature extraction techniques is no longer problematic. In addition, these results demonstrated that the WLD descriptor accurately captured motion near the camera. In recent years, long-term, short-term memory, abbreviated as LSTM, has been used to solve various issues in speech recognition, natural language processing, and motion recognition. LSTM was created to solve the gradient disappearance or explosion problem, which had previously plagued the deep learning research community. In the past, the entire deep learning research community was plagued by this crucial issue. The research paper's authors [36] present multiple models of autoencoders in which local spatiotemporal and depth features were investigated. The models were investigated. The first autoencoder acquires knowledge using conventional spatiotemporal features, while the second autoencoder acquires knowledge end-to-end via a convolutional feed-forward architecture. In the research [37], convolutional neural networks (CNNs) are trained to use semantic information to detect suspicious video events. This information is extracted directly from the videos. In the interim, the authors of the paper [38] could achieve the same outcome by employing a network that has been pre-trained on the ILSVRC benchmark dataset. The proposed sparse coding method employs recurrent neural networks to optimize the parameters and enhance the ability to predict anomalous events [39]. The research paper's authors used generative adversarial networks to reconstruct appearance and motion representations to identify anomalous video events [40]. This technique uses the optical flow map of normal frames to model the network. Eventually, deviations from the normal model are determined using measurements of local differences. A brute-force strategy for detecting video activity is proposed as a bidirectional C-LSTM architecture that takes frame difference as its input [41]. They encoded spatial information with the VGG16 architecture of convolutional neural networks (CNNs) [42], and they derived and encoded temporal dynamics with bidirectional convolutional LSTM. The authors propose a new technique for bidirectional temporal coding and maximal feature-by-feature pooling as an alternative to the current violence detection models based on spatial–temporal coding. This strategy employs data enhancement because, as stated previously, depth models require a substantial quantity of data. The author prioritized applying simpler models over representations based on deep learning in this work. These representations require voluminous training data to learn differentiated and compact event integrations.

3 Methodology

This section presents the methodology of the proposed framework for automating the surveillance system using deep learning and an attention mechanism. Figure 1 illustrates the entire architecture of the proposed approach. The proposed framework is composed of four main components: 3D convolution network (C3D)-based feature extraction, genetic algorithm-based feature selection, anomaly prediction, and visual attention detection.

3.1 Detecting and implementing violent event actions

This section provides a detailed explanation of the proposed scheme for detecting violent event actions, encompassing three key components: 3D CNN-based feature extraction from the video data, GA-based optimal learnable features, the optimization of GA hyperparameters, and evaluation criteria, along with the anomaly prediction score estimation scheme.



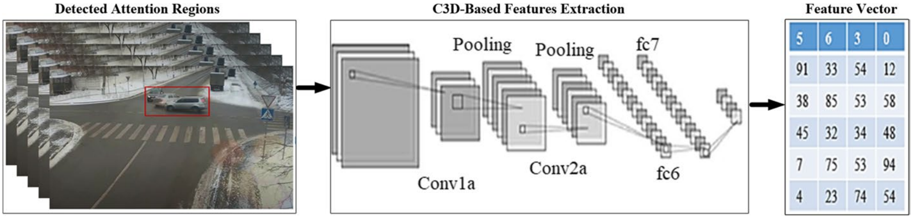
Predicted Score	Frame No.
78	$frame_i$
83	$frame_{i+1}$
88	$frame_{i+2}$
91	$frame_{i+3}$
79	$frame_n$

Fig. 1 The proposed framework for anomaly violence detection using C3D features and CNN regression model

3.1.1 Features extraction mechanism

3D Convolutional Neural Networks (CNNs) are widely used in computer vision for tasks such as classification, detection, and identification [43]. The 3D CNN model comprises layers such as aggregation, fully connected, and complicacy. In its natural state, each layer is connected to the previous layer via a core of a fixed size [44, 45]. This algorithm is based on the idea that a moving object is always associated with a cluster of pixels in the feature space. The model architecture uses various types of layers and stimulation purposes to achieve improved characterization structures compared to ergonomic engineering programmable software [46, 47]. Figure 2 depicts the overall proposed framework for features extraction. In this Figure, we have chosen the popular 3D convolution network (C3D) as the pre-trained feature extractor due to its high performance and efficiency.

Recent studies have shown that pre-trained Sports-1 M dataset models produce excellent classification and detection performance. During the training process, CNN 3D receives video clips as a general representation and fine-tunes the model to



5	6	3	0
91	33	54	12
38	85	53	58
45	32	34	48
7	75	53	94
4	23	74	54

Fig. 2 2 C3D-based features extraction and feature vector generation from input frames

display the unique characteristics of the video clip while still being able to broadcast the exclusive film clip. This learning strategy employs mixed sampling and cross-validation implicitly [48].

3.1.2 Genetic algorithm-based features selection

The genetic algorithm is an advanced optimization technique that has many applications in data mining and computer vision research. This model is an evolutionary search approach that imitates selection, mutation, and crossover in natural settings [49]. The genetic algorithm is a metaheuristic feature selection method that starts the search and finds several solutions to the problem. It selects the best answer from a list of responses to the problem and reduces the size of the feature vector by selecting the best features, making the training and validation process more efficient.

In the context of violence detection and subtype classification, the features for each sample (video frame) are extracted using the C3D method and represented as a vector with a dimension of 1×1000. The GA method takes the features and labels of video frames as input and returns a list of the best feature attributes. The reduced features vector consists of optimal features that not only improve the classification but also make the training and validation process more efficient by reducing the size.

In the first stage of the presented procedure, a population that is both random and uniform is produced, and the chances of mutation and crossover during each generation are set at 0.3. The model starts with an initial population that is entirely random, and each chromosome includes several gene characteristics, with each gene assigned a specific number. The chromosomes are represented graphically using the following equation:

$$L = \left\{ \left[C_i | C_i \notin (1, 0) \right] \right\}, n_i = 1 \tag{1}$$

where, L, C_i, and n_i represent the set of selected features, specific feature characteristics, and the total number of features extracted using a CNN, with a dimension of 1×1000 features, respectively. This notation clearly indicates that L represents a set of selected features where C_i does not belong to the closed interval [0, 1].

3.1.3 Hyperparameters of the genetic algorithm

In our proposed method, the Genetic Algorithm plays a crucial role in feature selection to enhance violence detection and subtype classification. We acknowledge the significance of disclosing hyperparameters for reproducibility. The key hyperparameters for our GA are as follows:

1. The GA begins with an initial population that is both random and uniform. This population is generated with a predefined size, which we set to [Specify the population size, e.g., 100 individuals] in our experiments.
2. During each generation, the GA employs crossover and mutation operations to explore the search space effectively. In our experiments, we set the probabilities of mutation and crossover to [Specify mutation rate, e.g., 0.3] and [Specify crossover rate, e.g., 0.3], respectively.
3. Each chromosome in our GA represents a potential solution (i.e., a feature subset) to the feature selection problem. We represent chromosomes as sequences of genes, where

each gene corresponds to a feature. The equation used for chromosome representation is as follows:

L = $[C_i | C_i \in \{0, 1\}]$ for i = 1 to 1000. Here, C_i takes a binary value (0 or 1) to indicate whether the corresponding feature is included or excluded in the feature subset.

3.1.4 Evaluation criteria

To assess the effectiveness of our GA-based feature selection method, we employ the following evaluation criteria:

1. The fitness function in our GA quantifies the quality of each feature subset. We calculate this fitness using a classification accuracy measure. The higher the accuracy, the better the feature subset is considered.
2. We determine the convergence of the GA by monitoring the change in the fitness of the best solution over generations. Convergence is achieved when the fitness values stabilize or exhibit minimal improvement over a certain number of generations.

3.1.5 Implementation methods of anomaly prediction

The proposed flowchart for visual attention-based anomalous prediction/detection is shown in Fig. 3. The visual features are extracted using the full connection FC6 layer of the C3D system. Prior to feature computation, each video frame is resized to 239,318 pixels, and the frame rate is adjusted to 29 frames/second. The C3D is computed for every 16-image frame video clip, and l2 normalization is applied before averaging the features of the 16 images to obtain the video clip features [47]. A three-level FC neural network is used to process these 4096-dimensional features. All 160 frames (i.e., all the C3D extracted files) are used to determine the existence of an exceptional scenario. The regression network produces scores for anomalous events, which can range from 0 to 1 or be binary and represent the probability of an anomalous event occurring within the limited fragment under investigation. Following the approach in [6], we use the MIL rank loss as a sparse and smooth constraint and treat each video clip as an instance of a package. The abnormal score sc(N) of the video must satisfy the following criteria. The C3D system's full connection FC6 film is used to export visual features. Before computing the characteristics, we enlarge each video frame to 239,318 pixels and adjust the frame frequency to 29 frames/second. Before applying the l2 normalization, for every video clip of the 16-image frame, we calculate the C3D. We average the 16 image clip features in a video clip to get its features [47].

