Abstract :

Time-to-real detection of potential threats has formed the base of maintaining public safety through violence detection systems. The present study dwells on deep learning models that offer improved prediction accuracy for human activity analysis with the changed scope of their AI use.
Recent studies use a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs such as LSTM) in many settings to classify violence in videos obtained from surveillance footage, attaining a substantial drop in the rate of classification errors. This research constitutes a hybrid deep learning framework that combines CNNs for extracting features and LSTMs for analyzing violent and non-violent scenes. It was also examined how well the GRUs identified and classified the scenes along with fully connected networks in parallel so as to increase the robustness of the model. CNNs extract spatial features, LSTMs and GRUs capture temporal dynamics. The experiment reaches 92% test accuracy, accomplishing an impressive advance in differentiating violent onset from benign events, achieving superior precision and recall values compared with state-of-the-art benchmarks. This finding delineates the contribution of deep learning in respect of the effective enhancement of automated surveillance, paving the path for safer environments through proactive presence of threat detection.

Introduction :

Physical violence affects almost everyone's life, not only them but also their families and beloveds. It also destabilizes their economy and the mental health because of the insecurities caused by violence. Some studies reveal shocking results, for example, physical violence is the major cause of mortality among peoples around the world. The study and research by Hillis et al.in 2016, which has been published in the American Academy of Pediatrics, shows that more than half of the children in continents like Asia, Africa, and North America have experienced violence assaults in the year 2015, and world wide, more than a half of all children below the age of 17 experienced such violence. According to a research/study by the European Union Agency for Fundamental Rights, every 1 out of 4 Europeans was a victim of violence and around 22 million were physically assaulted in 1 year. For these kinds of reasons, it’s important to address the issue of physical violence in societies.

Multiple long-term, medium-term, and short-term solutions have been proposed. Those for the long term consist of understanding reasons or situations that exacerbate violence, so that future generations can be educated to avoid them. Some studies have studied an increase in violence due to exposure to an environment of violence and aggressiveness, which leads to a tendency to use violence as a tool. It has also been shown that there is a direct correlation between aggression and a poor maternal relationship or low self-esteem. Moreover,

technologies, such as smartphones, contribute to an increase in violence, as does the consumption of aggressive video games.

Other studies have opted for medium-term solutions to face violence by trying to relate street population density and the urban landscape to crime rates. Mobile phone and social network data have been used to measure population density in urban areas and to relate it to the crime rate using statistical techniques focused on spatial analysis, although the fact that it is not possible to differentiate between indoor and outdoor populations is a drawback. Moreover, through the use of convolutional neural networks (CNNs), it has been discovered that crime rates are lower in green areas. Another study employed deep learning to assess images from Baidu Street View (the namesake of Google Street View in China) showing the number of people on the streets, as well as the type of urban landscape (buildings, green areas, etc.) and correlated them with the number of crimes. Another study also used Google Street View images to analyze the number of vehicles and the pavement, and related those features to the crime rate through the use of machine learning techniques.

The subject of this review is that the short-lived strategies and rapid prevention of physical assault are the ones that operate in real time.

As of yet, to provide post-crime evidence that is also used as an evidence to identify the offenders and used by insurance companies or police, images and security cameras have primarily been utilized. Although a considerable number of published papers are already available that have implemented real-time violence detection solutions.

These approaches are based on the use of security video cameras led by artificial intelligence (AI) algorithms.

It is a growing field whose development has been possible because of the growth of three main pillars: the rising adoption of images and security cameras, the technological evolution of big data platforms, and the development of artificial intelligence algorithms that allow for the evaluation of images and video.

The use of security cameras has expanded around the world for surveillance and safety purposes, among others. It is important to bring attention to that while the images and videos have a wide range of uses that can enhance the quality of life of citizens, there are already studies that confront the threat to personal privacy posed by large-scale recording.

Currently, big data platforms deliver us the ability to obtain and manage large-scale data collected from the environment that surrounds us.These tools enable us to record, store, and process, in real time, various forms of data, such as images and videos.

Finally, artificial-intelligence-based algorithms for image and video analysis are increasingly widespread in recent times; they have also advanced in diversity and accuracy.

