Automated Software Engineering (2022) 29:28
https://doi.org/10.1007/s10515-022-00323-3

UAV surveillance for violence detection and individual identification

Anugrah Srivastava[1] · Tapas Badal[1] · Pawan Saxena[1] · Ankit Vidyarthi[2] · Rishav Singh[3]

Received: 21 September 2021 / Accepted: 3 January 2022 / Published online: 2 March 2022

Abstract

Violence detection and face recognition of the individuals involved in the violence has an influence that's noticeable on the development of automated video surveillance research. With increasing risks in society and insufficient staff to monitor them, there is an expanding demand for drones square measure and computerized video surveillance. Violence detection is expeditious and can be utilized as the method to selectively filter the surveillance videos, and identify or take note of the individual who is creating the anomaly. Individual identification from drone surveillance videos in a crowded area is difficult because of the expeditious movement, overlapping features, and bestrew backgrounds. The goal is to come with a better drone surveillance system that recognizes the violent individuals that are implicated in violence and evoke a distress signal so that fast help can be offered. This paper uses the currently developed techniques based on deep learning and proposed the concept of transfer learning using deep learning-based different hybrid models with LSTM for violence detection. Identifying individuals incriminated in violence from drone-captured images involves major issues in variations of human facial appearance, hence the paper uses a CNN model combined with image processing techniques. For testing, the drone captured video dataset is developed for an unconstrained environment. Ultimately, the features extracted from a hybrid of inception modules and residual blocks, with LSTM architecture yielded an accuracy of 97.33% and thereby proved to be noteworthy and thereby, demonstrating its superiority over other models that have been tested. For the individual identification module, the best accuracy of 99.20% obtained on our dataset, is a CNN model with residual blocks trained for face identification.

Keywords Violence detection · Drone surveillance videos · Deep learning · LSTM · Transfer learning · Violent individual

T. Badal, P. Saxena, A. Vidyarthi, R. Singh have contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

With the increase in the crime rate, surveillance becomes all the more important. Action Recognition using Computer Vision Techniques has been an active research area where different human activities like walking, talking, sitting, etc., have been identified (Cao et al. 2017; Ramanathan et al. 2014; Penmetsa et al. 2014). In this context, violence detection can be a class of activity recognition. Violence recognition has developed significantly throughout the years and it has become the need of the hour due to increasing incidents occurring all over the world (Li and Chuah 2017; Choi et al. 2009). The dependence of human personnel monitoring these surveillance systems has decreased due to the emergence of automated smart surveillance systems (Mumtaz et al. 2018). These systems can be effectively trained for detecting violent activity (Fu et al. 2017; Singh et al. 2018).

Violence detection in parallel with violent individual identification can be done through CCTV monitoring which is the most basic form of surveillance deployed in public places (Bindemann et al. 2017; Aydin 2019; Goya et al. 2009). Monitoring by drones offers some advantages over surveillance by stationary cameras. We require more than one security camera positioned at multiple angles and locations to cover a 360-degree view of a whole region, but a single drone can do the job. To distinguish between violence and non-violence, various challenges need to be addressed. A proper dataset for learning violent and nonviolent behavior in various surroundings is essential to be applicable for real-time surveillance systems. Models that independently rely on handcrafted features which are typical for detecting human activity for years but these suffer from the limitation of being unable to adapt real-life data sets. Recently, enormous deep learning techniques are proposed. In the case of 2D CNN, analyzing a single frame to locate spatial features but proves to be ineffective as concerned with motion information in a video (Donahue et al. 2017). 3D CNN aims for feature extraction directly from the raw video frames of the input by applying convolutions on the sequence of video frames (Ding et al. 2014). These frames that contain people subsequently, it is possible to feed them into the 3D CNN that processes them. This functionality of 3D CNN makes it suitable for violence recognition in videos (Ji et al. 2013). Therefore, video is considered as several contiguous frames to obtain spatial-temporal features. The proposed study is broken into two portions. The first part of the paper mostly discussed violence detection from drone-captured data while the second section concentrated on violent individual identification based on the recognized faces. In the first section, pre-trained neural network models are utilized to extract features from video frames, that have been initially trained on the ImageNet dataset. Additionally, a combination of two models is also used for the extraction of the features. At the end of the day, the features derived from the combination of InceptionV3 and ResNet101V2 pre-trained models, with LSTM, proved to be significant.