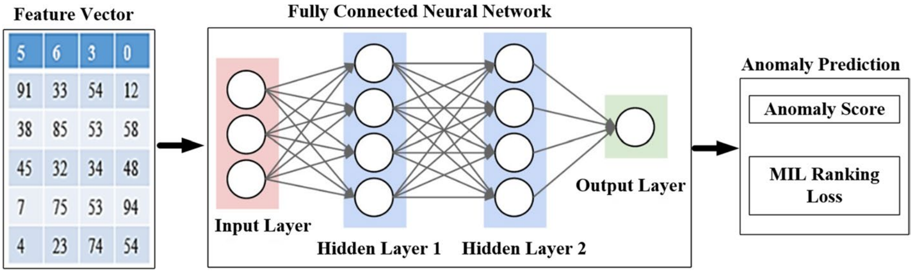


Fig. 3 Anomaly prediction based on full connection FC6 layer of the C3D system

A three-level FC neural network is fed these features (4096D), which can be seen in Fig. 3. We gradually deduce that all 160 frames (all C3D extracted files) are used to persuade the exception scenario of its existence. The regression network generates video exception scores. Scores ranging from 0 to 1 or binary can be interpreted as the likelihood of an anomalous occasion taking place inside the limited fragment under investigation. Inspired by the work [6], we use MIL rank loss as a sparse and smooth constraint and treat each video clip as an instance of a package. The following requirements must be met when capturing the abnormal sc(*N*) of video, as shown below.

$$\begin{cases} 0 \leq sc(N) < T, \text{ if } N \text{ is irregular}; f \quad (x) \\ T \leq sc(N) \leq 1, \text{ if } N \text{ is regular}; f \quad (x) \end{cases} \tag{2}$$

Equation 2 sets the criteria for determining the abnormality sc(*N*), based on the nature of the video clip *N*, which can either be "irregular" or "regular." The sc(*N*) reflects the likelihood of an abnormal event occurring within a specific segment under investigation. When *N* is identified as "irregular," the abnormality sc(*N*) must satisfy the condition that it falls within the range of 0 to *T*. This indicates that for irregular video clips, the score can take values greater than or equal to 0 and less than *T*. The function *f(x)* is employed to characterize this scenario. On the other hand, if *N* is considered "regular," the abnormality sc(*N*) should be constrained to the range of *T* to 1. This implies that for regular video clips, the score can take values greater than or equal to *T* and up to 1.

Table 1 shows the parameters employed for features extraction. The barrier *T* is responsible for classifying videos as normal or abnormal. For exception clips, the goal is to have a value less than or equal to 1, while for regular videos, the target should be less than or equal to 0 [50]. In most cases, *T* is set at 0.5. The rectifier linear activation unit (ReLU) activation function is used between fully connected (FC) layers, and there is a 50% dropout regularization between them [51]. For training, a 16-million frame video clip is used, with each frame representing various local regions from previous stages. To train the localization model, the UCF Crime dataset is used, which contains 800 normal films and 810 abnormal videos. The author describes 14 distinct types of occurrences in the dataset [62].

3.2 Visual attention detection

Visual attention detection is a technique used to identify regions of interest in an image or video. This can be achieved using deep learning algorithms that are trained to detect patterns and features in the data. The system can be trained to detect suspicious activity or objects in the video feed, and alert security personnel if necessary. However, to maintain the edges of the viewing area and remove noise effectively, a texture-based bilateral BG subtraction method is used. This method ensures a permanent BG structure and identifies the visual feature area as a spatial cognition point while considering the rest as indifferent

Table 1 The variables and hyperparameters used for extracting deep features using C3D

Parameters	Value
Size of the individual batch	29
Dropout ratio	50%
Video frame resolution	140 × 180
Number of layers responsible for features extraction	7

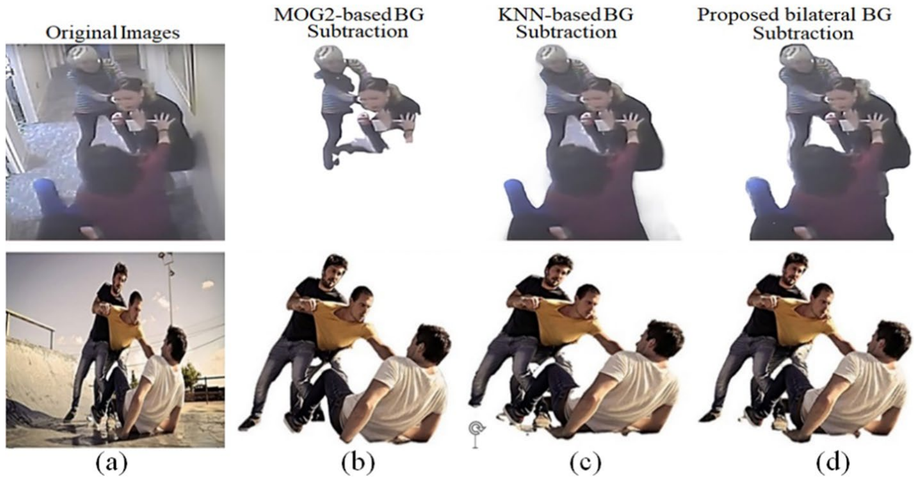


Fig. 4 Visual attention detection using various background subtraction methods. **a** Original image, **b** BG subtraction based on MOG2, **c** KNN-based BG subtraction, and (**d**) Proposed bilateral-based BG subtraction method

[52]. The approach utilized in our situation is capable of preventing the loss of edges in the viewing area. Figure 4 depicts the detection of fighting events using two commonly used background subtraction (BG) techniques, namely the hybrid Gaussian enhanced model (MOG2) and the K-Nearest Neighbor method (KNN). In these methods, the input images are analyzed on a per-pixel basis to distinguish between foreground and background regions. However, in certain scenarios, such as those involving shadow interference and intermittent motion, noise may be falsely identified as moving pixels, leading to incorrect results [53–55].

Figure 4(b and c) presents the application of two widely used background (BG) subtraction techniques, namely the hybrid Gaussian enhanced model (MOG2) and the K-Nearest Neighbor method (KNN), for detecting fighting events in a video. However, these methods may encounter certain scenarios, such as those involving shadow interference and intermittent motion, that result in the misclassification of noise as moving pixels, leading to erroneous detection. To overcome this challenge, we propose a two-sided texture-based approach that can effectively eliminate noise and extract the correct area. In both Fig. 4(b and c), the foreground regions containing noise and misclassified pixels, including most of the shadows that are falsely identified as moving pixels, are erroneously detected. Consequently, the crucial foreground regions are removed from the images because the pixels are erroneously classified as background.

In contrast, Fig. 4(d) demonstrates the efficacy of our proposed bilateral BG subtraction method, which accurately delineates the precise area of interest. By removing the noise and precisely extracting only the relevant area, our method achieves more precise and accurate results for detecting fighting events in videos.

Although prior studies have shown more complete split foreground objects [56], misclassified areas can affect the extraction of regions of interest (ROI). Therefore, in our proposed anomaly detection process, bilateral BG subtraction can be used to obtain visual areas of attention more accurately and efficiently [57]. The comparative analysis depicted in Fig. 4 is based on a UCF-Crime dataset containing various unusual activity categories. This method can be performed at up to 100 frames per second on the graphics processing unit (GPU) for input sizes of 240* 320 pixels. Consequently, this approach effectively identifies areas before using the deep learning pipeline to detect anomalies [58, 59].

To start, we apply bilateral filtering to frame I of the input image, then label the resulting grayscale image as "bilateral." We create a structure using non-overlaying slabs with two sides, where "bilateral" is split into slabs that are the same size as a pixel. In our system, the value of n is 4. Next, we generate a binary bitmap by calculating the average of each block and comparing it to the pixel values within the block.

$$Dist\left(BM_{bil_{obs}}, BM_{mod}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(b_{ij}^{bil_{obs}} + b_{ij}^{mod}\right) \tag{3}$$

This results in a $\mathrm{BM}_{\mathrm{bil}}$ bitmap, where pixels below the average are assigned a binary value of 0, and those above the average are assigned a binary value of 1. The rules we use to update the background (BG) model and determine appropriate learning rates are similar to those we have used previously. As each new frame is processed, we calculate the fencing distance to every block to determine if the detected background block is within the awareness zone. The matching bit number for the block's location (i,j) is represented as b_{ij}. Although the two-sided filter is slower than most filters, it keeps the edges of the active area clear. To process input frames, we use CUDA's texture memory and apply two-sided GPU filters instead of using global memory. The algorithm combines BG subtraction with two-sided filters. Two-sided filters are utilized in our approach to reduce the noise in the incoming public unstriped frames. The BG subtraction method is then employed to recover the foreground, which comprises the candidate note area to be recorded. Finally, a deep learning pipeline is employed to predict the attention areas of various exceptional events. Figure 5 illustrates the traffic event accident using the incremental generation of texture information for clarity. This Figure provides a simple example of how texture data can be generated on a single 44-pixel block.

