Intelligent Video Surveillance Systems for Real-Time Suspicious Activity Detection Using AI and Computer Vision: A Study

J. Arunnehru[1], Divya Midhunchakkaravarthy [2], S. Hemalatha[3]

[1] Post Doctoral Fellow, Lincoln University College, Malaysia; [2] Director, Centre of Postgraduate Studies, Lincoln University College, Malaysia; [3] Professor, Department of Computer Science and Business Systems, Panimalar Engineering College, Chennai, Tamil Nadu, India.

arunnehruj@gmail.com, divya@lincoln.edu.my, pithemalatha@gmail.com

---

Abstract: This comprehensive review explores the current state of artificial intelligence (AI) techniques for detecting suspicious activities in video footage, with a focus on physical assault. The rapid growth of surveillance systems has necessitated automated methods for identifying potential security threats in real-time. The study examines various intelligent techniques, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and transformer-based models, used for feature extraction and classification in violence detection algorithms. It analyzes the effectiveness of different approaches, such as skeleton-based methods and audio-based detection, and discusses the challenges faced in developing robust systems. The review also provides an overview of commonly used datasets for training and evaluating suspicious activity detection models. By synthesizing recent advancements in the field, this paper aims to offer insights into the most promising approaches for enhancing the accuracy and efficiency of AI-based suspicious activity detection systems, while also addressing ethical concerns related to surveillance and privacy.

Keywords: Artificial Intelligence (AI); Video Surveillance, Physical Assault Detection; Convolutional Neural Networks (CNNs); Long Short-Term Memory (LSTM) Networks; Transformer-based Models; Suspicious Activity Recognition

---

Introduction

The rapid growth of surveillance systems in urban areas has led to an overwhelming influx of video data, making traditional manual monitoring methods ineffective and unfeasible. This surge in data volume has necessitated the creation of automated systems capable of identifying suspicious activities in real-time. High-risk locations such as public transit, financial institutions, educational facilities, and roadways face various security and safety risks, including theft, accidents, vandalism, and terrorism. Intelligent techniques for suspicious activity detection involve using advanced algorithms and computational models to examine video feeds, recognize patterns, and identify unusual behaviors that might indicate potential security threats. These systems aim to augment human capabilities by continuously monitoring multiple video streams, swiftly processing visual information, and notifying security personnel. The use of intelligent techniques in detecting suspicious activity offers several advantages. Firstly, it enables round-the-clock surveillance without the limitations of human exhaustion or attention span. Secondly, these

systems can adapt and learn new patterns of suspicious behavior over time, improving their accuracy and reducing false alarms. However, challenges exist, including the need for robust algorithms to distinguish between normal and suspicious activities in diverse settings, the requirement for substantial computational resources to process high-quality video feeds in real-time, and ethical concerns related to surveillance and privacy. This research aims to explore the current state of anomalous activity detection using intelligent techniques, assess the effectiveness of various methods, and propose innovative approaches to enhance the accuracy and efficiency of these systems.

Suspicious activity significantly impacts various aspects of life, affecting direct victims, their families, society at large, and daily routines. The persistent state of insecurity also influences citizens' mental health [1] and the economy (shopping, travel, tourism, etc.) [2]. For instance, the world's greatest cause of mortality for people between the ages of 15 and 44 is physical assault [3]. Some studies reveal concerning statistics. According to Hillis et al. (2016) [4], a study published in the American Academy of Pediatrics, at least 50% or more of children in Asia, Africa, and North America experienced violence in 2015. Furthermore, over half of the world's children—roughly 1 billion people between the ages of 2 and 17—have witnessed such violence. In just one year, more than 22 million people in Europe were physically assaulted, and more than one in four were harassed, according to research from the European Union Agency for Fundamental Rights (FRA) [5]. These factors underscore the critical need to address the issue of suspicious activity in societies worldwide

Recent research has explored mid-range approaches to combat violence by examining the relationship between crime rates, street population density, and urban landscapes. Using statistical techniques focused on spatial analysis, researchers have examined social media and cell phone data to measure urban population density and investigate its relationship with crime rates [6]. However, these methods struggle to differentiate between indoor and outdoor populations. Additionally, studies employing convolutional neural networks (CNNs) have shown that crime rates tend to be lower in green spaces [7]. In a related investigation, deep learning techniques were applied to Baidu Street View images (China's equivalent to Google Street View) to assess street population, urban landscape features, and their correlation with crime occurrences [8]. Another study analyzed Google Street View images for vehicle presence and pavement characteristics, using machine learning to correlate these factors with crime rates [9].

These mid-term approaches, which analyze crime rates in relation to street traffic, population, and urban landscape types, offer concrete insights into areas more susceptible to physical assaults and inform the design of safer urban environments. However, this offline methodology doesn't provide immediate assistance to assault victims. The primary focus of this assessment is on quick techniques and short-term, real-time remedies for averting attacks. Traditionally, security cameras and photos have been used to gather evidence following a crime, helping to identify offenders and supporting police investigations, insurance claims, and legal actions [10]. Real-time methods for identifying violent occurrences have recently been developed by a number of research. The majority of these techniques make use of security cameras that have been programmed with algorithms based on AI [11, 12, 13]. The widespread use of surveillance systems and image capture methods [14, 15], the rise of big data platforms

[16, 17], and the development of AI algorithms intended to analyze video and image footage [18] have all contributed to the field's explosive growth [20]. Globally, the deployment of security cameras has grown due to various factors, including surveillance and safety concerns [14, 15]. It is important to note that the use of images and video has numerous applications that could potentially improve citizens' quality of life. Nevertheless, existing research has highlighted the possible threats to individual privacy associated with widespread recording [19]. Modern big data platforms allow for the collection and examination of substantial amounts of information from our surroundings [16, 20]. These systems enable real-time capture, storage, and analysis of diverse data types, including visual media [17].

Artificial intelligence (AI) - driven algorithms for image and video analysis have proliferated in recent years, exhibiting improved accuracy and adaptability [18]. It is worth considering whether AI is necessary for detecting aggressive behavior. The fact is that continuous human monitoring of visual data is either financially burdensome (due to personnel costs, facilities, etc.) or results in inadequate surveillance, with few individuals overseeing numerous scenarios, potentially resulting in fatalities [21]. Developing efficient violence detection systems requires the use of based on artificial intelligence image and video analysis approaches. When it comes to recognizing aggressive actions in video recordings, computer vision—specifically, action recognition—is crucial [22]. Computer vision, a subfield of artificial intelligence, allows robots to comprehend visual information and make defensible judgments. Action recognition is a subfield of this field that focuses on identifying certain motions and actions in video clips [23]. In AI-based violence detection, models are trained to identify trends and actions linked to violent incidents [3]. Algorithms trained on labelled video datasets containing both violence and non-violent acts are the foundation of this procedure. The computers are able to recognize violent conduct in fresh, unseen video material by learning certain patterns and characteristics [24]. These algorithms may identify violent acts in formerly unseen video material because they are taught to identify patterns and traits associated with violent conduct [24].

