Artificial Intelligence based Automatic Violence Detector in CCTV Footage

RUBA SOUNDAR K, MELWIN PRABHU R
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, INDIA

Abstract: In this modern society, CCTVs are everywhere to monitor employees, students, and public citizens in malls, hotels, theaters, etc.; Analyzing CCTV videos is mandatory for public security. Many places have cameras, not just one or two, they could have a hundred. Hence it needs human intervention to use it efficiently by checking all the time, which is a time-consuming process and not a much more efficient way to watch them all. Without humans watching, the camera won't be useful, it could just store the videos. This paper proposes a methodology to detect violence by automating surveillance cameras without the involvement of humans by applying deep learning techniques. This work builds a network to detect violence in videos by observing human movements. Proposed network is composed of Resnet50 (Residual Network) in combination with ConvLSTM (Convolutional Long Short-Term Memory).

Input frames are processed to extract humans and eliminate the background with the help of computer vision techniques. Upon background elimination, human poses are estimated with the help of MediaPipe solutions, and the frames are passed to the proposed network. Various experiments were conducted on the existing surveillance datasets empirically validate the efficiency of the proposed network by a 5% increase in accuracy compared to State-Of-the-Art (SOTA) violence detection methods. Once the violence is detected, the system will give out a violence alert message automatically thus enabling the respected authorities to take quick actions.

Keywords: Artificial Intelligence, Violence Detection, LSTM, Convolutional Networks

Received: May 11, 2025. Revised: June 14, 2025. Accepted: July 29, 2025. Published: October 14, 2025.

1. Introduction

E live on a beautiful planet, but in an ugly world. It's Wus, the people making it ugly by violent activities. everywhere. Even if we have sixth sense, we can't live mannerly without any rules. People always need someone like cops or something to watch over us. Even if we protest for justice or need, we can't maintain harmony. People protest for their basic needs, holding up their rights; they gather for the election campaign and temple festivals. There are many places like this where crowds gather and so is the violence. Maybe terrorism is not common, but people take the fight with some serious weapons. It may lead to measurable casualties, and some will lose their normal peaceful life.

So, the government takes every possible action to reduce violence everywhere. One of those is to monitor every public place through CCTV cameras. Private sectors have their CCTVs to avoid violence. To manage and get some worthy value from these, human involvement is necessary. Even with enough human resources, one can't handle all the cameras. Here comes this paper, which is software that always analyzes every camera. It detects violent activities by humans and alerts the system. It will only alert the abnormal activity like aggressiveness, and it won't consider normal activities like handshaking, hugging, dancing.

Proposed system uses deep learning algorithms, which are a powerful collection of neural network learning techniques. Neural networks are a lovely biologically inspired software framework that allows a computer to understand from data collected through observation. ResNet50 (Residual Network - 50) in combination with ConvLSTM (Convolutional Long Short-Term Memory) is utilized. Computer vision, a subset of deep learning, is used to process visual information. Live CCTV video input is processed frame by frame to extract humans with the help of computer vision. MediaPipe Solutions Google's open-source framework, used for media processing. Mediapipe pose is one of the solutions which is customizable that is used to estimate the pose of humans. Then the frames are passed to the network which can detect violence and alert the system if violence occurs.

Rest of the paper is organized as follows: Section 2 discusses various works related to the topic. Proposed system design and data preprocessing are discussed in section 3. While section 4 deals with experimental results section 5 concludes with the notable achievements of the work.

2. Related Works

In [1] Building on a Deep learning-based effective violent activity detection model, Rohit Halder and Rajdeep Chatterjee were able to assist authorities in detecting violent activity in real-time. To detect violent activities, a Convolutional Neural Network-based Bidirectional LSTM was deployed. The ability to forecast and localize the presence of a violent event in a frame is improved by knowing both the past and future path of a video clip. It needs to be tested further with more conventional datasets in which each or even many violent acts, including weapons, are difficult to detect.

In [2] Pin Wang et al. (2020) investigated aberrant behavior detection, focusing on the low precision and reliability of a brute - force attack detection method based on a convolutional neural network and path combination. This method employs artificial features, and profundity features to extract the video's spatiotemporal properties, which are then combined with the trajectory features using a convolutional neural network. By combining two features, one can enhance the accuracy of violent behavior identification. For these data sets, the deep network model suggested in this paper has poor recognition performance.