Figure 6 demonstrates how a picture can be created using various block sizes with recommended descriptive shapes. In Figs. 6a, too much detail has been removed, while in Figs. 6d, there is too much unimportant detail. The texture descriptor is valid, and the block size is a good choice, as demonstrated in Fig. 6.

Figure 7 showcases the proposed framework for detecting attention regions in video frames and predicting the anomaly score using a CNN regression model. The figure consists of three main phases: input frame loading, attention region detection, and anomaly score prediction. In the first phase, an input video frame is loaded into the framework. In the second phase, attention regions are detected from the input frame, and C3D features are extracted from these regions. This step helps to narrow down the focus to the relevant parts of the frame, thereby reducing the computational burden and improving the accuracy of anomaly detection. In the third and final phase, the proposed CNN regression model takes the optimal features and predicts the anomaly score for each frame in the video. This score reflects the degree of anomaly in the frame, with higher scores indicating a greater likelihood of an anomaly event occurring. By predicting the anomaly score for each frame in the video, the framework can effectively identify and flag any anomalous activity, enabling prompt response and action.

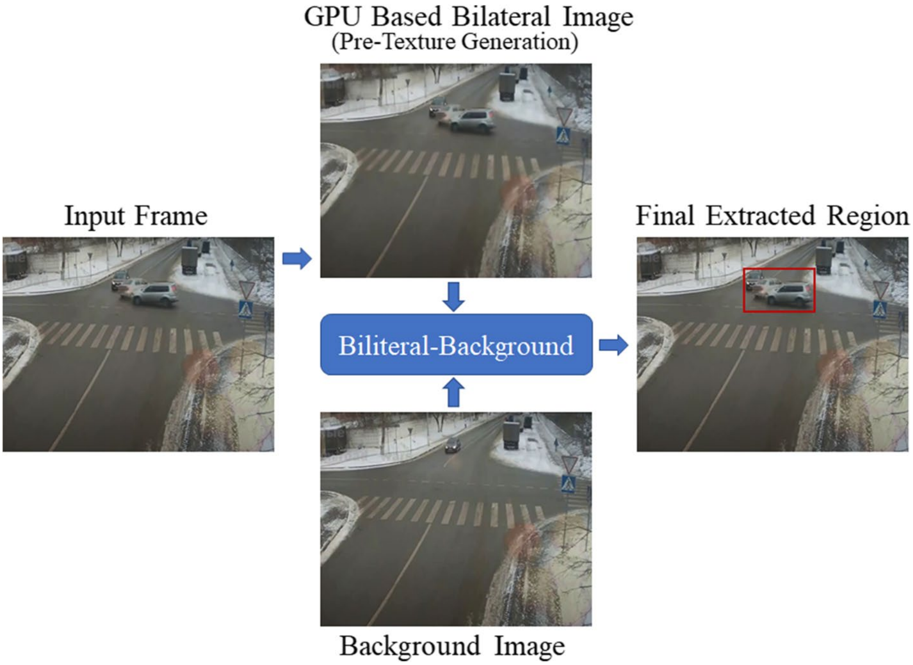


Fig. 5 The proposed GPU-based bilateral filter for attention region detection in video Frames

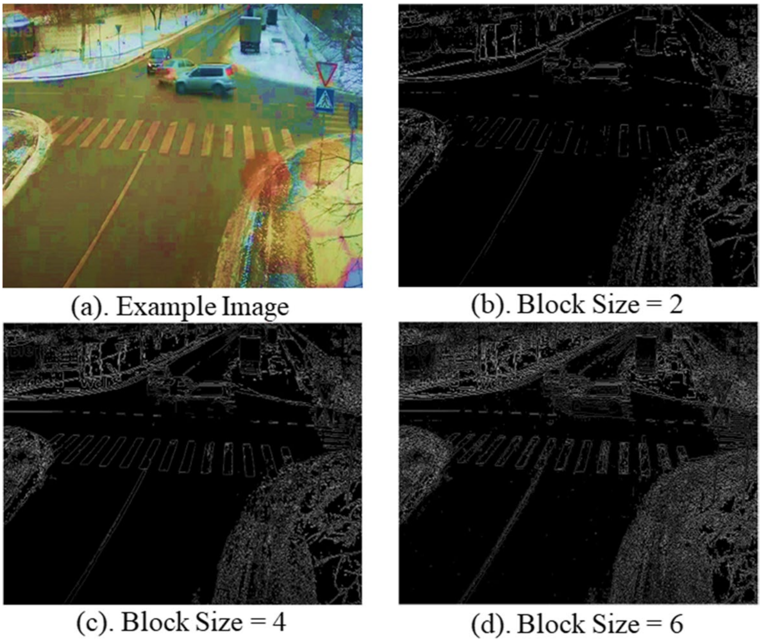


Fig. 6 The proposed texture descriptor visual results with a block size of 2, 4, and 6

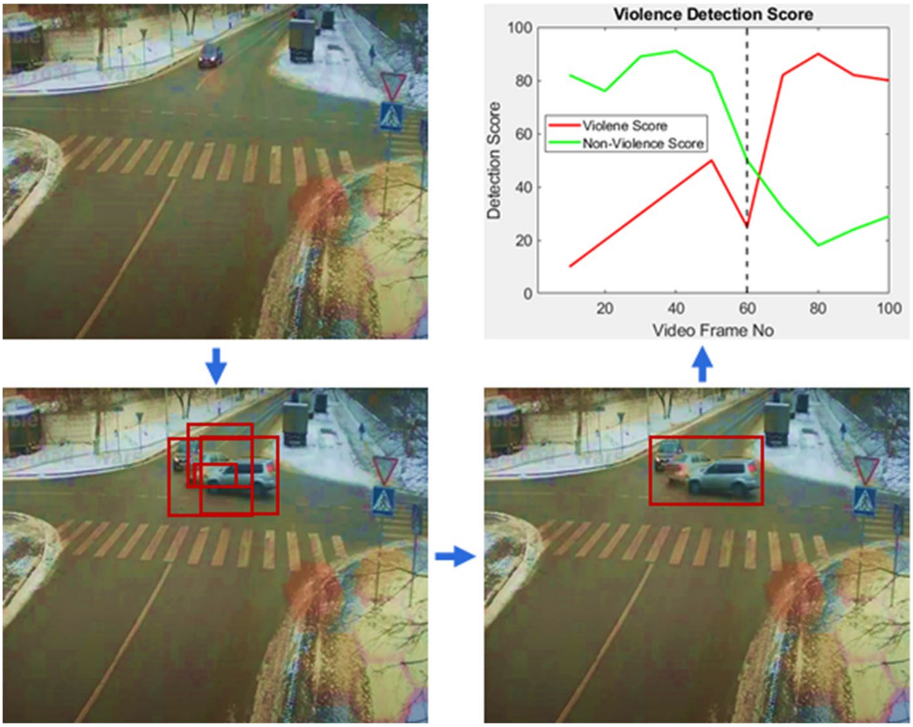


Fig. 7 An essential module of the proposed violence recognition system, showcasing the detection of visual attention regions and their use in a deep learning model for recognizing various violent scenes

3.3 Evaluation metrics

To evaluate the proposed framework, we employed the following evaluation metrics: true positive rate (TPR), false positive rate (FPR), accuracy, ROC curve, confusion matrix, precision, recall, and F-measure. The ROC curve visually represents a classifier's diagnostic ability by comparing TPR against FPR at various threshold values, and the AUC metric represents the overall performance of the classifier. The confusion matrix provides a detailed breakdown of correct and incorrect predictions, from which precision, recall, and F-measure can be calculated. A robust model should achieve high TPR and accuracy, low FPR and false alarm rates, and high precision, recall, and F-measure values for typical clips.

3.4 Computational complexity

In this study, we also evaluate the computational efficiency of our proposed method by examining the time complexity of its primary components: the attention mechanism and BG algorithm, the 3D Convolutional Neural Network (3D CNN) model for violence prediction, the Genetic Algorithm (GA) for feature selection, and the texture-based bilateral filter. As detailed in Table 2, the time complexity of each component depends on specific variables related to their operations.