The violent individual model which runs side-by-side has the complicated task of face recognition which is accomplished by using two models face detection and face recognition and the concept of one-shot learning for achieving better recognition performance. Face recognition using drone images is a difficult task because of the poor resolution, scale variation, and aerial footage. This model is constructed with state-of-the-art deep learning facial recognition. This model is a ResNet network with 29 convolutional layers (He et al. 2016). It is a simplified version of the ResNet-34 network from the paper Deep Residual Learning for Image Recognition (He et al. 2016), with a few layers deleted and the number of filters per layer cut in half. The network was trained from the start to the end on a dataset of approximately 3 million faces. The face scrub dataset (Kemelmacher-Shlizerman et al. 2016) and the VGG dataset (He et al. 2016). The overview of the proposed method is shown in Fig. 1.

The paper's contribution is summarized in the following points:

(1) Target dataset is created for testing the efficiency of the developed model using drones for an unconstrained environment and diversified views from different heights with violent and non-violence activities performed by single/multiple persons.

(2) In this study, violent detection using drone surveillance is proposed. Besides the single model architectures, various hybrid models to obtain the features extracted from videos are implemented for the violence detection model. The proposed work utilizes the state-of-art technique of transfer learning with LSTM architecture, which proved to be superior to other single model architecture for drone-captured videos.

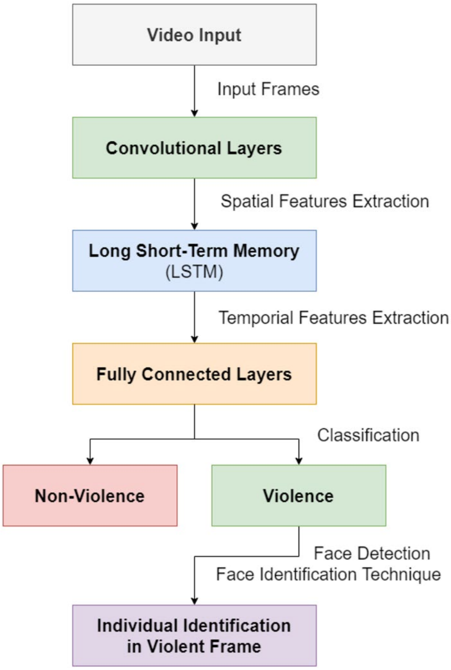
Video Input
Input Frames
Convolutional Layers
Spatial Features Extraction
Long Short-Term Memory (LSTM)
Temporial Features Extraction
Fully Connected Layers
Classification
Non-Violence
Violence
Face Detection
Face Identification Technique
Individual Identification in Violent Frame

Fig. 1 The figure illustrates the overview of the proposed method

(3) Furthermore, identification of individuals involved in violence frames using drone surveillance is done by processing the input face image and applying the face detection model to detect faces and extract the features. Later the facial features are run on a trained CNN model, for face recognition.
(4) In this literature, a comparative study of ten model architecture is manifested for violence detection in video.

The rest of this work is structured as follows: Sect. 2, Related work is discussed violent individual recognition. In Sect. 3, a detailed overview of the dataset is discussed. In Sect. 4, the Methodology with subsection Model Architectures is discussed. In Sect. 5, the Experiment Evaluation along with Experimental Result and Analysis is introduced briefly and the paper is concluded in Sect. 6 with a conclusion and future scope.