Societal concerns about suspicious activity have widespread consequences. Numerous studies have explored this issue, proposing various solutions with differing levels of directness. The most crucial and immediate solution for identifying individuals experiencing physical harm is the real-time detection of violence, serving as the ultimate protective measure. The increasing use of surveillance footage, the rise of massive data platforms, and developments in algorithms using artificial intelligence for analyzing videos and images are the main drivers of this capacity. In order to provide a thorough and current review of AI-based identification of suspicious actions in pictures, with a focus on physical assault, this work expands upon an organized mapping investigation [24]. This paper's thorough and up-to-date analysis of every step involved in based on artificial intelligence video violence identification is one of its main contributions. Key traits are categorized and ranked according to how frequently they appear in pertinent research publications. By examining the algorithms utilized in video suspicious activity identification, the combination of them, and the results obtained from the most popular datasets in recent literature, this work also distinguishes itself.

This research aims to address the challenge of detecting suspicious activities, particularly physical assault, in video footage using artificial intelligence techniques. Previous studies have explored various

approaches, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and transformer-based models for feature extraction and classification. However, gaps exist in the comprehensive analysis of different algorithms' effectiveness and their combinations. Some methods may have been omitted due to computational limitations or dataset constraints. This research contributes to society by enhancing real-time surveillance capabilities, potentially reducing crime rates and improving public safety.

Related work

This portion analyzes the content of earlier assessments, highlighting their advantages and disadvantages to aid in conducting a thorough and pertinent review of the current field. Yao and Hu [25] emphasized the challenges of identifying suspicious activity, pointing to its rarity and ambiguity as major barriers to acquiring records from the actual world. Their study did not, however, cite any relevant publications or offer information on the selection procedure, including the databases that were used, search parameters, duration, or applied filters. The difference between deep learning (DL) techniques and conventional methods for identifying hostility was the main topic of debate. The evaluation divided cutting-edge research into two primary groups: traditional frameworks and deep learning-based techniques. The classifiers and methods for extracting features were the two subcategories into which the conventional framework part was subdivided. It included studies using important feature extraction techniques, such as feature descriptors, motion-based, appearance-based, and trajectory-based approaches. It also looked at the use of classifiers like k-nearest neighbors (KNN) and support vector machines (SVM), pointing out both their benefits and drawbacks. The articles were arranged in a table in the paper according to the feature extraction methods used and the scene type (busy or non-crowded). It did not, however, go into depth about the machine learning techniques or video classification that were applied in each research.

A visual scheme and summary table that provided a thorough understanding of suspicious activity detection by delineating important phases and ideas were published by Siddique et al. [23]. The study included a clear research approach, including search terms, databases, time periods, and filtering criteria, even though it was not a systematic review. There is a requirement for a more thorough categorization, since the article divided aggressiveness detection systems into 19 groups, some of which only had one reference. Aspects including object identification, extraction of features, scene congestion and accuracy were all covered in detail in the summary table. Nevertheless, it lacked a comparative examination of algorithm performance and omitted information on classifier types and datasets. Additionally, the dataset table only included clip counts and publication years, and the feature extraction table did not classify feature types. Kaur and Singh [26] published a conference paper summarizing the most recent evaluations on aggression detection since 2016.

The examined research were categorized into two main groups: deep learning approaches and classical methods, whose were further subdivided into the extraction of features and classification techniques. Recent developments and potential methods in the field were highlighted in the study, especially those utilizing long short-term memory (LSTM) models and convolutional neural networks

(CNNs). Key issues in attack detection were discussed, such as object occlusion, illumination fluctuations, and potential misunderstanding with other activities. The study concluded that as no one solution could successfully handle all of the problems at once, each of these obstacles should be addressed separately.
In a conference paper, Kaur and Singh [26] summarized the most recent assessments of aggressiveness detection conducted since 2016. The examined papers were divided into two main categories: deep learning approaches and classical methods, which were further subdivided into classification and extraction of features techniques. The paper emphasized the potential of recent and promising techniques in the field, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) studies. The discussion addressed various challenges in assault detection, including the potential for confusion with other actions, changes in illumination, and overlapping objects. In conclusion, the paper suggested addressing these challenges on a case-by-case basis, as there was no universal solution to tackle them collectively.

There was no specific section on study technique or related work in Shubber et al. [3]. The study highlighted the higher accuracy of deep learning techniques by classifying the examined material into two categories: standard and deep learning approaches. Along with outlining four classification strategies and providing pertinent citations, it also included tables that summarized feature extraction techniques utilized in various research, describing both feature extraction procedures and classifiers. The authors highlighted that advancements in deep learning were facilitated by high processing power and abundant data availability. The datasets used in the examined research were described in one part, together with information on the number of clips of video, frame numbers, and content types—such as real-life footage, movies, and hockey games. It did not, however, indicate if crowded scenarios were included in these datasets. Tables also showed the accuracy percentages for each dataset from different investigations.
A thorough systematic review of suspicious activity detection from 2015 to 2021 was carried out by Omarov et al. [27]. Research methodology, basic principles, categorization of detection strategies, video characteristics and descriptors, datasets, assessment criteria, and obstacles in suspicious activity detection comprised the five main components of the review. The systematic review procedure, including research objectives, article selection standards, and filtering techniques, was covered in full in the study methodology section. Along with establishing key concepts, the report also explained how physical aggressiveness detection algorithms are trained and tested. This study's portion on physical aggressiveness evaluation factors stood out since it connected these measures to pertinent research, offering insightful information. In conclusion, not all facets of suspicious activity detection are covered by the reviews that have been published in the last three years. To provide a more thorough knowledge of AI-driven suspicious activity identification in video analysis, an updated assessment is necessary.

The process begins with using a video as input, which requires compiling a dataset containing both suspicious and normal scenes. Keyframe extraction, the following stage, is experimental and not found in every design is show in the Figure 1. By choosing frames that are most likely to include violent behavior, this method lowers the amount of video data that must be analyzed and, as a result, the processing demands. Depending on the particular attributes being retrieved, the data is then changed in the next step to conform to the demands of the suspicious activity recognition algorithm, the following stage concentrates on extracting features and algorithm learning, which might differ depending on various

algorithmic combinations. The classification of a scenario into violent or non-violent is finally done using a classifier [24].