The attention mechanism and BG algorithm have an $O(N)$ time complexity, where N represents the number of video frames being processed. This denotes a linear relationship between the number of frames and the computational time, which is ideal for real-time applications. The 3D CNN model, utilized for violence prediction, has a time complexity of $O(N * M^2 * K^2 * C^2$, where M is the dimension of the input matrix, K is the size of the kernel, and C is the number of channels. The CNN's complexity highlights its comprehensive processing to derive meaningful insights from the video frames. The Genetic Algorithm-based feature selection process exhibits a time complexity of $O(G * P * f)$, where G represents the number of generations, P is the population size, and F is the computational complexity of the fitness function. The complexity of the GA mirrors its extensive search across generations and populations to select the most optimal features. Lastly, the texture-based bilateral filter, which helps detect the attention region, has a time complexity of $O(W * H)$, where W and H are the width and height of the video frames, respectively. This complexity reflects the spatial operation of the filter across the frame dimensions.

The attention mechanism and BG algorithm exhibit an O(N) space complexity, where N is the number of video frames processed. This linear space complexity demonstrates that our system can manage increased data load by proportionally expanding memory usage. The proposed 3D CNN model for violence prediction has a space complexity of O(N*M^2*C), where M represents the input matrix's dimension, and C represents the number of channels. This complexity reflects the storage required for the multi-dimensional matrices used in the convolutional layers. The Genetic Algorithm-based feature selection operates with a space complexity of O(P*F), where P is the population size, and F is the size of the feature set. This component's space complexity underlines our method's capacity to handle extensive feature sets and large populations for more accurate feature selection.

Finally, the texture-based bilateral filter displays a space complexity of O(W*H), where W and H denote the width and height of the video frames, respectively. This reaffirms the scalability of our approach as the spatial dimensions of the video frames increase.

Table 2 Time complexity of components in the proposed scheme

Component	Time complexity
Attention mechanism and BG algorithm	$O(N)$
C3D model for violence prediction	$O(N * M^2 * K^2 * C^2)$
Genetic algorithm-based feature selection	$O(G * P * F)$
Texture-based bilateral filter	$O(W * H)$

Notations:

- N is the number of video frames,
- M is the dimension of the input matrix for the CNN,
- K is the size of the kernel in the CNN,
- C is the number of channels,
- G is the number of generations in the GA,
- P is the population size in the GA,
- F is the computational complexity of the fitness function in the GA,
- W and H are the width and height of the video frames, respectively, for the texture-based bilateral filter

Table 3 Space complexity of components in the proposed scheme

Component	Space complexity
Attention mechanism and BG algorithm	O(N)
C3d model for violence prediction	O(N*M^2*C)
Genetic algorithm-based feature selection	O(P*F)
Texture-based bilateral filter	O(W*H)

Notation:

N is the number of video frames,

M is the dimension of the input matrix for the CNN,

C is the number of channels,

P is the population size in the GA,

F is the size of the feature set for the GA,

W and *H* are the width and height of the video frames, respectively, for the texture-based bilateral filter

Table 4 This table displays numbers for the UCF-crime dataset, which has been detected and localized. The quantity between the brackets indicates the number of videos that are included in the training data

Category	Rate
Videos quantity	Violent: 948 (808), Normal: 949 (798)
Frame frequency	30 FPS
Original resolution	512×512x3
The typical overall number of frames	12.8 million
Length of dataset	126 h
Irregular programs checked	Fighting, Road Accidents, and Robbery

Examining space complexity assures that our proposed system can efficiently handle large-scale video surveillance data, an essential requirement in real-world urban surveillance systems (Table 3).

4 Results

The experiment utilized a core i7 CPU, 32 GB RAM, and a graphics card. To extract spatiotemporal features, Matlab 2018b was used as the framework along with a disciplined C3D model. Table 4 displays statistics for violent regions in the UCF-Crime big video repository used for model training and validation, while localized exception areas are illustrated in various categories. The UCF-crime dataset's surveillance video was obtained from LiveLeak and YouTube. To test the accuracy of the experiment, a reference set consisting of 3 types of data from the UCF-crime dataset was used. Each evaluated film was given a real label in a binary format, making it easy to completely evaluate the exam.

Figures 8, 9, 10, 11, 12, and 13 provide qualitative comparisons of unusual events in the UCF-Crime dataset, such as explosions, road accidents, theft, abuse, and normal events. Figure 8 presents visual results for normal and explosion violence recognition with four subplots. The first two subplots show input video frames, while the remaining two display