2 Related work

The paper presents the identification of the violent individual for surveillance video data created. Identification of a violent individual is essentially two-task research, violence recognition, and identifying the individual causing the violence. Rapid progress has been seen in Human activity recognition, with the development of CNN-based models (Zhang et al. 2016; Ha and Choi 2016; Wang et al. 2016; Ordóñez and Roggen 2016). Significant efforts for fight detection and violence detection are made by researchers.

Initial propositions of violence recognition systems were based on fire and blood detection, extracting the degrees of movement and perceiving sound input like screams or gunshots. Features extracted from audio were fed into Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) to detect gunshots or other sounds (Cheng et al. 2003). Violence scenes in movies were classified into 6 classes using Bayesian Networks (Giannakopoulos et al. 2007). Three of those classes are violent: shots, fights, and screams. A deep learning audio-based approach (Zaheer et al. 2015) was implemented on a custom-made scream database where interpolated MFCC features were fed as input for classification. KNN classifier (Cheng et al. 2003) was used by combining audio-visual features consisting of motion orientation variance, audio characteristic features, and average motion. Another approach to detecting aggression in videos (Zajdel et al. 2007) where a system called 'Cassandra' was developed in which motion features and scream signals were used for classification. SVM classifier based on audio features and frequency domain for violent activity detection is implemented (Giannakopoulos et al. 2006). These methods, although accurate, fail to be of any relevance as audio and color are absent in real-time systems. This increases the susceptibility to obtain false positives and decreases the reliability of the system (Naik and Gopalakrishna 2017).

Violence detection through handcrafted features like "bag-of-words" (Bermejo Nievas et al. 2011), Motion SIFT (MoSIFT) (Laptev and Lindeberg 2005), Space-Time Interest Points (STIP) (Laptev and Lindeberg 2005), Motion features, Motion blobs (Deniz et al. 2014) were implemented for this purpose. A model is proposed to implement a long-term temporal structure for violence recognition (Zhou et al. 2017) by combining the results obtained from giving different inputs like optical flow images, acceleration images, and RGB images for spatial networks. Convolutional Neural Network functionality was extended for the classification of 1 million videos consisting of 487 classes (Karpathy et al. 2014). Violence Detection using 3D convolutional networks and spatial-temporal features (Ullah et al. 2019). They implemented it in three phases, starting with a 2D CNN that could recognize frames with individuals in them and feed that information into a 3D CNN model consecutively. They used the SoftMax classifier to categorize the video as violent or non-violent. On the 3 different data sets considered, it was observed that other well-known methods are outperformed by the developed method. A sensor network approach (Baba et al. 2019) to recognize violence in videos was implemented in which time-domain classifiers and deep neural networks are utilized for accommodating both temporal and spatial features.

By taking advantage of transfer learning with LSTM networks combined with CNN, the model was implemented (Dandagpl et al. 2019). CNN and LSTM together achieved better performance than pure pre-trained models. ImageNet models like VGG16, VGG19, and ResNet50 were used in combination with fully connected layers, LSTM network, and Spatial Transformation Networks (Sumon et al. 2020). The STN overcomes the limitations of a regular CNN in terms of handling input variations like scale, viewpoint variation, etc. The working of the spatial transformation network which applies attention to the extracted features and performs transformations like scaling, rotation, and translation is explained (Jaderberg et al. 2015).

The face recognition field, effort, and research are made by many major universities and companies. Kohonen is credited with being a forerunner of the most widely used facial recognition technology, which used a basic neural net utilizing a network (Kohonen 2012) of Eigenfaces by calculating the eigenvectors via the autocorrelation matrix of the face images. A real-time face detection system (Xu 2012) requires the size and position of each person included in the picture or video. Tracking the correspondence between different faces is also necessary for the frame. Several face detection algorithms and methods can be learned from the most frequently cited article Robust Real-time Object Detection (Wu et al. 2004) which makes face detection workable. A combination of tracking and face detection model is proposed in the article (Li et al. 2008) that received the Best Student Paper CVPR 2007. Recently, various approaches, algorithms, and techniques have been either combined with conventional LBP or updated LBP to obtain facial recognition and improve facial recognition accuracy. The method of convolutional neural network (CNN) based face detection with deep face tracking and well-known CNN face identification algorithm (Saypadith and Aramvith 2019) was proposed and the Deep Learning System and Embedded GPU System were used. Besides, an Improved Real-Time Face Recognition Local Binary Pattern Histogram (LBPH) (Deeba et al. 2019) was used to achieve real-time facial recognition.