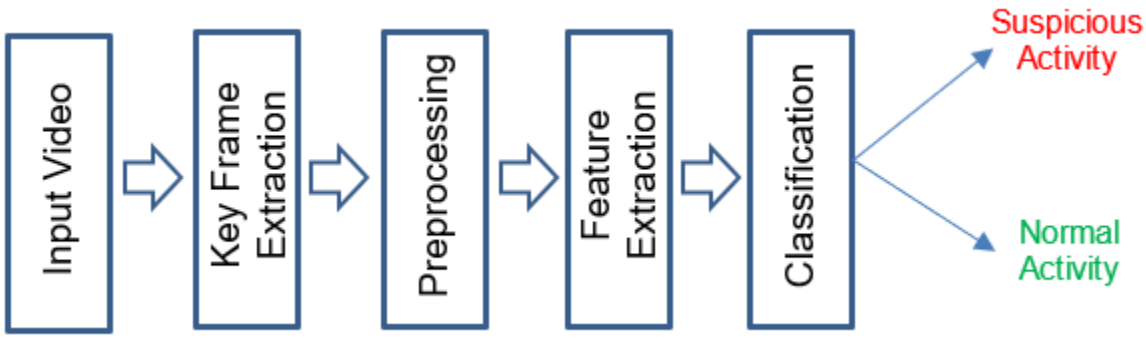

Figure 1: Basic steps of video surveillance for suspicious activity detection

The research articles under examination have been categorized according to the classification techniques employed for detecting suspicious activities. This study's categorization of various suspicious activity detection algorithms is primarily based on the type of algorithm used (including traditional feature-based methods, convolutional neural networks (CNN), long short-term memory networks (LSTM), and transformer models). Additionally, the study explored the datasets utilized for suspicious activity detection.

CNN + LSTM: This category includes articles that use a combination of long short-term memory (LSTM) and convolutional neural networks (CNNs), regardless of pre-training status. CNNs extract spatial features, which are then input into the LSTM to extract temporal features. The LSTM receives input from features extracted by both the CNN and LSTM in two distinct phases [28, 29].
CNN: Studies that solely use convolutional neural networks (CNNs) to identify suspicious activity in videos fall under this category. In order to classify video frames, CNNs examine spatial characteristics. Nevertheless, certain research using CNN-3D and CNN-4D have trouble successfully capturing both spatial and temporal patterns, as will be covered in subsequent sections [30]. Additionally popular is the use of CNNs that have already been trained to identify suspicious activity. To create baseline weights, these models are first trained on large picture datasets like ImageNet. Then, using transfer learning to increase accuracy and efficiency, they are refined to extract pertinent characteristics and identify violent patterns in video sequences [31].

Traditional feature extraction: This group consists of studies that have employed manual techniques for extracting features, either solely for feature extraction or for both extraction and training. These methods rely on mathematical approaches rather than machine learning or deep learning techniques [32,33]. Some studies in which computers examine bone structures by recognizing people's body postures in movies to identify violent acts are under the "skeleton-based" (deep learning or manually) category. These approaches explicitly concentrate on body movement and posture pattern to identify the presence of

suspicious activity, in contrast to other approaches that prioritize spatial feature extraction (e.g., CNNs) or changes in motion, texture, or illumination (e.g., manual techniques) [34, 35].

Transformer: This category includes research that uses transformer-based architectures to identify violent content in video [36, 37].

LSTM: Research that uses LSTM networks to record the temporal patterns connected to violent events in films falls under this category [38].

Methods on CNN + LSTM

A very quick and effective real-time violent detection system was created by Talha et al. [28] and tested on the creators' own devices. An LSTM processes the spatial information that are extracted by a CNN in the system. For categorization, the CNN also has two completely linked layers. Although performance numbers were not made public, Madhavan [29] presented a technique designed to address categorization challenges brought on by changing weather and illumination circumstances. This method could also aid in resolving issues related to video classification's restricted pixel allocation.

In their study, Ullah et al. [39] employed Mask R-CNN, an extension of Faster R-CNN, for feature frame selection. This object detection-based model was used to identify vehicles and people, with the added capability of recognizing and labeling object segments within images. For feature extraction, the researchers used two CNNs: DarkNet and another CNN that uses optical flow as a residual input. After that, a multilayer long short-term memory (M-LSTM) network was fed the retrieved characteristics.

Vijeikis et al. [40] created a computationally efficient and lightweight model by combining a CNN and LSTM, although it exhibited slightly lower accuracy. Halder and Chatterjee [41] achieved exceptional results in identifying violent activities using a bidirectional LSTM based on a lightweight convolutional neural network. In order to extract spatial characteristics from the INRA person dataset, Traoré and Akhloufi [42] used a pre-trained VGG-16 model. These features were then processed by a bidirectional gated recurrent unit (BiGRU), a kind of recurrent neural network (RNN). Three completely linked layers were used for the classification, with softmax activation being used in the last layer. Similarly, Ref. [43] used an LSTM to collect temporal characteristics and VGG-16 for spatial feature extraction.

Without using frame subtraction, Asad et al. [44] created a suspicious activity detection model by examining two consecutive video frames captured at intervals t and t + 1. Each frame's high-level and low-level characteristics were extracted by the model using two pre-trained CNNs. Wide dense remaining blocks (WDRBs) were used to combine feature maps from the two frames in order to improve feature representation. An LSTM network was then used to analyze these combined characteristics in order to identify temporal correlations. An actual time graph was created to alert people to continuing violent activities, and a scene was labelled as violent when its detected degree of suspicious activity exceeded a certain threshold. The pre-trained CNN MobileNetV2 was used by Contardo et al. [45] to extract spatial characteristics from video frames. Two distinct LSTM models a temporal Bi-LSTM and a temporal

ConvLSTM were then used to analyze these extracted characteristics in order to assess and contrast how well they detected suspicious activity. In order to evaluate the effectiveness of each method, Gupta and Ali [46] used a pre-trained VGG-16 network for extraction of features, processing its output independently by an LSTM and a Bi-LSTM.

Islam et al. [47] incorporated numerous parameters to characterize the datasets, including class count, videos per class, frame rate, video length, average frame count per video, resolution, and location count. Their focus was on detecting sexual and physical assault, utilizing pre-trained VGG-16 and VGG-19 networks with LSTMFrom each packet, frames were chosen at random by Jahlan and Elrefaei [48], who then chose a particular area inside the frame to transform the pictures into square forms. They calculated the differences between successive frames instead of putting individual frames straight into the CNN. To combine spatial components, they used a convolutional LSTM and a lightweight CNN called automated mobile neural architecture search (MNAS). After testing three classifiers and normalizing the training features prior to classification, SVM was found to be the most successful.

Mumtaz et al. [49] used an LSTM to receive the output of a previously trained VGG-19 CNN for feature extraction. This process management approach was combined with a new deep learning-based structure for suspicious activity detection, and control charts were added for the analysis of surveillance video data. An LSTM processed the spatial information extracted by Sharma et al. [50] using Xception, a pre-trained CNN. Singh et al. [51] divided the most advanced methods into three categories: deep learning, machine learning, and motion-based. They used two CNNs to extract low-level features and local motion, which were then sent to an LSTM to capture global temporal patterns. Body motions, object edges or lines that are and appearance-invariant traits like changes in lighting, weather, and other environmental factors were all included in the retrieved features.

Srivastava [52] proposed two distinct algorithms: a suspicious activity detection algorithm and a facial recognition system that could be beneficial in violent situations. The latter focused on identifying suspicious activity using drone cameras. An LSTM received the output of a CNN block that had been pre-trained on ImageNet for the purpose of extracting spatial features. Three combinations of chosen models and seven distinct algorithms were used in total. Traoré and Akhloufi used two EfficientNet CNNs, one trained with RGB input and the other with optical flow, both of which were pre-trained on ImageNet. After an LSTM processed the characteristics that were collected from these CNNs, a classifier that included a fully connected layer (FCL) with an activation function that was sigmoid was used.