In [3By eliminating the predicted backdrop from the raw video, Linhao Li et al. (2020) accomplished surveillance object recognition. The Generalized Shrinkage Thresholding Operator (GSTO) is meant to conserve more information by combining the benefits of three basic shrinkage operators. A refined dictionary learning operation is then used to find identity texture patches for the frequently evolving patterns. Finally, by integrating thresholding technique with a spatial and temporal continuity requirement, foreground items are recognized. GSTO is more effective and controllable than traditional shrinkage operators. It is excellently suited to the task of matrix-based background modelling. Interferences such as falling snow, shadows, and irregular object motion prevent us from precisely locating all foreground pixels.

In [4], Ismael Serrano et al. (2018) investigated how to recognize violent actions in surveillance footage scenarios. They offer a hybrid "custom - made" feature framework that gives improved accuracy and computational efficiency than the pooled feature learning method. For the binary recognition challenge, the suggested strategy outperforms alternative “handmade" features and 3D Convolutional Neural Network approaches.

In [5], they worked on object recognition and body position estimation for computer vision which is a subset of deep learning and robotic vision systems. This system uses an RGB-D image as an input to generate colored point cloud data and extract scene attributes. Detected feature descriptors are built using the CSHOT descriptor algorithm, which is based on local shape and texture information. Then, to detect correlations between the scene and the colored point, a two-stage matching procedure is used. They filter out similar inaccuracies and estimate the object's initial 3D posture using the Hough voting algorithm.

In [6], they worked on tracking people and overcoming the problem of the track-by-detection strategy. They proposed a pose-guided tracking-by-detection framework. To make up the missing parts of humans, they used pose-guided person location prediction. They used pose heat maps to cope with person-specific intra-class distractors. To increase the number of people represented, they used Pose Graph Convolutional Networks (PoseGCN).

In [7], they worked on network architecture for human pose estimation and handled pose estimation, location, and semantic information separately. Semantic knowledge is invaluable for a network's depth and breadth to be acceptable. Information about the location uses high-resolution features to get better results by fusing. They have processed different resolutions features which are at various kinds of levels. So, it is called the Parallel Pyramid network.

In [8], They concentrated on quick position prediction while avoiding the resource constraints of the edge device. They created a model of the fast and lightweight pose network (FLPN) for posture estimation, as well as a novel lightweight bottleneck block for lowering computational costs, which can combine the simple network and lightweight constriction into a cost-effective method for accurate pose estimation. They employed structural similarity assessment to adjust the proper ratio of core feature maps and lower the model size.

In [9], They worked on getting high accuracy pose estimation by iterating between optimal position and attitude of a camera; They also used geometric relationships to get the actual value of the camera's attitude; Iterating between ideal and camera altitude. Another method for obtaining an exact camera position is to utilize an estimated camera altitude as the initial value of an evolutionary method.

In [10], They worked to overcome the problem of capturing viewpoints and flexibility of human poses. They used a base network for initial estimation, View In-variant hierarchical correction network to learn the 3D pose refinement, View In-variant discriminative network (VID) to impose high-level restrictions.

3. Proposed Work

Figure 1 shows various building blocks in the proposed methodology. OpenCV library, which is a computer vision solution to subtract the background of live CCTV videos is used.

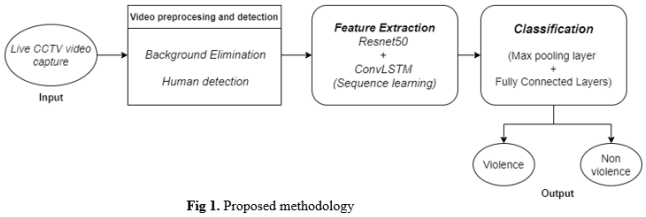

Fig 1. Proposed methodology

3.1 Background elimination:

It is intended to extract humans, everything else will be eliminated. Background modeling is an efficient way to obtain foreground objects in video frames using Statistical approaches. The probability function of the background image will be estimated. Blocks, patterns, and pixels can all be used to extract features. Pixel-wise features tend to produce noisy classification performance because they don't retain local relationships, whereas pattern-wise and block-wise features are apt to be unaffected by small alterations. Extraction of characteristics from comparable pixels can be used to segment the background. Another option is to use Euclidean distance to measure the correlation between pixels in both frames' regions. The similarity threshold will then determine whether each pixel is foreground or background. A dehazing system is made up of a few components.

For that, the KNN background subtraction algorithm computes the Euclidean distance between each segment in the image and defined training region. A function named "BackgroundSubtractorKNN" is used for this. It gives an output image of a 2D matrix which is black and white. By doing the bitwise NOT operation of that output image, one can get the background model. CCTV is fixed on some spot, and it can rotate about some angle. The background will be static and immobile, unlike foreground human movements. Hence a background model is created that knows the clear background. Now one can use this model, which is a background frame to get the perfect foreground by applying the frame differencing methodology Fig.2. It defines that 8 consecutive frames are taken to get subtracted with the background model. Then add all the resultant foreground images to get clear foreground with reduced noise and perfect human movement.