The drone surveillance video will be used to evaluate the proposed work's performance. The task of identifying violent individuals is challenging because drone-recorded images can be distorted by lighting changes, poor resolution, and changes of size and conflicts as several events take place. In the first place, the proposed work estimates whether the video involves violence using the concept of transfer learning with LSTM architecture. Later, a CNN model is proposed for identifying a violent individual.

3 Dataset

Huge datasets for human action recognition have been created over the years, but a dataset for violent individual recognition using a drone has not been made explicitly. It is critical to have a suitable dataset to execute better work. The study experimented on three different datasets. The overview of the dataset is shown in Table 1.

3.1 Source dataset

The data set that's been utilized for this study in the process of training the system and obtaining the transfer values are the Real-Life Violence Situations Dataset and Hockey dataset as shown in Fig. 2, and further for testing the system we have developed the dataset.

3.1.1 The hockey dataset

The Hockey dataset is the first of its type, meticulously crafted benchmarks dataset for violent activity recognition (Bermejo Nievas et al. 2011). The dataset has 1000 video clips having 50 frames with an individual frame having a size of 720 × 576 and resolution 360 × 288. The dataset is separated into two groups, fight, and non-fight, with 500 clips in each category, and are labeled accordingly. The dataset is captured from hockey games played in the National Hockey League (NHL) of real-life violent events. The hockey dataset is difficult to work with because of the rapid camera motion used in filming non-fight footage of real-time hockey games.

Table 1 Table represents the computational details of used datasets

Characteristics	Training		Testing
	Hockey (Bermejo Nievas et al. 2011)	RSLV (Soliman et al. 2019)	
Violence video	500	1000	80
Non-violence video	500	1000	70
Average duration	1–2 s	3–7 s	5–10 s
Resolution	–	(480–720)p	1080p

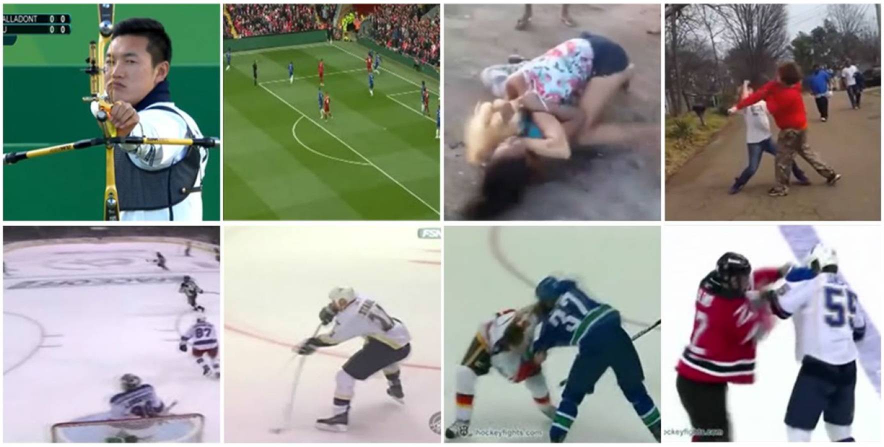

Fig. 2 Samples from Real-Life Violence Situations (Soliman et al. 2019) (Row 1) and Hockey datasets (Bermejo Nievas et al. 2011) (Row 2). Non-violent(Column 1, Column 2) frames and violent frames (Column 3, Column 4)

3.1.2 Real-life violence situations(RLVS)

The Real-life violence situations dataset (Soliman et al. 2019) contains 1000 violence and 1000 non-violence videos with varying resolution(480p–720p) with 3–7 seconds duration and 5 seconds average duration. The range of video frame width is between 224 and 1920, while the range of height is between 224 and 1080 with a video of 397 × 511 average size. These videos were collected from various sources and social networking platforms to cover a wide range of scenarios. From YouTube videos to real-life street fights, fights in the football field, political riots which are labeled as violent videos, this dataset contains videos taken in both indoor and outdoor environments. Videos of human activities like eating, walking, etc. were collected for the nonviolent directory which is labeled as non-violent videos. The street fights are captured using either a mobile camera or surveillance camera which are positioned at a height of 2.5–3 m above the ground.