A CNN and a separable convolutional LSTM (SepConvLSTM) were integrated into a dual-input framework by Islam et al. [53]. While one input examined frame differences between subsequent frames (i and i + 1), the other processed RGB video with background suppression. The categorization algorithms and the kinds of data that were sent to the fully linked layers varied across the three iterations of this design that were put into practice. Although optical flow pictures do not capture full scene details, Mugunga et al. [21] used them to capture spatio-temporal aspects and save processing expenses. After that, the collected features were input into a Bi-ConvLSTM, which outperformed unidirectional ConvLSTMs by learning both short- and long-term relationships.

Convolution Neural Networks (CNN)

Mahmoodi et al. [54] introduced an image segmentation technique (SSMI) to minimize computational expenses by limiting the number of frames processed by CNN. A single 3D-CNN architecture was utilized to extract spatio-temporal features, with fully connected layers for classification. To boost effectiveness, Ahmed et al. [55] deployed CNN-v4, which emphasized frame selection. Ji et al. [56] presented the Human Violence Dataset, comprising 1930 video clips featuring firearms violence and physical aggression. The two-stream CNN model classified violence levels (L1, L2, L3) by independently processing spatial and temporal data, employing a confusion matrix. Ehsan et al. [57] developed Vi-Net, a CNN that employed fully connected layers with softmax activation for classification. Jayasimhan et al. [58] suggested a 3D-CNN followed by a 2D-CNN to capture temporal and spatial information without transfer learning. Kim et al. [59] enhanced people monitoring and created a 3D-CNN for fall and violence detection. Monteiro and Durães [60] implemented a two-path ResNet model that combined motion and appearance features with global average pooling, using the AVA dataset.

A classifier based on C3D was presented by Talha et al. [61], who also divided the body of research into four categories: multimodal learning, supervised and unsupervised learning, knowledge distillation, and deep learning. Using GradCAM to improve interpretability, Appavu [62] implemented a keyframe selection strategy that included spatial, temporal, & spatio-temporal branches. Using a pre-trained CNN for the extraction of features, Adithya et al. [63] found that a five-layer 3D-CNN produced the best results. In order to minimize false positives, Bi et al. [64] used ResNet18 for feature extraction and Deeplab-V3plus for image segmentation.

Jain and Vishwakarma [65] introduced dynamic images to highlight salient motion by extracting features using ResNetV2. Liang et al. [66] implemented a SlowFast network with ResNet for feature extraction, using YOLOv5 for frame extraction and DeepSort for tracking, mapping violent events onto real-world coordinates. Mumtaz et al. [67] proposed Deep Multi-Net (DMN) for efficient suspicious activity detection by integrating pre-trained AlexNet and GoogleNet.

Que et al. [68] improved the accuracy of identifying start and end frames in long-duration videos for suspicious activity detection. Santos et al. [69] employed the X3D neural network, pre-trained with Kinetics-400. Sernani et al. [70] used the AIRTLab dataset to evaluate robustness and highlight the benefits of transfer learning. Shang et al. [71] introduced MAF-Net, a multimodal model integrating RGB, stream, and auditory features for suspicious activity recognition.

Magdy et al. [30] used optical flow methodologies to compare CNN-3D and CNN-4D, with the latter capturing complex spatio-temporal relationships. Hua et al. [72] implemented a residual attention module to enhance human pose estimation. Liu et al. [73] employed a multi-feature deep convolution network to maintain temporal consistency in human pose estimation across video frames.

Traditional Feature Extraction based Approaches

In their study, Wintarti et al. [32] selected 20 frames from each video at random for feature selection, utilizing PCA to reduce dimensionality and DWT for frequency analysis. They then employed an SVM to categorize the extracted features. Mohtavipour et al. [74] divided related research into deep learning and handcrafted approaches, implementing a CNN with three inputs: spatial (grayscale video), temporal (optical flow), and spatio-temporal (motion energy images). Lohithashva [33] developed an SVM classifier trained on features extracted using local orientation patterns (LOOP). For manual feature extraction, Jaiswal [75] used a fuzzy histogram of optical flow orientations in conjunction with local binary patterns (LBP). AdaBoost was then used to train these extracted features, and Ensemble RobustBoost aggregation was used for classification. Hu et al. [76] applied TOP-ALCM to extract co-occurrence patterns by segmenting video frames into X-T, Y-T, and X-Y planes. The extracted features were then classified using either an SVM or a CNN.

Zhou [34] examined skeleton-based methods that used 3D-CNNs like HRNetW32 to extract spatio-temporal features. These features were converted into patch vectors with position embeddings, which were subsequently input into the TokenPose Model and a transformer layer for classification. Saliency diagrams were used to illustrate the keypoint selection process. Hung et al. [35] grouped related works into four categories: They investigated multimodal learning strategies, supervised learning, unsupervised learning, and deep learning. An SVM was developed for classification after deep learning techniques were used to extract skeleton-related data including count, acceleration, and distance.

Naik and Gopalakrishna [77] categorized related research into three groups: optical-flow, space-time interest points, and CNN-based methods. They used DeepPose to determine body positions and trained an LSTM to capture temporal relationships. Narynov et al. [78] classified skeleton-based methods as either top-down or bottom-up. They used a pre-trained CNN (PoseNet) to extract and track skeletons, differentiating between striking and kicking actions. Srivastava et al. [79] implemented suspicious activity detection on drone-captured frames by using a CNN to derive keypoints and an SVM to classify suspicious activity into six features. Su et al. [80] introduced a skeleton-based approach that utilized geometric X, Y, and Z maps, with Z representing time. This method trained SPIL (Skeleton Point Interaction Learning) for classification by tracking head movements

Transformer-Based Approaches

Akti et al. [36] employed the Vision Transformer (ViT) algorithm, a neural network architecture combining self-attention with transformer-based vision models. This method divided images into segments, extracting features while preserving positional data. These areas were then analyzed for temporal relationships before classification. The researchers also introduced a dataset comprising images and recordings sourced from the internet.

Ehsan et al. [37] focused on extracting feature frames using YOLO to identify individuals in videos and remove background elements. They utilized STAT, a GAN-based algorithm, to transform temporal motion features into spatial image frames, and employed the Farneback method for optical flow calculations. The discriminator converted motion variation characteristics into visual representations and

assessed the veracity of the pictures produced by STAT's generator. While the discriminator attempted to distinguish between actual and artificial images during training, the generator aimed to create extremely realistic images. After training, the STAT generator was unable to reconstruct violent actions, enabling behavior classification based on discrepancies. However, it successfully translated normal motion features into normal images.

Kumar et al. [81] introduced an optimized transformer model for action recognition, inspired by recent advancements in video vision transformers. The model refined temporal relationships using tubelet embedding and captured spatial features from input frames. The processed frames underwent multiple transformer layers. Through efficient preprocessing, the study showed that transformer models can be effectively trained on smaller datasets, despite typically requiring large datasets. The authors also presented a new dataset for violence detection.