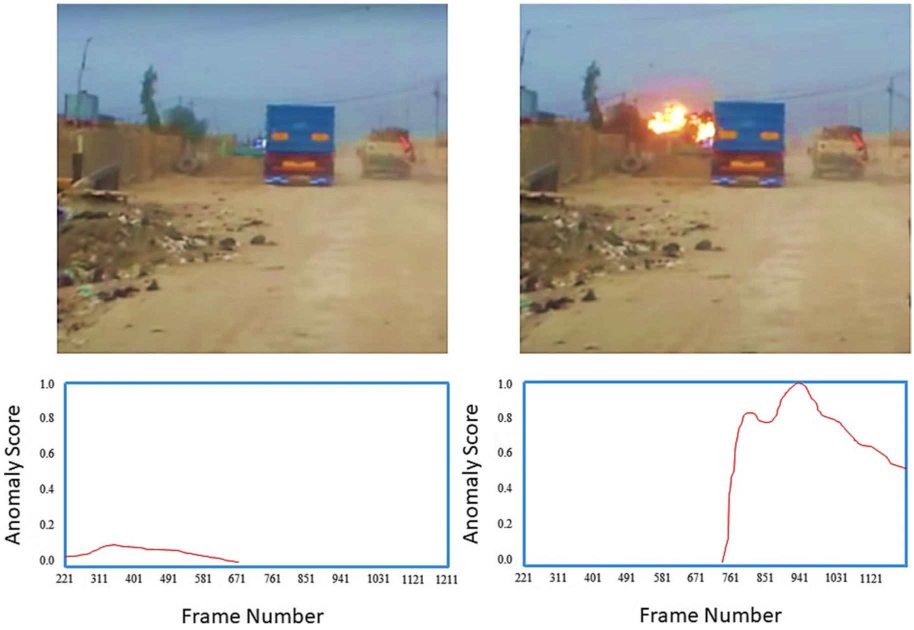


Fig. 8 Visual result of normal and explosion violence recognition

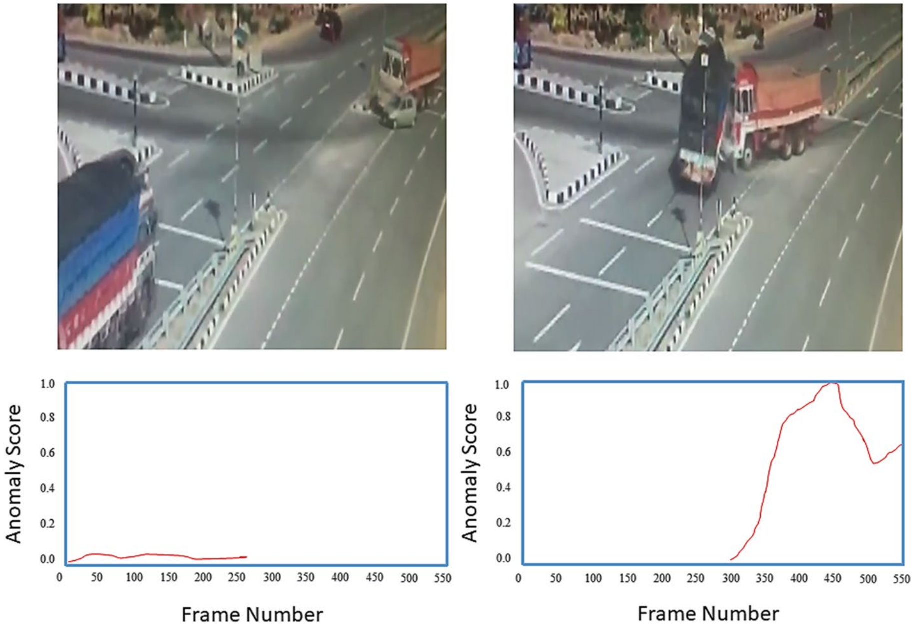


Fig. 9 The visual result of normal and road accident violence recognition

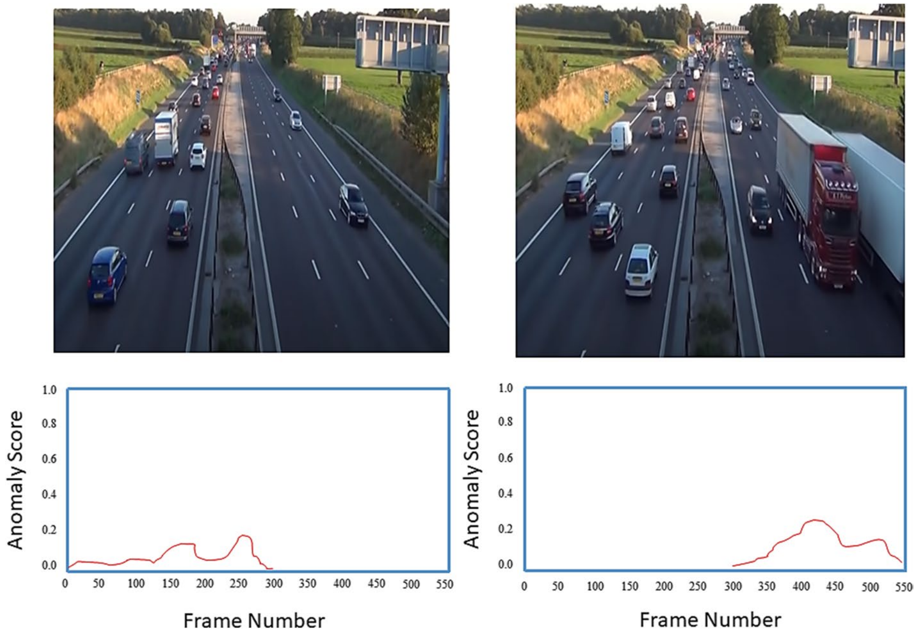


Fig. 10 The visual result of a normal video scene (no violence detected)

anomaly scores calculated by the proposed model. Anomaly scores for normal and explosive video frames are shown in the third and fourth subplots, respectively, highlighting an increase in the anomaly detection score for the explosive frame.

Figure 9 illustrates a visual comparison between normal and road accidents, as recognized by a violence recognition system. The first two subplots depict normal traffic and road accident at a junction. In contrast, the third and fourth subplot displays an anomaly score estimated by the proposed model for road accident detection. The anomaly scores for normal frame ranged from 1 to 5%, indicating no anomalies. However, the anomaly scores for normal frame rose to 100%, demonstrating a high probability of an anomaly.

Figure 10 demonstrates vehicle anomaly detection using the proposed system. In the first subplot, an input frame with no anomalies is accurately detected. The second subplot shows a complex scene with multiple closely spaced and overlapping vehicles. Despite this complexity, the model correctly identifies this frame as normal, with no increase in the anomaly score, indicating the absence of anomalies..

In Fig. 11, a scene depicts a man stealing from a store. In the first image, where no stealing occurs, there is no increase in the stealing score. However, in the second image from the theft scene, the anomaly score rises, indicating an anomaly in the video. These images effectively and confidently identify thefts, as demonstrated in the highlighted frames from 800 to 1400.

Figure 12 depicts the visual results of normal and abuse violence recognition through subplots. The first and second subplot shows a normal and physical abuse scene between a patient and a nurse, as detected by the proposed system. The third and fourth subplot exhibits the estimated anomaly detection score, The anomaly detection score is higher in the fourth subplot, indicating an increase in anomaly detection due to the occurrence of physical abuse. The concept of BG subtraction demonstrates that our method is

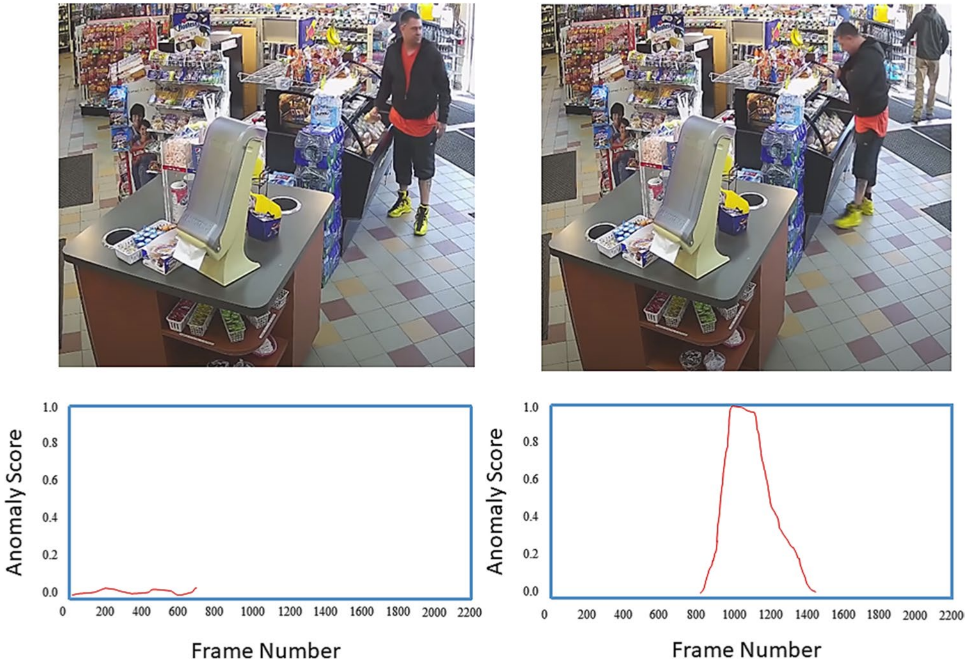


Fig. 11 The visual result of normal and theft violence recognition

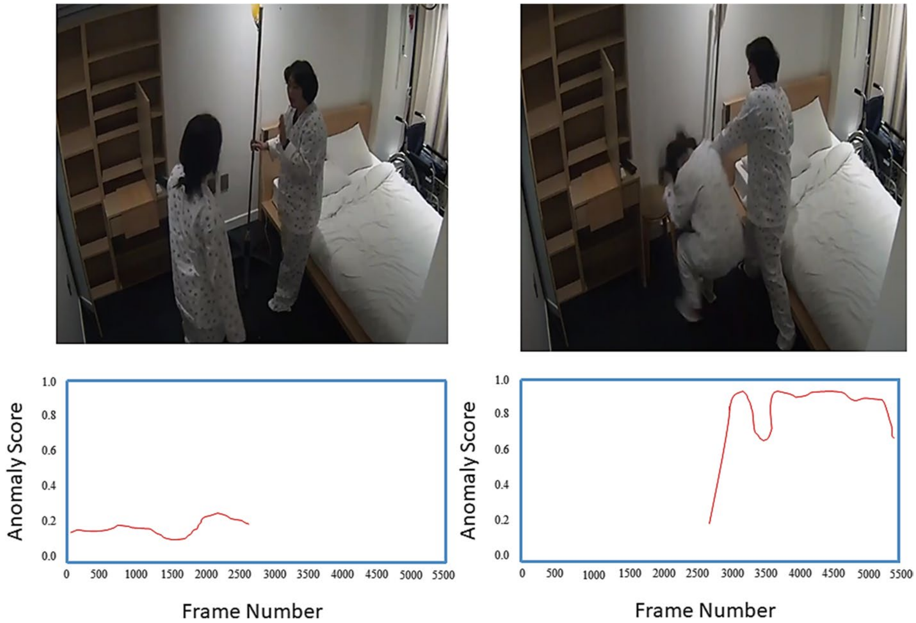


Fig. 12 The visual result of normal and abuse violence recognition

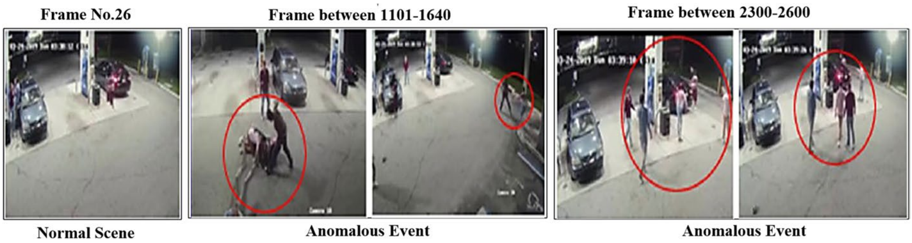


Fig. 13 Manually segmented the violent and normal frames of the UCF-Crime dataset

capable of locating anomalous areas. In industrial applications, focusing on the appropriate abnormal area rather than performing comprehensive audits can help safety managers focus on the appropriate abnormal area.

Figure 13 illustrates the attentional regions assigned to the machine learning pipeline, highlighted in red. Regions that are not of interest will appear blurred and will not contribute significant visual features during retrieval, training, and reasoning processes, as the brain does not allocate attention to them.