The challenging features of both Hockey and Real-Life Violence Situations datasets are making them the most suited source for the purpose of recognizing violence.

3.2 Target dataset

In testing data, the violent sequences of the dataset are taken from a developed dataset from the paper (Srivastava 2021) and other violent and nonviolent sequences are added in the dataset. The testing dataset has been proposed to influence research work in a violence detection field hence after the publication of the study, the dataset will be made available to researchers. To our knowledge, this is the first of this kind of dataset in which the violence and non-violence scenes are captured using drone surveillance for diversified views and the unconstrained environment from different heights. The

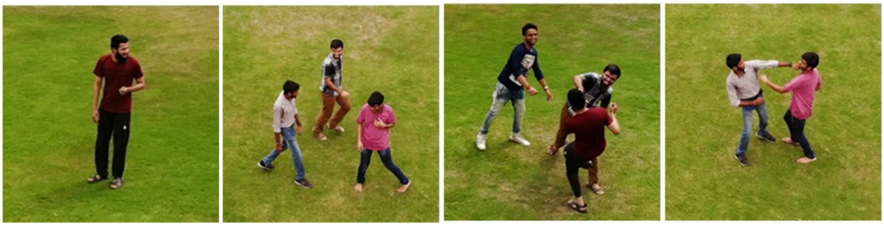

Fig. 3 Samples from testing dataset. Non-fight frames (Column 1, Column 2) and fight frames (Column 3, Column 4)

activities are carried out by people between 17 to 30 age groups of both genders. It has 150 video clips with a resolution of 1080p at 60 fps, captured from 3-6 meter height from the ground using a drone (DJI Mavic Pro). The datasets, violent videos have a great diversity with dual conflict to crowd fight as shown in Fig. 3. The distinctiveness and irregularities the drone-captured videos suffer like poor resolution, shadows, blurring, illumination, scale variation, and brightness, benefit the testing set to check the model accuracy precisely.

4 Methodology

The section provides a detailed description of the approach proposed to recognize violent individuals via drone surveillance and the workflow of the section is as follows:

(1) It is discussed a complete overview along with the algorithm of the entire system design.
(2) Following that, the proposed technique of the system's transfer learning-based violence detection model with LSTM architecture is presented.
(3) Finally, the individual identification method is explored in the context of violence.

4.1 System architecture

Raw video data need to be organized and manage accurately. The system consists of the function used to get 20 frames and horizontally flip each frame. Therefore, a total of 40 frames from one and all video files are captured and the frames are transformed to a suitable format as per the input's requirement of the respective model. The shape of the datasets videos-

$$DataSet(V_n, F, W, H, num\_channels) \tag{1}$$

where Vn denotes the number of videos in the dataset, F denotes the number of video frames, W denotes the width of the frame and H represents the height of the frame.Number of channels(*num_channels*) is set to 3(RGB). The shape of the input required by the respective pre-trained model is observed and accordingly, the frame size is resized $[num\_channels \times (img\_size \times img\_size)_m]$. Hence, the shape of the pre-processed datasets videos-