LSTM-Based Approaches

Ullah et al. [38] developed an algorithm for detecting suspicious activity in industrial settings. This method incorporated object detection for fragmentation. To address gradient vanishing, two recurrent neural networks were implemented: LSTM and GRU. The suggested suspicious activity detection system's cloud distribution architecture was described in full in the article.

Suspicious Activity Datasets

Given how seldom physical aggressiveness is in comparison to routine activities like sports or transportation, it might be difficult to detect suspicious activity on video footage. Even while more violent incidents are now available due to security cameras, they are still very rare. Developing datasets for model training that comprise both physical attacks and comparable non-violent occurrences is essential to resolving this problem. The physical antagonistic datasets used in certain investigations are described in this section, arranged according to how frequently they are utilized. The most commonly used dataset is the Hockey Fights dataset [82], which contains National Hockey League (NHL) game footage. Following this is the Action Movies dataset [82], featuring scenes from action films. The Violent Flow dataset [83], also known as the violent crowd dataset, consists of authentic YouTube videos depicting crowd violence. The Real World Fight-2000 (RWF-2000) dataset [11] showcases real-world conflicts captured in surveillance footage. The Real Life Violence Situations (RLVS) dataset [84] comprises violent scenarios extracted from YouTube. UCF-Crime Selected (UCFS) [85] includes untrimmed surveillance videos from 13 categories, such as assault, robbery, and gunshot.

The BEHAVE dataset [86] features extended video recordings categorized as either violent or non-violent. The Surveillance Camera dataset [87] contains genuine interior and outdoor footage from security cameras. The XD-Violence Selected (XD-V) dataset [88] encompasses various media types, including movies, CCTV, and hand-held cameras, covering a range of incidents like abuse, car accidents, explosions, rioting, and shootings. The AIRTLab dataset [70] consists of 180 clips recorded from two viewpoints under

natural lighting conditions. The Industrial Surveillance dataset [38] includes Google and YouTube recordings of violent activities in workplace settings such as stores and offices.

The Human Violence dataset [56] aggregates violent behaviors from YouTube film trailers and classifies them into four categories: blood, weapon possession, physical contact, and fighting. Drone surveillance footage taken from different angles and altitudes is included in the Target dataset [52]. The dataset in [77] comprises self-recorded videos documenting two violent actions (punching and kicking) performed by 20 individuals in different contexts. Srivastava et al.'s dataset [79] consists of drone images captured at heights of 2–18 meters, featuring actors aged 17 to 30. The HD dataset [79] is made up of high-definition video segments containing timestamps indicating violent incidents.

The Conflict Event dataset [89] classifies videos into 51 action groups. Adithya et al.'s dataset [63] focuses on drone-captured violence. Kumar et al. [81] have combined RLVS clips, YouTube videos, and GitHub content in their dataset, which includes non-violent actions like walking, exercising, and waving. VSD2015 [90] expands on the LIRIS-ACCEDE dataset, featuring action movie snippets. AVA-Kinetics [91] contains 80 human action categories, with 13 related to violence. MediaEval VSD-2014 [92] incorporates action movie sequences with gunfire, explosions, car chases, blood, and fights. Mahalle et al.'s dataset [93] transforms YouTube violence videos into audio, covering 31 aggression types such as gunshots and yelling. The Social Media Fight Images dataset [36] sources images and video frames from Google and social platforms. Naryanov et al.'s dataset [78] emphasizes real-world aggressive behavior from online and social network recordings. Rachna et al.'s dataset [94] comprises YouTube, stock footage, and self-recorded violent videos. The Violent Clip Dataset [71] includes 37 Hollywood films and 30 YouTube excerpts. Hung et al.'s dataset [35] features staged scenarios involving interactions between elderly individuals with mobility issues. These datasets are outlined in Table 1. The availability of diverse datasets from various sources aids in the study of suspicious activity detection, enabling the creation of models tailored to specific requirements.

Table 1. Overview of Dataset Characteristics in Selected Articles

Dataset	Reference	Year	No. of Clips	Frame rate	Resolution
BEHAVE	[86]	2010	4	25	640 x 480
Action Movies	[82]	2011	200	30	512 x 720
UCF-Crime Selected (UCFS)	[85]	2018	1900	25	-
AIRTLab	[70]	2020	350	30	1920 x 1080
Human Violence	[56]	2021	1930	30	1280 x 720
Violent Clip Dataset (VCD)	[71]	2022	7279	30	-
Target	[79]	2022	150	60	1080p

Conclusion

This research focuses on using artificial intelligence (AI) to detect physical assault in video footage, contributing significantly to the field through a comprehensive analysis of various algorithms and their combinations. The study emphasizes the importance of integrating spatial and temporal feature extraction methods, as well as the effectiveness of skeleton-based and audio-based approaches. It highlights the necessity of diverse and representative datasets for training robust models, addressing a common limitation in AI development. The research's multi-modal approach, combining computer vision, audio analysis, and skeleton-based methods, allows for a more comprehensive understanding of suspicious events. This holistic approach is essential for addressing the complexities of real-world scenarios and potentially leads to higher detection accuracy and fewer false positives. Key findings underscore the importance of integrating different AI techniques and data modalities, demonstrating the potential to substantially impact public safety and crime prevention efforts by enhancing real-time surveillance capabilities. The study also identifies important areas for future work, including addressing computational limitations for real-time implementation, expanding dataset diversity, and exploring novel combinations of AI techniques. By recognizing the need for representative data encompassing various types of physical assault across different environments and demographics, the research addresses potential biases and improves model generalizability. This focus on data quality and diversity is crucial for creating models that can effectively adapt to real-world situations. In conclusion, this research makes a valuable contribution to AI-based suspicious activity detection in video footage, laying a strong foundation for advancing public safety and crime prevention through comprehensive analysis, emphasis on dataset diversity, and identification of key areas for future research.

References

1. Muarifah, A.; Mashar, R.; Hashim, I.H.M.; Rofiah, N.H.; Oktaviani, F. Aggression in Adolescents: The Role of Mother-Child Attachment and Self-Esteem. Behav. Sci. 2022, 12, 147. [Google Scholar] [CrossRef] [PubMed]
2. Long, D.; Liu, L.; Xu, M.; Feng, J.; Chen, J.; He, L. Ambient population and surveillance cameras: The guardianship role in street robbers' crime location choice. Cities 2021, 115, 103223. [Google Scholar] [CrossRef]
3. Shubber, M.S.M.; Al-Ta'i, Z.T.M. A review on video violence detection approaches. Int. J. Nonlinear Anal. Appl. 2022, 13, 1117–1130. [Google Scholar]
4. Hillis, S.; Mercy, J.; Amobi, A.; Kress, H. Global prevalence of past-year violence against children: A systematic review and minimum estimates. Pediatrics 2016, 137, e20154079. [Google Scholar] [CrossRef] [PubMed]
5. Crime, Safety and Victims' Rights: Fundamental Rights Survey. 2021. Available online: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2021-crime-safety-victims-rights_en.pdf (accessed on 1 February 2024).
6. Vomfell, L.; Härdle, W.K.; Lessmann, S. Improving crime count forecasts using Twitter and taxi data. Decis. Support Syst. 2018, 113, 73–85. [Google Scholar] [CrossRef]