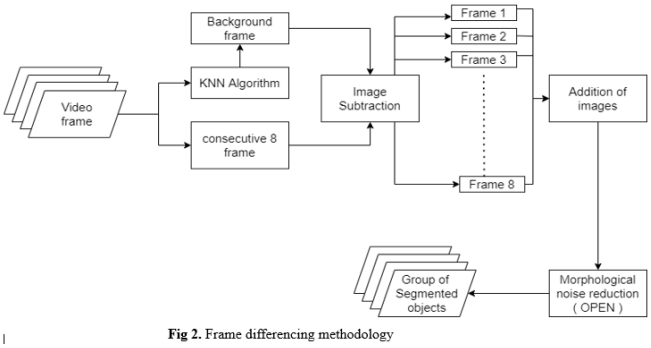

Fig 2. Frame differencing methodology

Finally, a morphological opening operation is performed to erode and dilate an image at the same time, with the same structural element used for both operations. Fractals and thin lines can be removed from such an image while the morphology and structure of big real objects are preserved.

KNN is more efficient for local density estimation, which estimates the density value at point x based on the distance between x and its kth nearest neighbor. It is a non-parametric function so it can adapt to any underlying probability distribution function. The equation is eq.1.

$$[t]p\left(x|x_i\right) \approx \frac{1}{NV}\sum_{m=1}^{N} b^m k\left(\frac{||x_i - x||}{D}\right) \qquad \text{(eq.1)}$$

Here K is a Kernel function; If u < 1/2, K(u) will be 1 else 0; D is Bandwidth; N is the number of Samples; V is Dd. If a video sequence sample is labeled as foreground, then bm =0.

The background model deals only with samples that satisfy bm. In eq.2, T is the threshold that is related to value V.

3.2 Human Detection and pose estimation:

Now, the output is detecting moving things which will be humans, vehicles, etc... that require color images so that it could be processed further in Neural Network. To do so, 2D matrix of the original image for each dimension is taken and converted into a grayscale image. Now, three 2D grayscale matrices are obtained. Bitwise AND operation is performed among them. It will give humans only video. The human detected image undergoes further processing to estimate the human pose. This would relieve proposed network's workload while training and testing. This work leveraged BlazePose research, which is also used in the ML Kit Pose Detection API, to create media pipe, an ML solution for elevated body poses monitoring that infers 33 3D landmarks and background image patches on the full body from RGB video frames. Proposed solution achieves real-time performance on the newest smart phones, computers, and laptops, whereas current state-of-the-art systems rely mostly on powerful desktop systems for inference.

A two-step detector-tracker ML pipeline is used in the solution, which has been demonstrated to be effective in MediaPipe Hands and MediaPipe Face Mesh solutions. The pipeline initially locates the individual region-of-interest (ROI) well within frame using a detector. The tracker estimates the postural landmarks and segmentation masking within the ROI that use the ROIcropped frame as input. It's worth mentioning that the detector is only used in video use cases when it's truly unavoidable, such as for the first frame or when the tracker can't recognize body pose presence in the frame before it. For all other frames, the pipeline simply estimates the ROI using the previous frame's pose landmarks. In MediaPipe Pose, the landmark model predicts the placement of 33 pose landmarks, ranging from the eyes to the foot index. The results of pose estimation are in Fig 5.

3.3 Feature extraction and classification

The pose estimated frame is passed to the network on training for feature extraction. The network is Resnet-50 with an average pooling layer and a fully connected layer at last which is doing the classification. Live videos are processed frame by frame. The network gives out the percentage of violence in the frame. It can be put in the frame for indication of violence.