The recognition of violent acts in videos falls within the area of computer vision, particularly in the field of action recognition. Computer vision is a branch of artificial intelligence that facilitates computers to understand and derive decisions from visual data. Action recognition is an area that aims at detecting specific actions across video sequences.

Applying AI to detect violence in videos requires training models to identify patterns and activities reflective of violent behavior.

AI-driven violence identification in videos operates by training algorithms using labeled video datasets, where violent and non-violent actions are identified and labeled. These algorithms are trained to identify patterns and features linked to violent actions , enabling them to detect such actions in new, unseen videos.

Overall, physical aggression is a critical social issue with extensive impact. Many research projects have been implemented to solve this concern through various approaches, each providing solutions with varying levels of approach. The real-time recognition of violent acts stands out as the fastest and crucial solution, acting as the definitive protective measure in detecting individuals affected by physical harm. This achievement owes its practicality to a number of critical factors: the growing reliance on security footage and videos, the enhancement of large data platforms and the advancement of algorithms to be able to analyze images and videos supported by artificial intelligence.

The goal of this research is to further develop a systematic mapping study with the aim of presenting a complete and timely overview of AI-powered video violence detection, targeting physical assault as a primary focus. The key contribution of this work lies in the formation of an extensive and recent review, addressing all the stages involved in video violence detection using artificial intelligence. Lastly, a factor that differentiates this paper is its evaluation of different algorithm types used in video violence detection, the set of algorithms employed together, and the outcomes achieved for the most frequently used datasets in the state-of-the-art literature.

Literature Review :

Various Literature works are taken under considerations and studies in order to create an algorithm that is effective and robust for violence detection under various conditions. These literatures are some that try to achieve similar goals and try different algorithms and models that are available and add some new approaches on top of that. They are as following

1. Talha et al. [35] proposed a highly green and fast actual-time violence detection gadget examined on the personal devices of its builders. The system is predicated on a CNN for extracting spatial functions, which might be in the end fed into an LSTM. additionally, the CNN incorporates completely connected layers for classification functions. Madhavan [36] brought a version

geared toward overcoming demanding situations related to class in varying weather and lighting fixtures conditions, though overall performance metrics have not been provided. This version addressed problems like dedicating minimum pixels for video type. Ullah et al. [48] employed masks R-CNN, an extension of faster R-CNN, to choose characteristic frames through detecting and labeling objects such as humans and motors within photographs. For function extraction, they applied DarkNet and any other CNN that integrated optical float as residual input. These features were processed using a multilayer lengthy short-term reminiscence (M-LSTM) community.

2. Vijeikis et al. [49] evolved a lightweight and fast version combining CNN and LSTM, although it established slightly lower accuracy. Halder and Chatterjee [50] applied a lightweight CNN-based totally bidirectional LSTM to successfully identify violent activities, reaching notable effects. Traoré and Akhloufi [51] applied a pre-trained VGG-sixteen with the INRA character dataset for spatial characteristic extraction. those features were fed right into a bidirectional gated recurrent unit (BiGRU), accompanied by means of 3 absolutely linked layers, with the final layer employing softmax activation. In addition, Ref. [52] also used VGG-sixteen for spatial function extraction and an LSTM for temporal feature evaluation.

3. Asad et al. [53] presented a violence detection version that processed consecutive video frames at instances t and t+1 as inputs, in preference to their distinction. two pre-skilled CNNs have been used for excessive- and occasional-degree feature extraction from these frames. extensive dense residual blocks (WDRBs) then blended these features, and an LSTM captured temporal styles among them. The gadget also plotted an actual-time graph to signify violence tiers, raising an alarm while the output handed a threshold. Contardo et al. [54] employed MobileNetV2, a CNN pre-skilled on ImageNet, to extract spatial features. These outputs had been fed into two kinds of LSTMs—temporal Bi-LSTM and temporal ConvLSTM—to evaluate their relative performance.