7. Jing, F.; Liu, L.; Zhou, S.; Song, J.; Wang, L.; Zhou, H.; Wang, Y.; Ma, R. Assessing the impact of street-view greenery on fear of neighborhood crime in Guangzhou, China. Int. J. Environ. Res. Public Health 2021, 18, 311. [Google Scholar] [CrossRef] [PubMed]
8. Yue, H.; Xie, H.; Liu, L.; Chen, J. Detecting people on the street and the streetscape physical environment from Baidu street view images and their effects on community-level street crime in a Chinese city. ISPRS Int. J. Geo-Inf. 2022, 11, 151. [Google Scholar] [CrossRef]
9. Hipp, J.R.; Lee, S.; Ki, D.; Kim, J.H. Measuring the built environment with google street view and machine learning: Consequences for crime on street segments. J. Quant. Criminol. 2021, 38, 537–565. [Google Scholar] [CrossRef]
10. Shukla, H.; Pandey, M. Human Suspicious Activity Recognition. Int. Innov. Res. J. Eng. Technol. 2020, 5. [Google Scholar] [CrossRef]
11. Cheng, M.; Cai, K.; Li, M. RWF-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4183–4190. [Google Scholar]
12. Jaafar, N.; Lachiri, Z. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. Expert Syst. Appl. 2023, 211, 118523. [Google Scholar] [CrossRef]
13. Zhou, W.; Min, X.; Zhao, Y.; Pang, Y.; Yi, J. A Multi-Scale Spatio-Temporal Network for Violence Behavior Detection. IEEE Trans. Biom. Behav. Identity Sci. 2023, 5, 266–276. [Google Scholar] [CrossRef]
14. Afra, S.; Alhajj, R. Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people. Phys. A Stat. Mech. Its Appl. 2020, 540, 123151. [Google Scholar] [CrossRef]
15. Vosta, S.; Yow, K.C. A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras. Appl. Sci. 2022, 12, 1021. [Google Scholar] [CrossRef]
16. Alonso, R.S.; Sittón-Candanedo, I.; Casado-Vara, R.; Prieto, J.; Corchado, J.M. Deep reinforcement learning for the management of software-defined networks and network function virtualization in an edge-IoT architecture. Sustainability 2020, 12, 5706. [Google Scholar] [CrossRef]
17. Ageed, Z.; Zeebaree, S. A Comprehensive Survey of Big Data Mining Approaches in Cloud Systems. Qubahan Acad. J. 2021, 1, 29–38. [Google Scholar] [CrossRef]
18. Ding, D.; Ma, Z.; Chen, D.; Chen, Q.; Liu, Z.; Zhu, F. Advances in video compression system using deep neural network: A review and case studies. Proc. IEEE 2021, 109, 1494–1520. [Google Scholar] [CrossRef]
19. Kostka, G.; Steinacker, L.; Meckel, M. Between Privacy and Convenience: Facial Recognition Technology in the Eyes of Citizens in China, Germany, the UK and the US (10 February 2020). Available online: https://ssrn.com/abstract=3518857 (accessed on 1 February 2024).
20. Ali, O.; Shrestha, A.; Soar, J.; Wamba, S.F. Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. Int. J. Inf. Manag. 2018, 43, 146–158. [Google Scholar] [CrossRef]
21. Mugunga, I.; Dong, J.; Rigall, E.; Guo, S.; Madessa, A.H.; Nawaz, H.S. A frame-based feature model for violence detection from surveillance cameras using ConvLSTM network. In Proceedings of the

2021 6th International Conference on Image, Vision and Computing (ICIVC), Qingdao, China, 23–25 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 55–60. [Google Scholar]
22. Negre, P.; Alonso, R.S.; Prieto, J.; Arrieta, A.G.; Corchado, J.M. Review of Physical Aggression Detection Techniques in Video Using Explainable Artificial Intelligence. In Proceedings of the International Symposium on Ambient Intelligence; Springer: Cham, Germany, 2023; pp. 53–62. [Google Scholar]
23. Siddique, M.; Islam, M.S.; Sinthy, R.; Mohima, K.; Kabir, M.; Jibon, A.H.; Biswas, M. State-of-the-Art Violence Detection Techniques: A review. Asian J. Res. Comput. Sci. 2022, 13, 29–42. [Google Scholar]
24. Negre, P.; Alonso, R.S.; Prieto, J.; Dang, C.N.; Corchado, J.M. Systematic Mapping Study on Violence Detection in Video by Means of Trustworthy Artificial Intelligence. 2024. Available online: https://ssrn.com/abstract=4757631 (accessed on 1 February 2024).
25. Yao, H.; Hu, X. A survey of video violence detection. Cyber-Phys. Syst. 2023, 9, 1–24. [Google Scholar] [CrossRef]
26. Kaur, G.; Singh, S. Violence detection in videos using deep learning: A survey. In Advances in Information Communication Technology and Computing: Proceedings of AICTC 2021; Springer: Singapore, 2022; pp. 165–173. [Google Scholar]
27. Omarov, B.; Narynov, S.; Zhumanov, Z.; Gumar, A.; Khassanova, M. State-of-the-art violence detection techniques in video surveillance security systems: A systematic review. PeerJ Comput. Sci. 2022, 8, e920. [Google Scholar] [CrossRef] [PubMed]
28. Talha, K.R.; Bandapadya, K.; Khan, M.M. Violence Detection Using Computer Vision Approaches. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 544–550. [Google Scholar]
29. Madhavan, R.; Utkarsh; Vidhya, J. Violence Detection from CCTV Footage Using Optical Flow and Deep Learning in Inconsistent Weather and Lighting Conditions. In Proceedings of the Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik, India, 23–24 April 2021; Revised Selected Papers, Part I 5. Springer: Berlin/Heidelberg, Germany, 2021; pp. 638–647. [Google Scholar]
30. Magdy, M.; Fakhr, M.W.; Maghraby, F.A. Violence 4D: Violence detection in surveillance using 4D convolutional neural networks. IET Computer Vision 2023, 17, 282–294. [Google Scholar] [CrossRef]
31. Chen, Y.; Zhang, B.; Liu, Y. ESTN: Exacter Spatiotemporal Networks for Violent Action Recognition. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 44–48. [Google Scholar]
32. Wintarti, A.; Puspitasari, R.D.I.; Imah, E.M. Violent Videos Classification Using Wavelet and Support Vector Machine. In Proceedings of the 2022 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 10–11 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 01–05. [Google Scholar]
33. Lohithashva, B.; Aradhya, V.M. Violent video event detection: A local optimal oriented pattern based approach. In Proceedings of the Applied Intelligence and Informatics: First International