3.4 Network architecture

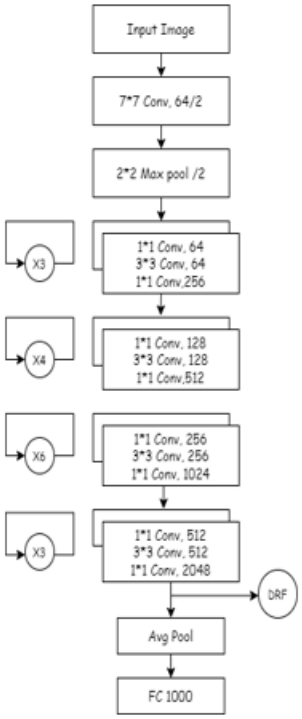

Fig 3. Resnet-50 with skip connection

Resnet-50 as shown in Fig.3 in combination with convLSTM is used. As final dense layers are removed, it outputs the feature vector or also known as transfer values, and not the predicted class. Then these transfer values with a fixed dimension of (1,20,2048), into an LSTM layer. An avg-pooling layer of size 4 and regular dropout layers, dropping 50% information each time is also included. Finally, the chain of fully connected dense layers of size 1024, 50, 16, and finally a binary output perception with a sigmoid activation function. ReLU Activation between each fully connected dense layer is introduced. Batch Normalization layer is placed just before dense layers to prevent the internal confounders shift problem. Binary cross-entropy as loss function and RMSprop as optimizer is utilized with 20% of the data being chosen for validation and the remaining 80% for training. After 5 epochs, the learning rate is half of what it was at the start. Regarding validity loss, I've reached a stalemate. I experimented with hyper-parameter adjustments on the Hockey dataset before applying it to the other datasets. To avoid memorization, 20 epochs are performed and stopped after 5 rather than 10 as in the final optimum network training. For train-test data, 80-20 split is done. The baseline hyper-parameters are where tuning is started. Each hyperparameter is evaluated separately, and the best value is chosen for further evaluations. The following is the sequence in which the hyper-parameters should be executed, in priority order: Learning rate, sequence length, augmentation usage, dropout rate, and CNN network training mode are all factors to consider (retrain or static). In Table 1, presents the different hyperparameters evaluated in each iteration.

Table 1. Hyper Tuning Parameters

CNN architecture	*Res Net 50*	*Inception V3*
Learning rate	1e-4	1e-3
Use augmentation	True	False
Number of frames	20	30
Dropout	0	0.5
Train type	Retrained CNN	Static CNN

Full names of authors are preferred in the author field but are not required. Put space between authors' initials.

4. Experimental Results

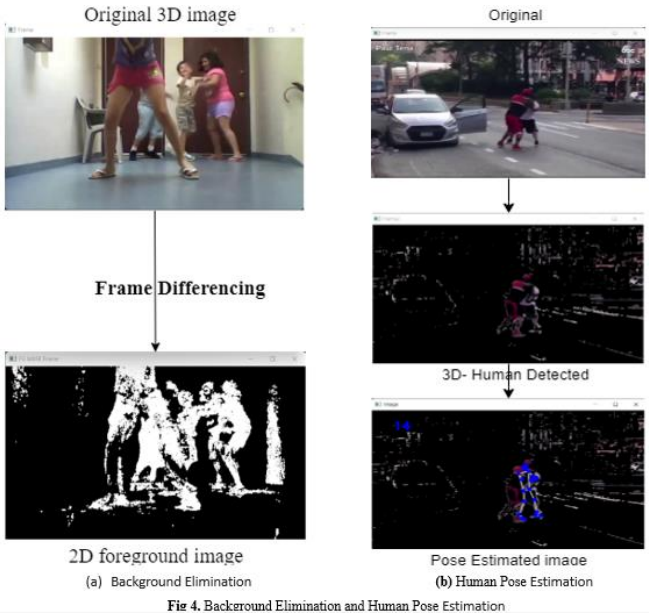

Fig 4. Backeround Elimination and Human Pose Estimation

Hyper-parameter mentioned in Table 1 were applied on the proposed network model for the videos available in "CCTV footage" dataset (RWF-2000. Dataset consists of 2000 videos, 1200 videos for training, and 800 videos for validation. The output of the frame differencing applied frame is shown in Fig.4(a). The output of 3 channeled images by the proposed algorithm and human pose estimated image is shown in Fig.4(b).

The violence detected frame, which is the output of proposed network, is shown in Fig.5(a). The sample output for the normal scenario image is shown in Fig.5(b). The hyper-tuning test accuracy for each of the hyper-parameter's values. Proposed network which is convLSTM in combination with Resnet50 gives the best performance of 92% accuracy, the InceptionV3 gives the accuracy of 87% which is not far from the proposed network. The learning rate of 0.0001 resolved with this accuracy, nothing higher than this didn't give this much accuracy. The learning rate of 0.001 gives very poor results of 40% accuracy on this network. The length size of the sequence improves the accuracy by 2.5%. The dropout of 50% did not improve the model performance and resulted in only 87% The augmentation increases the accuracy by 5%. Finally, as expected, the CNN weights were not retrained and had bad results of 60% accuracy.