We draw a comparison between the methodology we provided and Sultani et al. prior work [6]. Their techniques are limited to a specific type of abnormal event, like a battle situation, as proposed in [10], despite many research studies utilizing location to detect anomalies. The suggested approach is compared to the full-frame technique to analyze a large number of exception occurrences. Figure 14 shows the confusion matrix of the proposed model on the UCF-Crime dataset, obtained using a testing set consisting of

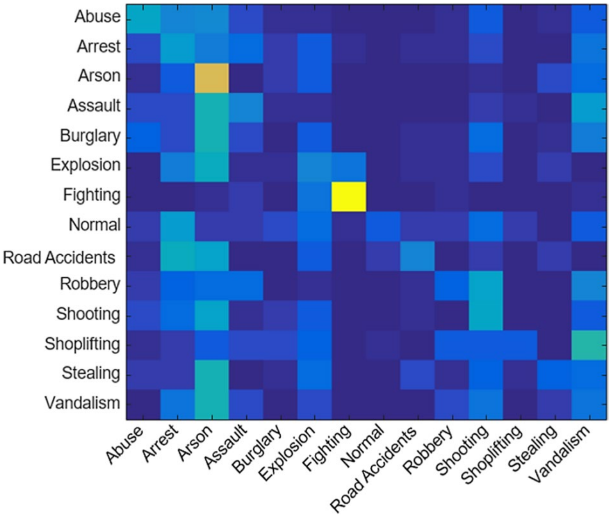


Fig. 14 Confusion matrix for our proposed violence recognition model

30% of the total dataset. The confusion matrix displays the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each class.

The proposed positioning method can be used to achieve higher accuracy on multiple tested videos, as shown in Table 5. In addition, this table also shows the accuracy achieved in percentage between entire frame violence recognition and proposed localized region-based violence detector. Table 5 presents the evaluation of a proposed violence recognition system, as well as a comparison with a state-of-the-art system [6]. For the existing system, the accuracy is reported as 97.93%, 99.90%, and 77.70% for robbery, fighting, and road accidents, respectively, while the overall accuracy is reported as 91.84%. For the proposed model, the accuracy for recognizing violent activities is reported as 97.00%, 99.00%, 96.35%, 95.00%, 94.50%, 93.80%, 97.20%, 96.70% and 96.89% for Robbery, Fighting,

Table 5 Evaluation of the proposed model using a validation set and comparing the performance with a state-of-the-art violence recognition system

Technique	Categories	Accuracy	Precision	Recall	F1-score	AUC score
Existing system [6]	Robbery	97.93%	97%	98%	97.5%	98.5%
	Fighting	99.90%	99.5%	99.9%	99.7%	99.8%
	Road Accident	77.70%	76%	80%	78%	79.0%
	Average	91.84%	90.8%	92.6%	91.7%	92.4%
Proposed model	Robbery	97.00%	97.20%	97.5%	97.35%	97.8%
	Fighting	99.00%	98.80%	99.1%	98.95%	99.2%
	Road Accident	96.35%	96.10%	96.7%	96.40%	96.6%
	Arson	95.00%	94.70%	95.3%	95.41%	95.2%
	Vandalism	94.50%	93.30%	92.7%	92.50%	92.6%
	Burglary	93.80%	96.50%	94.1%	98.10%	94.0%
	Explosion	97.20%	97.00%	97.3%	93.15%	97.4%
	Shooting	96.70%	96.50%	96.8%	96.65%	96.9%
	Average	96.89%	96.26%	96.19%	96.06%	96.21%

Road Accident, Arson, Vandalism, Burglary, Explosion and Shooting, respectively, with an overall accuracy of 96.89%. The accuracy of each test video is calculated using each fragment containing an exception event. Robberies, for example, can be found in video clips 14 through 15, fights in clips 4 through 18, and traffic accidents in clips 5 through 7. As a result, the accuracy should be assessed across multiple fragments and an average score calculated. We propose methods that, on average, achieve consistent accuracy across all videos tested as shown in Table 5. We also extracted the appropriate C3D features of approximately 135 test filmed clips from the UCF-Crime datum to compare the learning model's accuracy. To calculate accuracy, multiply lots of videos examined by the properly anticipated quantity. 133 of 135 videos were correctly labeled, and 134 of 135 videos were correctly classified using the visual attention learning we presented.

Several events can occur in a CCTV video in real-world scenarios. For example, a thief tries to steal something and is retaliated against by the victim. As a result, we chose a visual representation of a video that had been tested on YouTube at random. A thief attempts to rob an assembled multitude inside a train junction but is unsuccessful, and the people successfully defend themselves against their belongings. Like earlier UCF-Crime assessments is shown in Fig. 13, we analyze respectively filmed clips in depth by examining and determining their correctness. Although previous work failed to locate an assembly of persons and the bandits were forced to flee the spot, the proposed local approach detects two separate pieces of anomalies, as shown in the qualitative measures below. Ground truth is manually annotated by inspecting each frame individually. Frame No from 1–1100 contains normal video frames. The video frames Between 1101 and 1640 belong to a violent class, similarly, the video frames from 2300 to 2600 contain violent scenes. The accuracy of the violent scene classification can be seen in Table 5.

The performance of the model in terms of accuracy and ROC can be evaluated using the measures from the confusion matrix, as demonstrated in Fig. 15. These measures are crucial for assessing the model's ability to correctly classify instances of anomalies and normal behavior in surveillance videos.

Table 6 presents a performance comparison of several anomaly detection models for surveillance video analysis, including Anomaly-Event Detection (AED), Spatio-Temporal Low-rank Fusion (STLF) Autoencoder, Convolutional 3D (C3D) Multiple Instance Learning (MIL-DeepRank), Graph Convolutional Network (GCN) Anomaly Detection, C3D-based Local Adaptive Weighted Surrogate (CLAWS) Anomaly Detection, and the proposed Anomaly Detection Model that utilizes a C3D backbone and Evolutionary Search (ES) technique. The evaluation of each model was based on two metrics, namely False

Fig. 15 Shows the performance comparison of the proposed model with state-of-the-art using a ROC curve

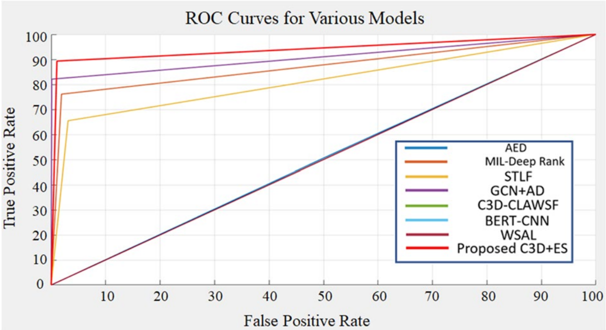


Table 6 Performance comparison of anomaly detection models for surveillance videos based on false alarm rate and AUC metrics

Authors / Year	Methods	False alarm (%)	AUC (%)
Lu et al., / 2013 [1]	Anomaly-Event Detection (AED)	27.20	50.60
Sultani et al., / 2018 [2]	MIL-DeepRank (C3D)	1.89	76.12
Hasan et al., / 2016 [3]	Spatio-Temporal Low-rank Fusion (STLF) Autoencoder	3.10	65.51
Zhong et al., / 2019 [4]	Graph Convolutional Network (GCN) Anomaly Detection (AD)	0.10	82.12
Zaheer et al., / 2020 [5]	C3D-based Local Adaptive Weighted Surrogate (CLAWS) Anomaly Detection	NA	83.03
BERT-CNN., / 2022 [6]	CNN Features with BERT Transformer Model for Anomaly Detection	NA	86.71
CamNuvem., / 2022 [7]	Weakly Supervised Anomaly Detection	NA	88.75
Proposed method	Attention C3D + ES	0.0105	89.33

Alarm rate and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

False Alarm rate indicates the percentage of normal frames that were wrongly classified as anomalous, while AUC measures how well the model differentiates between anomalous and normal frames overall. Among all the models, the proposed Anomaly Detection Model, which employed a C3D backbone and Evolutionary Search (ES) technique, achieved the lowest False Alarm rate of only 0.021% and the highest AUC of 89.33%. This outstanding performance suggests that the proposed model is a suitable choice for video surveillance analysis since it can accurately detect anomalous events with minimal false alarms. The other models also showed good performance, with AUC values ranging from 50.60% to 88.75%. However, they had higher False Alarm rates compared to the proposed model, which may limit their practical application in real-world scenarios. Overall, the results of Table 6 highlight the importance of evaluating the performance of anomaly detection models using multiple metrics and the potential of C3D-based models and evolutionary search techniques for accurate and efficient video surveillance analysis.