4. Gupta and Ali [55] examined each LSTM and Bi-LSTM architectures on functions extracted from the use of a pre-educated VGG-16 network to decide the higher-performing model. Islam et al. [56] focused on detecting sexual and physical assaults while providing designated parameters of the datasets, which includes elegance remember, video depend, frames in step with 2d, video duration, average frames per video, decision, and vicinity statistics. Their model leveraged pre-educated VGG-sixteen and VGG-19 networks, with outputs fed into LSTMs. Jahlan and Elrefaei [57] used random body selection from each packet and transformed the images into squares via

cropping unique regions. In preference to inputting character frames, they processed the difference between consecutive frames (i and that i+1). They followed MNAS (cellular Neural architecture seek), a lightweight CNN, paired with convolutional LSTM for temporal and spatial characteristic extraction. The extracted capabilities were normalized between zero and 1 earlier than being categorized by the usage of 3 exclusive classifiers, with SVM turning in the high-quality outcomes.

5. [Jan 2019]Dinesh Jackson Samuel R, Fenil E, Gunasekaran Manogaran , Vivekananda G.N, Thanjaivadivel T, ,Jeeva S , Ahilan A worked on this paper, a real time violence detection system is proposed which deals with the very huge input stream data and recognises the violence using human intelligence simulation. The input to the system is the enormous amount of real time video streams from different sources which are processed in Spark framework. In the Spark framework, the frames are separated and the features of individual frames are extracted by using the HOG (Histogram of Oriented Gradients) function. Then the frames are tagged according to features such as violence model, human part model and negative model that are applied during training of the BD-LSTM network to recognize scenes of violence. Bidirectional LSTM can have access to the information in both forward and reverse direction. Thus the output is generated in context to information both pertaining past and future. The network was trained with a violent interaction dataset, which included 2314 videos with 1077 fight ones and 1237 no-fight ones. We have also built a dataset by collecting 410 video clips that contain non-violence scenes and 409 video clips that include violence scenes, gathered from the football stadium.

6. Srivastava [61] suggested two algorithms, which are described below. The first was the violence detection algorithm. The second was a facial identification algorithm useful in case of violence. This algorithm emphasized the detection of violence using drone cameras. Spatial features were extracted using a CNN block that used pre-trained architectures of ImageNet, whose output is passed to an LSTM. In total, seven different algorithms were employed, together with three combinations of some of them. Traoré and Akhloufi have used two CNNs they named EfficientNet, respectively pre-trained on ImageNet. One of these had optical flow, and the second had RGB. The outputs of both CNNs then were fed into an LSTM, and finally, to a classifier that includes an FCL layer with a sigmoid activation layer.

7. Mahmoodi et al. [63] applied the SSMI image segmentation technique in their work to eliminate the need for all video frames to be fed to the CNN. This was done with an application of the temporal-spatial-focused feature extractor using single 3D-CNN architecture, which then used fully connected layers as the classifier. Ahmed et al. [64] applied a CNN-v4 and pointed out that unlike other state-of-the-art works using uncompressed frames of the video, this

work only focused on using selected characteristic frames-the reason being that CNNs applied to the entire video are computationally too demanding. Ji et al. [65] presented the Human Violence Dataset-a new dataset comprising 1930 clips from movie trailers on YouTube. It took into account not only physical aggression but also gun violence. Two-stream CNN models are a specific neural network architecture working with two independent streams of visual information: spatial and temporal streams. The temporal stream takes a single image input while the spatial stream takes 10 optical flow frames. After feature extraction with the CNN, a machine-learning procedure trains the classifier optimizing the weights to quantify the level of violence in the videos. This article quantified violence levels in the videos with relevance to a ranking score. Due to this, three levels of violence (L1, L2, and L3) were calculated with the help of a confusion matrix.

8. Baba et al. [19] introduced a lightweight CNN model such as MobileNet and SqueezeNet model on two publicly available datasets, namely the BEHAVE and ARENA datasets. For the purpose of optimizing the performance model, a time domain filter was utilized to differentiate between the violent scenes of the video within a certain period. The overall experimental result was able to achieve an accuracy of 77.9% with no misclassification of the violent class and a high false-positive rate of 26.76% of the nonfight clips. According to the authors, one of the shortcomings of the proposed method is the inability of the model to detect violence in the crowd.