$$DataSet(V_n, 40, W(img\_size)_m, H(img\_size)_m, 3) \tag{2}$$

The proposed work objective is to develop a surveillance system that can monitor automatically and detect the person causing violence using the surveillance videos and assists the authorities to spot these events and take the necessary measures. The computerized video surveillance system first learns features and then trains on those learned features. The system extract frame features using seven well-known ImageNet models VGG16, VGG19, ResNet101V2, DenseNet201, InceptionV3, MobileNet, NASNetLarge. Besides the seven models, three combinations of two models each with similar parameters are used for extracting the features. For combination, the final transfer values are calculated using the basic technique of simple mean transfer values from both models. Parameters are the weight matrices that play an important role in model predictive power. The weights that are learned during training can also be defined as parameters. Different ImageNet models have a different number of parameters and unique characteristics, based on which models are selected for the transfer learning process. The system is developed in such a way, that in a short video sequence it helps the user by determining whether the crime occurred or not. The process entails extracting a set of video frames and delivering them to ImageNet models that have already been trained and obtaining the transfer values as an output from penultimate SoftMax layers. The features taken from the models that have already been trained are stored in the cache files used from training in one file and the ones used for testing in another one. Load the transfer values which are saved in the disk, then another network architecture is trained using LSTM neurons, which are a special sort of neuron. These neurons contain memories and can comprehend the video's temporal information. The LSTM model architecture has been designed with one LSTM layer followed by the dense layer. The hyperbolic tangent function is traditionally used by the LSTM layer as the activation. The last layer consists of the softmax classifier. If they detect violence in a single frame at a particular time, it will be recognized as a violent video. The Fig. 4 represents complete model architecture of proposed system.

The captured violence frames are processed to capture the 68-point features, followed by an algorithm for face detection. The face recognition model assisted by facial features is used to identify faces once faces have been identified. The identified faces metadata will be extracted to identify the person and if the frame data match within a certain tolerance, the label associated with that frame is displayed.