This table compares the performance of different methods for anomaly/event detection [1] using video data. The methods vary in terms of their approach, including Anomaly-Event Detection (AED), MIL-DeepRank (C3D) [2], Spatio-Temporal Low-rank Fusion (STLF) Autoencoder [4], Graph Convolutional Network (GCN) Anomaly Detection [5], C3D-based Local Adaptive Weighted Surrogate (CLAWS) Anomaly Detection [6], and the proposed method which uses Attention C3D + ES [7].

The table shows the false alarm rate and AUC score for each method, which are metrics used to evaluate the effectiveness of the anomaly detection models. A lower false alarm rate indicates that the model is less likely to generate false positives (i.e., identifying normal events as anomalies), while a higher AUC score indicates that the model is better at distinguishing between normal and anomalous events. The proposed model has also been systematically evaluated and has shown better performance by comparing it with two state-of-the-art models that are based on graph CNN and clustering-based unsupervised learning [60, 61].

Overall, the proposed method has the lowest false alarm rate (0.0105) and the highest AUC score (89.33%), indicating that it performs better than the other methods in detecting anomalies in video data. It's important to note, however, that the performance of these methods may vary depending on the specific dataset and task they are applied to.

Some of Advantages in proposed attention mechanism and BG (Background) algorithm Inference Speed exhibit linear time complexity, denoted as O(N), where N represents the number of video frames processed. This linear relationship between the number of frames and computational time is highly advantageous for real-time applications. In our experiments, this characteristic allowed us to process video frames rapidly, ensuring timely violence prediction and incident detection. The Efficient Time Complexity using the GA for feature selection operates with a time complexity of O(G*P*F), where G represents the number of generations, P is the population size, and F is the computational complexity of the fitness function. Despite its extensive search across generations and populations, the GA demonstrates efficient performance in selecting optimal features. Experimental results confirm that our approach efficiently narrows down the feature set, enhancing inference speed without compromising predictive accuracy. Advantages in Computational Efficiency.

Both the attention mechanism and BG algorithm have advantages in Computational Efficiency by achieving linear space complexity, O(N), where N represents the number of video frames processed. This means that our system can effectively manage increased data loads by proportionally expanding memory usage. In practical terms, it ensures that our method can be deployed in scenarios involving large-scale video surveillance data, which is crucial for urban surveillance systems. The GA-based feature selection process has a space complexity of O(P*F), where P is the population size, and F is the size of the feature set. This space efficiency allows our method to handle extensive feature sets and large populations while keeping memory requirements manageable. Our experiments confirm that this approach balances computational efficiency with feature selection accuracy. The texture-based bilateral filter, used for attention region detection, has a space complexity of O(W*H), where W and H denote the width and height of video frames, respectively. This scalability ensures that our approach can efficiently process video frames with varying spatial dimensions.

5 Conclusion

This article presents a computer vision model that utilizes deep learning to ensure safety and provide assistance to individuals in an intelligent city surveillance system. The proposed model has the capability to recognize various types of violent and normal situations. This technology can be deployed by law enforcement agencies to prevent and address any suspicious activity that may endanger public safety and property. The study found that identifying a local attention area in each video segment is useful for detecting anomalies. The results demonstrate that the proposed approach is highly accurate in identifying a wide range of events such as traffic accidents, robberies, and fights. Moreover, combining robust BG subtractions can help identify the most crucial area of interest. The proposed visual attention model has an impressive precision rate of 99%. However, simply installing video surveillance cameras is not enough to reduce criminal activity. To provide rapid assistance to victims and actively pursue offenders, a mechanism must be implemented that involves continuous and thorough monitoring of video surveillance footage, which may require significant human resources. Therefore, a real-time automated system for detecting anomalous individual interactions may be a promising solution to these issues. This approach can be applied in various public spaces, including schools, colleges, airports, bus stops, hospitals, and train stations. For example, it can detect a traffic accident and automatically contact an ambulance. By leveraging the intermediate outcomes of the adaptive video classification system, a precise and real-time anomaly detection system can be developed. The proposed method surpasses other existing approaches in terms of both precision and speed. In future research, we aim at improving the system by expanding our efforts to optimize human-AI collaboration for monitoring and decision-making processes related to anomaly violence detection in video. Additionally, we plan to extend the system's capabilities to seamlessly integrate data from multiple surveillance cameras, allowing for the comprehensive tracking of individuals and objects across various camera feeds. This approach will significantly enhance situational awareness within urban environments and contribute to more effective violence detection.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No.62172366), and "Pioneer" and "Leading Goose" R & D. Program of Zhejiang Province (2023C01150). This research work was also supported by the Cluster grant R20143 of Zayed University, UAE.

Data availability The datasets used in this research work are already online available on the following link: https://www.crcv.ucf.edu/projects/real-world/.

Declarations

Conflicts of interest The authors of this manuscript declare no conflicts of interest.

References

1. Cárdenas AA, Amin S, Sastry S (2008) Secure control: towards survivable cyber-physical systems. Proc - Int Conf Distrib Comput Syst:495–500. https://doi.org/10.1109/ICDCS.Workshops.2008.40
2. Ghazal S, Khan US, Saleem MM, Rashid N, Iqbal J (2019) Human activity recognition using 2D skeleton data and supervised machine learning. IET Image Process 13(13):2572–2578. https://doi.org/10.1049/iet-ipr.2019.0030
3. Ding W, Liu K, Belyaev E, Cheng F (2018) Tensor-based linear dynamical systems for action recognition from 3D skeletons. Pattern Recognit 77:75–86. https://doi.org/10.1016/j.patcog.2017.12.004
4. Dong J, Jiang W, Huang Q, Bao H, Zhou X Fast and robust multi-person 3D pose estimation from multiple views
5. Wang X, Yang LT, Song L, Wang H, Ren L, Deen MJ (2021) A tensor-based multiattributes visual feature recognition method for industrial intelligence. IEEE Trans Ind Inf 17(3):2231–2241. https://doi.org/10.1109/TII.2020.2999901
6. Tan W, Yao Q, Liu J (2022) Overlooked video classification in weakly supervised video anomaly detection. *arXiv preprint arXiv:2210.06688*. https://doi.org/10.48550/arXiv.2210.06688
7. Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89(1–2):31–71. https://doi.org/10.1016/s0004-3702(96)00034-3
8. Irfanullah, Hussain T, Iqbal A, Yang B, Hussain A (2022) Real time violence detection in surveillance videos using convolutional neural networks. Multimed Tools Appl: 1–23.https://doi.org/10.1007/s11042-022-13169-4
9. Landi F, Snoek CGM, Cucchiara R (2019) Anomaly locality in video surveillance. [Online]. Available: http://arxiv.org/abs/1901.10364
10. Xu Q, See J, Lin W (2019) Localization guided fight action detection in surveillance videos. Proc - IEEE Int Conf Multimed Expo 2019-July:568–573. https://doi.org/10.1109/ICME.2019.00104
11. Jain M, Van Gemert J, Jegou H, Bouthemy P, Snoek CGM (2014) Action localization with tubelets from motion. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*:*740–747. https://doi.org/10.1109/CVPR.2014.100
12. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. 8.1–8.12. https://doi.org/10.5244/c.29.8
13. Wu S, Moore BE, Shah M (2010) Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit:2054–2060. https://doi.org/10.1109/CVPR.2010.5539882
14. Basharat A, Gritai A, Shah M (2008) Learning object motion patterns for anomaly detection and improved object detection. 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR. https://doi.org/10.1109/CVPR.2008.4587510
15. Cui X, Liu Q, Gao M, Metaxas DN (2011) Abnormal detection using interaction energy potentials. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit:3161–3167. https://doi.org/10.1109/CVPR.2011.5995558
16. Antić B, Ommer B (2011) Video parsing for abnormality detection. Proc IEEE Int Conf Comput Vis:2415–2422. https://doi.org/10.1109/ICCV.2011.6126525
17. Hospedales T, Gong S, Xiang T (2009) A Markov clustering topic model for mining behaviour in video. Proc IEEE Int Conf Comput Vis (Iccv):1165–1172. https://doi.org/10.1109/ICCV.2009.5459342
18. Zhu Y, Nayak NM, Roy-Chowdhury AK (2013) Context-aware activity recognition and anomaly detection in video. IEEE J Sel Top Signal Process 7(1):91–101. https://doi.org/10.1109/JSTSP.2012.2234722
19. Gnouma M, Ejbali R, Zaied M (2020) Video anomaly detection and localization in crowded scenes. Adv Intell Syst Comput 951(10):87–96. https://doi.org/10.1007/978-3-030-20005-3_9
20. Kratz L, Nishino K (2009) Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. 2009 IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2009, no. June, pp. 1446–1453. https://doi.org/10.1109/CVPRW.2009.5206771
21. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 FPS in MATLAB. Proc IEEE Int Conf Comput Vis:2720–2727. https://doi.org/10.1109/ICCV.2013.338

22. Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit:3313–3320. https://doi.org/10.1109/CVPR.2011.5995524
23. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences: supplementary material. Cvpr, pp. 1–31, [Online]. Available: http://arxiv.org/abs/1604.04574
24. Cheng KW, Chen YT, Fang WH (2015) Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 07:2909–2917. https://doi.org/10.1109/CVPR.2015.7298909
25. Cong Y, Yuan J, Liu J 2002_Studi Tingkah Laku Pelolosan Kerapu Macan (Epinephelus fuscoguttatus) PADA BUBU (skripsi).pdf
26. Dutta JK, Banerjee B (2015) Online detection of abnormal events using incremental coding length. Proc Natl Conf Artif Intell 5:3755–3761. https://doi.org/10.1609/aaai.v29i1.9799
27. Ionescu RT, Smeureanu S, Popescu M, Alexe B (2019) Detecting abnormal events in video using narrowed normality clusters. Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vision, WACV 2019, pp. 1951–1960. https://doi.org/10.1109/WACV.2019.00212
28. Kim J, Grauman K (2009) Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. 2009 IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2009, no. June, pp. 2921–2928. https://doi.org/10.1109/CVPRW.2009.5206569
29. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. 2009 IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2009, no. 1, pp. 935–942. https://doi.org/10.1109/CVPRW.2009.5206641
30. Leyva R, Sanchez V, Li C, Member S (2017) Feature sets for online performance. 26(7): 3463–3478
31. Ahmed SA, Dogra DP, Kar S, Roy PP (2019) Trajectory-based surveillance analysis: a survey. IEEE Trans Circuits Syst Video Technol 29(7):1985–1997. https://doi.org/10.1109/TCSVT.2018.2857489
32. Zhang T, Jia W, Gong C, Sun J, Song X (2018) Semi-supervised dictionary learning via local sparse constraints for violence detection. Pattern Recognit Lett 107:98–104. https://doi.org/10.1016/j.patrec.2017.08.021
33. Pękalska E, Tax DMJ, Duin RPW (2003) One-class LP classifier for dissimilarity representations. Adv Neural Inf Process Syst
34. Zhang T, Jia W, Yang B, Yang J, He X, Zheng Z (2017) MoWLD: a robust motion image descriptor for violence detection. Multimed Tools Appl 76(1):1419–1438. https://doi.org/10.1007/s11042-015-3133-0
35. Zhang T, Jia W, He X, Yang J (2017) Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans Circuits Syst Video Technol 27(3):696–709. https://doi.org/10.1109/TCSVT.2016.2589858
36. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) "Learning temporal regularity in video sequences. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-Decem:733–742. https://doi.org/10.1109/CVPR.2016.86
37. Hinami R, Mei T, Satoh S (2017) Joint detection and recounting of abnormal events by learning deep generic knowledge. Proc IEEE Int Conf Comput Vis 2017-Octob:3639–3647. https://doi.org/10.1109/ICCV.2017.391
38. Smeureanu S, Ionescu RT, Popescu M, Alexe B (2017) Deep appearance features for abnormal behavior detection in video. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 10485 LNCS:779–789. https://doi.org/10.1007/978-3-319-68548-9_70
39. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked RNN framework. Proc IEEE Int Conf Comput Vis 2017-Octob:341–349. https://doi.org/10.1109/ICCV.2017.45
40. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) DITEN, University of Genova DISI, University of Trento Carlos III University of Madrid. Icip, pp. 1577–1581
41. Hanson A, Pnvr K, Krishnagopal S, Davis L (2019) Bidirectional convolutional LSTM for the detection of violence in videos. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11130 LNCS:280–295. https://doi.org/10.1007/978-3-030-11012-3_24
42. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14
43. Zivkovic Z (2004) Improved adaptive Gaussian mixture model for background subtraction. Proc - Int Conf Pattern Recognit 2:28–31. https://doi.org/10.1109/icpr.2004.1333992
44. Zivkovic Z, Van Der Heijden F (2006) Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognit Lett 27(7):773–780. https://doi.org/10.1016/j.patrec.2005.11.005

45. Curtis JB, Zumberge JE, Brown SW, Park N (2013) Evaluation of Niobrara and Mowry formation petroleum systems in the Powder River, Denver and Central Basins of the Rocky Mountains, Colorado and. no. March, pp. 31–33
46. Yeh CH, Lin CY, Muchtar K, Kang LW (2014) Real-time background modeling based on a multi-level texture description. Inf Sci (NY) 269:106–127. https://doi.org/10.1016/j.ins.2013.08.014
47. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. Proc IEEE Int Conf Comput Vis 2015 Inter:4489–4497. https://doi.org/10.1109/ICCV.2015.510
48. Pinhanez CS (1999) Representation and recognition of action in interactive spaces. Media Arts Sci Progr
49. Khan UA, Javed A, Ashraf R (2021) An effective hybrid framework for content based image retrieval (CBIR). Multimed Tools Appl 80(17):26911–26937. https://doi.org/10.1007/s11042-021-10530-x
50. Koller D, Weber J, Malik J (1994) Robust multiple car tracking with occlusion reasoning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 800 LNCS:189–196. https://doi.org/10.1007/3-540-57956-7_22
51. Ivanov YA, Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22(8):852–872. https://doi.org/10.1109/34.868686
52. Ren H, Liu W, Olsen SI, Escalera S, Moeslund TB (2015) Unsupervised behavior-specific dictionary learning for abnormal event detection 28.1–28.13. https://doi.org/10.5244/c.29.28
53. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst 156:117–127. https://doi.org/10.1016/j.cviu.2016.10.010
54. Zhang Y, Lu H, Zhang L, Ruan X, Sakai S (2016) Video anomaly detection based on locality sensitive hashing filters, vol 59. Elsevier
55. Kooij JFP, Liem MC, Krijnders JD, Andringa TC, Gavrila DM (2016) Multi-modal human aggression detection. Comput Vis Image Underst 144:106–120. https://doi.org/10.1016/j.cviu.2015.06.009
56. Saleemi I, Shafique K, Shah M (2009) Probabilistic modeling of scene dynamics for applications in visual surveillance. IEEE Trans Pattern Anal Mach Intell 31(8):1472–1485. https://doi.org/10.1109/TPAMI.2008.175
57. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process Image Commun 47:358–368. https://doi.org/10.1016/j.image.2016.06.007
58. Jian M, Lam KM, Dong J (2014) Illumination-insensitive texture discrimination based on illumination compensation and enhancement. Inf Sci (NY) 269:60–72. https://doi.org/10.1016/j.ins.2014.01.019
59. Lin CY, Muchtar K, Lin WY, Jian ZY (2020) Moving object detection through image bit-planes representation without thresholding. IEEE Trans Intell Transp Syst 21(4):1404–1414. https://doi.org/10.1109/TITS.2019.2909915
60. Zhong JX, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:1237–1246. https://doi.org/10.1109/CVPR.2019.00133
61. Zaheer MZ, Mahmood A, Astrid M, Lee SI (2020) CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 12367 LNCS:358–376. https://doi.org/10.1007/978-3-030-58542-6_22
62. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479-6488. https://doi.org/10.48550/arXiv.1801.04264

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Muhammad Shoaib[1,2] · Babar Shah[3] · Tariq Hussain[4,5] · Bailin Yang[4,5] · Asad Ullah[2] · Jahangir Khan[2] · Farman Ali[6]

✉ Bailin Yang
ybl@mail.zjgsu.edu.cn

✉ Farman Ali
farmankanju@gmail.com

Muhammad Shoaib
mshoaib@cecos.edu.pk

Babar Shah
Babar.Shah@zu.ac.ae

Asad Ullah
asadullah.csit@suit.edu.pk

Jahangir Khan
jahangir.csit@suit.edu.pk

[1] Department of Computer Science, CECOS University of IT and Emerging Sciences, Peshawar 25000, Pakistan

[2] Department of Computer Science and Information Technology, Sarhad University of Science & Information Technology, Peshawar 25000, Pakistan

[3] College of Technological Innovation, Zayed University, 19282 Dubai, United Arab Emirates

[4] School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China

[5] School of Mathematics and Statistics, Zhejiang Gongshang University, Hangzhou 310018, China

[6] Department of Computer Science and Engineering, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, South Korea