Fig 5. Sample Output

During the evaluation, it is observed that many hyperparameters had serious effects on the model results. Proposed model which is convLSTM in combination with Resnet50 and the inceptionV3 models had significantly better results than the VGG19. This can be easily explained by the fact that both architectures had significantly better results in image classification with much fewer parameters while having more networks. The Resnet50 has 168 layers whereas the InceptionV3 has 159 layers. This makes proposed model better results than InceptionV3 in case of violence. But In ImageNet, InceptionV3 surpasses proposed model, displaying 1.5 percent higher accuracy. While retraining CNN networks enhances their performance significantly, training them on violent data allows them to tailor and uncover meaningful patterns of violence depending on human position and speed, which they then output to the ConvLSTM layer. As previously stated, the initial learning rate had a significant impact on the learning process; in comparison to 0.001, the lower commencing learning rate of 0.0001 increased the network's learning. It shows that a high learning rate creates drastic changes in network weights, obstructing the network's capacity to converge in the proper direction, whereas a low learning rate pressures the network to update its weight slowly but cautiously into the right direction of dropping. Because the problem and datasets are domain-specific, the dropout did not help the network in this trial scenario. When multimedia data is more heterogeneous, with varied video quality and camera placement, and as more classes are accessible, such as type of abuse, number of people involved, violence tool, extent of damage, etc., generalization will be crucial. The data augmentation process helped the model deal with the small amount of labeled data; the augmentation increased the number of samples and helped the model to find meaningful patterns in the frames. For the analysis of the optimized results, first "Movies" dataset is used, and the model reached 100% accuracy. It is concluded that it is a relatively "easy" dataset to classify because the learning reduction only occurred once and nearly by the end of the training session. The optimized model fitted on the "Violence CCTV Footage" dataset had arrived at the most reduced score out of all the datasets settling at 91.4% precision. Videos in this dataset contain enormous groups where even in the "violent" recordings most of the group is an onlooker and doesn't intercede in the brutal demonstration.

References

[1] Rohit Halder and Rajdeep Chatterjee "CNN‑BiLSTM Model for Violence Detection in Smart Surveillance" SN Computer Science (2020) 1:201 DOI:/10.1007/s42979-020-00207-x

[2] Pin Wang, Peng Wang, En Fan "Violence detection and face recognition based on deep learning" Pattern Recognition Letters doi: 10.1016/j.patrec.2020.11.018

[3] Linhai Li, Zhen Wang, Qinghua Hu, Senior Member, IEEE, Yongfeng Dong," Adaptive Non-Convex Sparsity based Background Subtraction for Intelligent Video Surveillance" IEEE Transactions on Industrial Informatics VOL. 14, Issue: 8, AUGUST 2020 10.1109/TII.2020.3009111.

[4] Ismael Serrano, Oscar Deniz, Jose L. Espinosa-Aranda, Gloria Bueno "Fight Recognition in video using Hough Forests and 2D Convolutional Neural Network" IEEE Transactions on Image processing, Vol.27, Issue:10, 2018 10.1109/TIP.2018.2845742

[5] Chi-Yi Tsai and Shu-Hsiang Tsai " C. -Y. Tsai and S. -H. Tsai, "Simultaneous 3D Object Recognition and Pose Estimation Based on RGB-D Images," in IEEE Access, vol. 6, pp. 28859-28869, 2018, doi: 10.1109/ACCESS.2018.2808225.

[6] Qian Bao, Wu Liu, Yuhao Cheng, Boyan Zhou, and Tao Mei "Pose-Guided Tracking-by-Detection: Robust Multi-Person Pose Tracking" IEEE Transactions On Multimedia, VOL.23, 2021. Doi: 10.1109/TMM.2020.2980194

[7] Lin Zhao, Nannan Wang, Chen Gong, Jian Yang, and Xinbo Gao "Estimating Human Pose Efficiently by Parallel Pyramid Networks" IEEE Transactions On Image Processing, VOL. 30, 2021 Doi: 10.1109/ TIP.2021.3097836

[8] Haopan Ren, Wenming Wang, Kaixiang Zhang, Dejian Wei, Yanyan Gao, And Vue Sun "Fast and Lightweight Human Pose Estimation"DOI: 10.1109/ ACCESS.2021.3069102

[9] Q. Wang and Z. Zhao, "An Accurate and Stable Pose Estimation Method Based on Geometry for Port Hoisting Machinery," in IEEE Access, vol. 7, pp. 39117-39128, 2019, doi: 10.1109/ACCESS.2019.2907222.

[10] G. Wei, C. Lan, W. Zeng and Z. Chen, "View Invariant 3D Human Pose Estimation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4601-4610, Dec. 2020, doi: 10.1109/TCSVT.2019.2928813.