Conference, AII 2021, Nottingham, UK, 30–31 July 2021; Proceedings 1. Springer: Berlin/Heidelberg, Germany, 2021; pp. 268–280. [Google Scholar]
34. Zhou, L. End-to-end video violence detection with transformer. In Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19–21 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 880–884. [Google Scholar]
35. Hung, L.P.; Yang, C.W.; Lee, L.H.; Chen, C.L. Constructing a Violence Recognition Technique for Elderly Patients with Lower Limb Disability. In Proceedings of the International Conference on Smart Grid and Internet of Things; Springer: Cham, Germany, 2021; pp. 24–37. [Google Scholar]
36. Aktı, Ş.; Ofli, F.; Imran, M.; Ekenel, H.K. Fight detection from still images in the wild. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 550–559. [Google Scholar]
37. Ehsan, T.Z.; Nahvi, M.; Mohtavipour, S.M. An accurate violence detection framework using unsupervised spatial–temporal action translation network. Vis. Comput. 2023, 40, 1515–1535. [Google Scholar] [CrossRef]
38. Ullah, F.U.M.; Muhammad, K.; Haq, I.U.; Khan, N.; Heidari, A.A.; Baik, S.W.; de Albuquerque, V.H.C. AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. IEEE Trans. Ind. Inform. 2021, 18, 5359–5370. [Google Scholar] [CrossRef]
39. Ullah, F.U.M.; Obaidat, M.S.; Muhammad, K.; Ullah, A.; Baik, S.W.; Cuzzolin, F.; Rodrigues, J.J.; de Albuquerque, V.H.C. An intelligent system for complex violence pattern analysis and detection. Int. J. Intell. Syst. 2022, 37, 10400–10422. [Google Scholar] [CrossRef]
40. Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient violence detection in surveillance. Sensors 2022, 22, 2216. [Google Scholar] [CrossRef]
41. Halder, R.; Chatterjee, R. CNN-BiLSTM model for violence detection in smart surveillance. SN Comput. Sci. 2020, 1, 201. [Google Scholar] [CrossRef]
42. Traoré, A.; Akhloufi, M.A. 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In Proceedings of the International Conference on Image Analysis and Recognition, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 152–160. [Google Scholar]
43. Aarthy, K.; Nithya, A.A. Crowd Violence Detection in Videos Using Deep Learning Architecture. In Proceedings of the 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 16–17 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [Google Scholar]
44. Asad, M.; Yang, J.; He, J.; Shamsolmoali, P.; He, X. Multi-frame feature-fusion-based model for violence detection. Vis. Comput. 2021, 37, 1415–1431. [Google Scholar] [CrossRef]
45. Contardo, P.; Tomassini, S.; Falcionelli, N.; Dragoni, A.F.; Sernani, P. Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In Proceedings of the RTA-CSIT 2023: 5th International Conference Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, 26–27 April 2023. [Google Scholar]
46. Gupta, H.; Ali, S.T. Violence Detection using Deep Learning Techniques. In Proceedings of the 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 25–27 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 121–124. [Google Scholar]

47. Islam, M.S.; Hasan, M.M.; Abdullah, S.; Akbar, J.U.M.; Arafat, N.; Murad, S.A. A deep Spatio-temporal network for vision-based sexual harassment detection. In Proceedings of the 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE), Dhaka, Bangladesh, 21–23 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6. [Google Scholar]
48. Jahlan, H.M.B.; Elrefaei, L.A. Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video. Arab. J. Sci. Eng. 2021, 46, 8549–8563. [Google Scholar] [CrossRef]
49. Mumtaz, N.; Ejaz, N.; Aladhadh, S.; Habib, S.; Lee, M.Y. Deep multi-scale features fusion for effective violence detection and control charts visualization. Sensors 2022, 22, 9383. [Google Scholar] [CrossRef]
50. Sharma, S.; Sudharsan, B.; Naraharisetti, S.; Trehan, V.; Jayavel, K. A fully integrated violence detection system using CNN and LSTM. Int. J. Electr. Comput. Eng. (2088-8708) 2021, 11, 3374–3380. [Google Scholar] [CrossRef]
51. Singh, N.; Prasad, O.; Sujithra, T. Deep Learning-Based Violence Detection from Videos. In Proceedings of the Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021), Mizoram, India, 25–26 June 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 323–332. [Google Scholar]
52. Srivastava, A.; Badal, T.; Saxena, P.; Vidyarthi, A.; Singh, R. UAV surveillance for violence detection and individual identification. Autom. Softw. Eng. 2022, 29, 28. [Google Scholar] [CrossRef]
53. Islam, Z.; Rukonuzzaman, M.; Ahmed, R.; Kabir, M.H.; Farazi, M. Efficient two-stream network for violence detection using separable convolutional lstm. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8. [Google Scholar]
54. Mahmoodi, J.; Nezamabadi-pour, H.; Abbasi-Moghadam, D. Violence detection in videos using interest frame extraction and 3D convolutional neural network. Multimed. Tools Appl. 2022, 81, 20945–20961. [Google Scholar] [CrossRef]
55. Ahmed, M.; Ramzan, M.; Khan, H.U.; Iqbal, S.; Khan, M.A.; Choi, J.I.; Nam, Y.; Kadry, S. Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning; Tech Science Press: Henderson, NV, USA, 2021. [Google Scholar]
56. Ji, Y.; Wang, Y.; Kato, J.; Mori, K. Predicting Violence Rating Based on Pairwise Comparison. IEICE Trans. Inf. Syst. 2020, 103, 2578–2589. [Google Scholar] [CrossRef]
57. Ehsan, T.Z.; Mohtavipour, S.M. Vi-Net: A deep violent flow network for violence detection in video sequences. In Proceedings of the 2020 11th International Conference on Information and Knowledge Technology (IKT), Tehran, Iran, 22–23 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 88–92. [Google Scholar]
58. Jayasimhan, A.; Pabitha, P. A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset. In Proceedings of the 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 15–16 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5. [Google Scholar]
59. Kim, H.; Jeon, H.; Kim, D.; Kim, J. Lightweight framework for the violence and falling-down event occurrence detection for surveillance videos. In Proceedings of the 2022 13th International

Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 19–21 October 2022; IEEE: Piscataway, NJ, USA; 2022; pp. 1629–1634. [Google Scholar]
60. Monteiro, C.; Durães, D. Modelling a Framework to Obtain Violence Detection with Spatial-Temporal Action Localization. In Proceedings of the World Conference on Information Systems and Technologies, Galicia, Spain, 16–19 April 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 630–639. [Google Scholar]
61. Zhang, Z.; Yuan, D.; Li, X.; Su, S. Violent Target Detection Based on Improved YOLO Network. In Proceedings of the International Conference on Artificial Intelligence and Security, Qinghai, China, 15–20 July 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 480–492. [Google Scholar]
62. Appavu, N. Violence Detection Based on Multisource Deep CNN with Handcraft Features. In Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, 29 April–1 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [Google Scholar]
63. Adithya, H.; Lekhashree, H.; Raghuram, S. Violence Detection in Drone Surveillance Videos. In Proceedings of the International Conference on Smart Computing and Communication, Nashville, TN, USA, 26–30 June 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 703–713. [Google Scholar]
64. Bi, Y.; Li, D.; Luo, Y. Combining keyframes and image classification for violent behavior recognition. Appl. Sci. 2022, 12, 8014. [Google Scholar] [CrossRef]
65. Jain, A.; Vishwakarma, D.K. Deep NeuralNet for violence detection using motion features from dynamic images. In Proceedings of the 2020 third international conference on smart systems and inventive technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 826–831. [Google Scholar]
66. Liang, Q.; Cheng, C.; Li, Y.; Yang, K.; Chen, B. Fusion and visualization design of violence detection and geographic video. In Proceedings of the Theoretical Computer Science: 39th National Conference of Theoretical Computer Science, NCTCS 2021, Yinchuan, China, 23–25 July 2021; Revised Selected Papers 39. Springer: Berlin/Heidelberg, Germany, 2021; pp. 33–46. [Google Scholar]
67. Mumtaz, A.; Bux Sargano, A.; Habib, Z. Fast learning through deep multi-net CNN model for violence recognition in video surveillance. Comput. J. 2022, 65, 457–472. [Google Scholar] [CrossRef]
68. Qu, W.; Zhu, T.; Liu, J.; Li, J. A time sequence location method of long video violence based on improved C3D network. J. Supercomput. 2022, 78, 19545–19565. [Google Scholar] [CrossRef]
69. Santos, F.; Durães, D.; Marcondes, F.S.; Lange, S.; Machado, J.; Novais, P. Efficient violence detection using transfer learning. In Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems, Salamanca, Spain, 6–8 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 65–75. [Google Scholar]
70. Mori, P.; Falcionelli, N.; Tomassini, S.; Contardo, P.; Dragoni, A.F. Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access 2021, 9, 160580–160595. [Google Scholar] [CrossRef]

71. Shang, Y.; Wu, X.; Liu, R. Multimodal Violent Video Recognition Based on Mutual Distillation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Shenzhen, China, 4–7 November 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 623–637. [Google Scholar]
72. Hua, G.; Li, L.; Liu, S. Multipath affinage stacked—Hourglass networks for human pose estimation. Front. Comput. Sci. 2020, 14, 1–12. [Google Scholar] [CrossRef]
73. Liu, S.; Li, Y.; Hua, G. Human pose estimation in video via structured space learning and halfway temporal evaluation. IEEE Trans. Circuits Syst. Video Technol. 2018, 29, 2029–2038. [Google Scholar] [CrossRef]
74. Mohtavipour, S.M.; Saeidi, M.; Arabsorkhi, A. A multi-stream CNN for deep violence detection in video sequences using handcrafted features. Vis. Comput. 2022, 38, 2057–2072. [Google Scholar] [CrossRef]
75. Jaiswal, S.G.; Mohod, S.W. Classification Of Violent Videos Using Ensemble Boosting Machine Learning Approach With Low Level Features. Indian J. Comput. Sci. Eng. 2021, 12, 1789–1802. [Google Scholar] [CrossRef]
76. Hu, X.; Fan, Z.; Jiang, L.; Xu, J.; Li, G.; Chen, W.; Zeng, X.; Yang, G.; Zhang, D. TOP-ALCM: A novel video analysis method for violence detection in crowded scenes. Inf. Sci. 2022, 606, 313–327. [Google Scholar] [CrossRef]
77. Naik, A.J.; Gopalakrishna, M. Deep-violence: Individual person violent activity detection in video. Multimed. Tools Appl. 2021, 80, 18365–18380. [Google Scholar] [CrossRef]
78. Narynov, S.; Zhumanov, Z.; Gumar, A.; Khassanova, M.; Omarov, B. Detecting School Violence Using Artificial Intelligence to Interpret Surveillance Video Sequences. In Proceedings of the Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, 29 September–1 October 2021; Proceedings 13. Springer: Berlin/Heidelberg, Germany, 2021; pp. 401–412. [Google Scholar]
79. Srivastava, A.; Badal, T.; Garg, A.; Vidyarthi, A.; Singh, R. Recognizing human violent action using drone surveillance within real-time proximity. J. Real-Time Image Process. 2021, 18, 1851–1863. [Google Scholar] [CrossRef]
80. Su, Y.; Lin, G.; Zhu, J.; Wu, Q. Human interaction learning on 3d skeleton point clouds for video violence recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 74–90. [Google Scholar]
81. Kumar, A.; Shetty, A.; Sagar, A.; Charushree, A.; Kanwal, P. Indoor Violence Detection using Lightweight Transformer Model. In Proceedings of the 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 26–28 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [Google Scholar]
82. Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R. Violence detection in video using computer vision techniques. In Proceedings of the Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, 29–31 August 2011; Proceedings, Part II 14. Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339. [Google Scholar]

83. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–6. [Google Scholar]
84. Soliman, M.M.; Kamal, M.H.; El-Massih Nashed, M.A.; Mostafa, Y.M.; Chawky, B.S.; Khattab, D. Violence Recognition from Videos using Deep Learning Techniques. In Proceedings of the 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 8–10 December 2019; pp. 80–85. [Google Scholar]
85. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488. [Google Scholar]
86. Blunsden, S.; Fisher, R. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. Ann. BMVA 2010, 4, 4. [Google Scholar]
87. Aktı, Ş.; Tataroğlu, G.A.; Ekenel, H.K. Vision-based fight detection from surveillance cameras. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6. [Google Scholar]
88. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339. [Google Scholar]
89. Cheng, S.T.; Hsu, C.W.; Horng, G.J.; Jiang, C.R. Video reasoning for conflict events through feature extraction. J. Supercomput. 2021, 77, 6435–6455. [Google Scholar] [CrossRef]
90. Sjöberg, M.; Baveye, Y.; Wang, H.; Quang, V.L.; Ionescu, B.; Dellandréa, E.; Schedl, M.; Demarty, C.H.; Chen, L. The MediaEval 2015 Affective Impact of Movies Task. In Proceedings of the MediaEval, Wurzen, Germany, 14–15 September 2015; Volume 1436. [Google Scholar]
91. Li, A.; Thotakuri, M.; Ross, D.A.; Carreira, J.; Vostrikov, A.; Zisserman, A. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv 2020, arXiv:2005.00214. [Google Scholar]
92. Demarty, C.H.; Penet, C.; Soleymani, M.; Gravier, G. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. Multimed. Tools Appl. 2015, 74, 7379–7404. [Google Scholar] [CrossRef]
93. Mahalle, M.D.; Rojatkar, D.V. Audio based violent scene detection using extreme learning machine algorithm. In Proceedings of the 2021 6th international conference for convergence in technology (I2CT), Maharashtra, India, 2–4 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8. [Google Scholar]
94. Rachna, U.; Guruprasad, V.; Shindhe, S.D.; Omkar, S. Real-Time Violence Detection Using Deep Neural Networks and DTW. In Proceedings of the International Conference on Computer Vision and Image Processing; Springer: Cham, Switzerland, 2022; pp. 316–327. [Google Scholar]