9. CNN and LSTM combination for violence discovery in videotape features was created lately. CNN is used for point birth at the frame position. Collected features are classified as violent or peaceful by exercising a variant of LSTM in the workshop by Sudhakaran and Lanz[22], Soliman et al.[23], Letchmunan et al.[24], and Sumon et al.[25]. By combining CNN and LSTM, the videotape's spatiotemporal features are localized, allowing for original stir analysis. Several pre-trained CNN models, videlicet VGG16 by Soliman et al. [23], Sumon et al.[25], VGG19 by Letchmunan et al.[24], Sumon et al.[25] and ResNet50 by Sumon et al.[25] were used to separate the spatial features, before their bracket as violent or peaceful events/occasions.

10. Hanson et al.[29] proposed a model separated into three corridor spatial encoders, temporal encoders, and classifiers. The authors formerly used ConvLTSM infrastructures supplemented with bidirectional temporal encodings. The model presented is called a one-sluice model, as it uses only one format for input; nonetheless, some workshops presented models that use both formats for their input. These models are called convolutional multistream models, where the type of each videotape sluice is anatomized as well.

11. Shakil Ahmed Sumon, Tanzil Shahria, Niyaz Bin Hashem, Raihan Goni, and Rashedur M. Rahman have investigated distinctive techniques to discover the saliency of the highlights from different pre-trained models in detecting violence in videos. A dataset was created that compared violent and non-violent videos of different settings. Three ImageNet models; VGG16, VGG19, and ResNet50 are being used to prize features from the frames of the vids. In one of the trials, the pulled features have been fed into a completely connected network that detects violence in frame position. Moreover, in another trial, we nourished the pulled highlights of 30 outlines to a long short-term memory(LSTM) organized at a time. likewise, we've applied attention to the features pulled from the frames through spatial motor networks which also enable metamorphoses like gyration, translation, and scale. Along with these models, we've designed a custom convolutional neural network(CNN) as a point extractor and a pre-trained model which is originally trained on a movie violence dataset. In the end, the features pulled from the ResNet50 pre-trained model proved to be more salient in detecting violence. These ResNet50 features, in combination with LSTM, give a delicacy of 97.06 which is better than the other models we've experimented with.

12. Rohit Halder and Rajdeep Chatterjee introduced a featherlight computational model for the better bracket of violent and non-violent conditioning. Deep literacy grounded in an effective violent exertion discovery model can help the authorities in detecting violent exertion in real time. The estimated results can be hereafter transferred to store and dissect the captured videotape to automate the crime monitoring system. Convolutional Neural Network-grounded Bidirectional LSTM has been used to describe violent conditioning and also compared with other approaches. Our proposed model gives 99.27, 100, and 98.64 bracket rigor.

13. Hua et al. [84] proposed enhancing human pose estimation by incorporating a residual attention module into the stacked hourglass network, addressing resolution loss in initial images. Their architecture improved both resolution and accuracy, leading to more precise pose estimation in images with complex backgrounds. Liu et al. [85] introduced a novel approach for maintaining temporal consistency in human pose estimation across video frames. Their method utilized structured-space learning and halfway temporal evaluation within a three-stage multi-feature deep convolutional network, ensuring stable and accurate long-term pose estimation.

14. Magdy et al. [37] divided video data into frame packets and applied an optical flow-based technique to detect motion areas. Their study compared CNN-3D and CNN-4D architectures, concluding that CNN-3D effectively analyzed short time spans, while CNN-4D provided a more advanced understanding of

complex spatio-temporal relationships. The CNN models were pre-trained using ImageNet.

15. Santos et al. [81] asserted that 3D-CNNs surpass conventional CNNs in extracting both temporal and spatial information. Their work leveraged a pre-trained X3D neural network using the Kinetics-400 dataset. Sernani et al. [82] underscored the importance of robustness in violence detection and introduced the AIRTLab dataset, specifically designed to assess algorithm resilience. Their study presented three different deep learning architectures, where pre-trained 3D-CNNs, trained on the Sports-1M dataset, demonstrated that transfer learning enhances efficiency compared to training models from scratch.

16. Que et al. [80] focused on detecting violence in long-duration videos using pre-trained CNNs. Their study emphasized accurately determining the start and end of violent episodes, an area often overlooked in prior research. The second phase of their approach employed a deconvolution technique to pinpoint specific video segments at the frame level, precisely identifying when violent events occurred. The study also highlighted the need for improved preprocessing and detection methods for marking the beginning and end of violent activities.