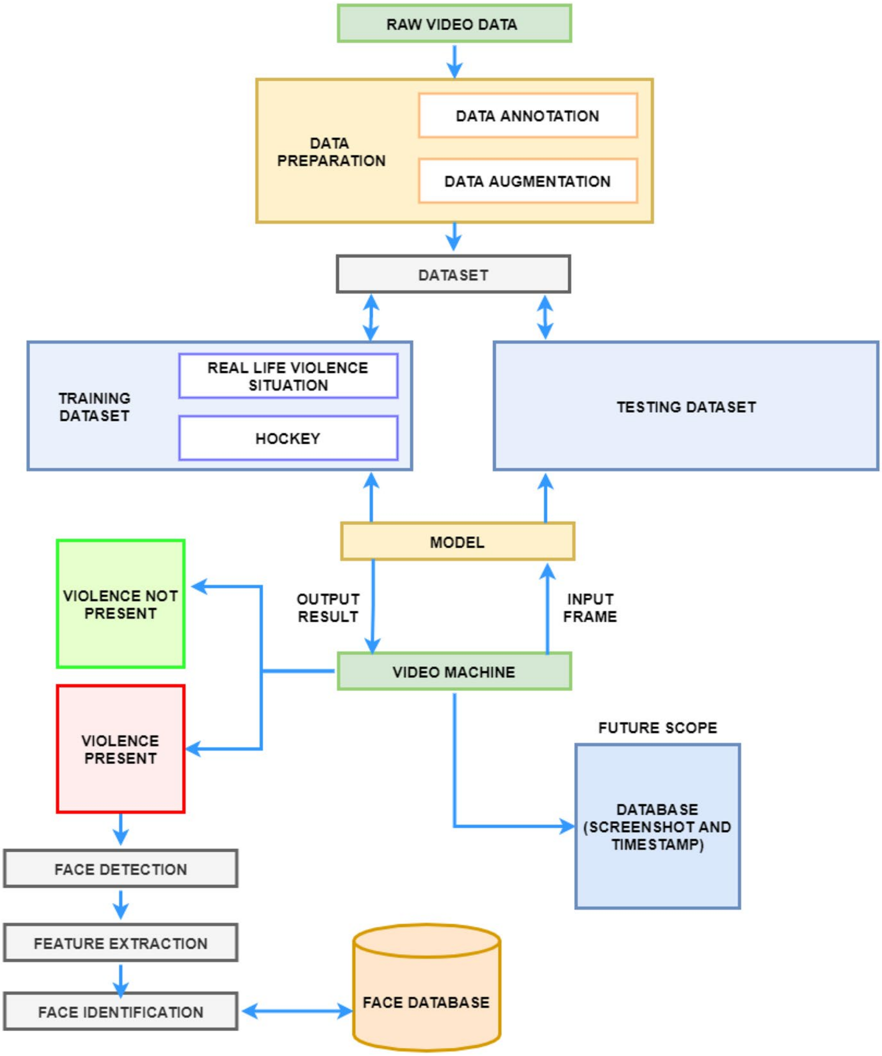

Fig. 4 The figure illustrates complete model architecture of proposed system

Algorithm 1 Violent Individual Identification

Part 1 – Violence Recognition

Require: V_a represents video clips of $DataSet$ videos, W denotes the width of frame and H denotes the height of frame. num_c denotes the number of channels

$DataSet(V_a, W, H, num_c)$

Require: The system apply function $F(x)$ on video clips V_a to extract and resize the frames to width $W(img\_size)_m$ and height $H(img\_size)_m$ according to the input shape requirements by the respective pre-trained model M_{CNN} from source ImageNet dataset. $(Frame)_n$ is the total number of frames of video V_a.

Require: $(transfer\_value)_m$ denotes the number of extracted features from each frame and $(transfer\_value\_size)_m$ is the size of penultimate layer of the selected model. $(rnn\_size)_m$ is the units of LSTM as the dimensionality of the output shape. $(L)_m$ is the LSTM features.

Require: $(Output)_m$ represents a sequence of fully connected dense layers with softmax activation function for the purpose of the violence and nonviolence classes classification.

```
for each $V_a \in DataSet$ do
    $F(x)[V_a] \rightarrow (V_a, (Frame)_n, W(img\_size)_m, H(img\_size)_m, num_c)$
    for each $Frame \in (Frame)_n$ do
        To obtained the Transfer Value
        if Single Model used then
            $(transfer\_value)_m \leftarrow M_{CNN}(Frame, (transfer\_value\_size)_m)$
        else if Hybrid Model used then
            $(transfer\_value)_m \leftarrow \frac{1}{2}\sum_{i=1}^{2} M_{CNN}(Frame, (transfer\_value\_size)_m)_i$
        end if
    end for
    for each $(transfer\_value)_m$ do
        $L_m \leftarrow LSTM((rnn\_size)_m, (transfer\_value)_m)$
    end for
    for each $L_m$ do
        $(Output)_m \leftarrow softmax((L)_m)$
    end for
end for
```

Part 2 – Individual Identification

```
Require: $V_{violence}$ represents the violence videos, $(Frame)_{violence}$ represents
    the each frames of the violence video $V_{violence}$, $Face(f)$ represents function
    to detect all the $face_n$ in the frame using 68-point feature/ 5-point
    features. $(CNN)_{model}$ is used to generate normalized encoding of the
    faces detected by the function $Face(f)$.

Require: $I_{database}$ is the face encoding stored in a database and $I_{generated}$ is
    the encoding of the face generated by the CNN model that is captured
    from a frame of video.

for each $(Frame)_{violence} \in V_{violence}$ do
    $Face(f)[(Frame)_{violence}] \rightarrow face_n$
    $(CNN)_{model}[face_n] \rightarrow I_{generated}$

    for each Face encoding generated $I_{generated}$ do
        $tolerance\_value = \sqrt{\sum_{(i,j)}(I_{1(i,j)} - I_{2(i,j)})^2}$
        if $(tolerance\_value)_{image} \leq 0.5$ then
            Output-The label is assign to face encoding.
        else
            Output-The label 'unknown' is assign to face encoding.
        end if
    end for
end for

A bounding box is created around the face with the corresponding (person
name/id or 'unknown') label below its bounding box.
```

4.2 Transfer learning for violence detection model

A Deep Learning-Based Technique, Convolutional Neural Network (CNN) stack is used for feature extraction from the frames of the input video sequence in combination with LSTM which is used to predict the conclusion to the video sequence. In the deep learning domain, predictive modeling problems use feature learning techniques which are appealing as a compound task of image recognition is done by learning complex underlying data representation, as compared to hand-crafted feature descriptors. The specific problem is learned and learned features are then acquired. A new task with another problem is solved by reutilizing the learned features. This prominent technique is known as transfer learning. Object classification and categorization are domains where the approach has been successfully used. The deep learning CNN model requires a large labeled dataset for training. The complex and demanding task of data annotation can be avoided by using transfer learning. However, the insufficiency in providing the amount of data to learn optimal deep features brings about an issue of significant overfitting. The approach of transfer learning plays a vital role in overcoming the problem of overfitting for small datasets. A constrained dataset is used to develop a network architecture with a target task using pre-trained existing network architecture and learned features in transfer learning.

Figure 5 represents the blocks of convolutional layers followed by dense fully connected SoftMax layers, pre-trained on more than a million ImageNet images with 1000 class output. The target task architecture is created using the source task network by utilizing transfer learning, trained on Hockey and Real-Life Violence Situations Dataset with output classes used as an input for the LSTM network for violent and non-violent activities. The LSTM network is one of the Recurrent Unit's variations. The Recurrent Unit is the basic building block in a Recurrent Neural Network (RNN). A recurrent neuron has an internal state which serves as a kind of memory and is updated every time the unit receives a new input. Two types of features are extracted sequentially.

4.2.1 Transfer learning features

The features extracted consist of spatial features of the frame which are extracted using the transfer Learning technique, from pre-trained models convolution layers on the ImageNet dataset.

When Single Model used to obtain the Transfer Value

$$(transfer\_value)_m = (V_a, 40, (transfer\_value\_size)_m) \tag{3}$$

When Combination Model used to obtain the Transfer Value

$$(transfer\_value)_m = \frac{1}{2}\sum_{i=1}^{2} Model(V_a, 40, (transfer\_value\_size)_m)_i \tag{4}$$

where $(transfer\_value)_m$ represents the number of extracted features from each frame for each model respectively and $(transfer\_value\_size)_m$ is the size of the layer immediately before the final layer of classification of the selected model.

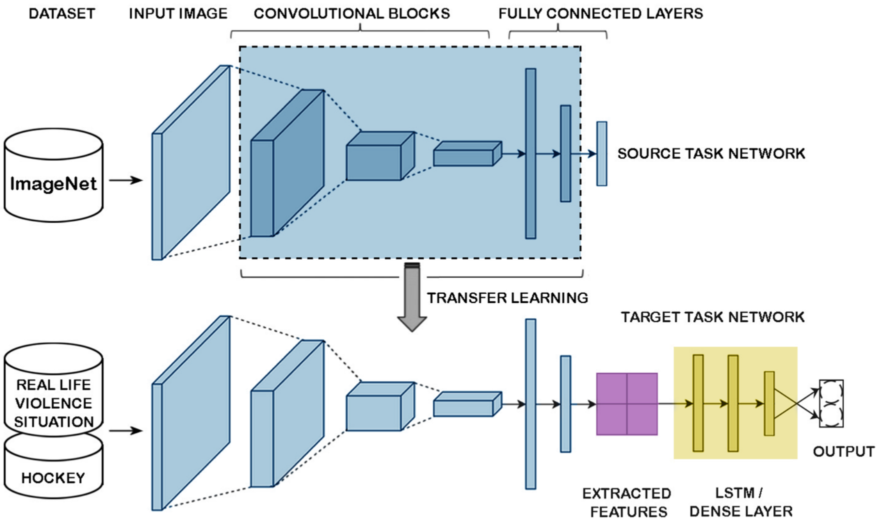

Fig. 5 The figure illustrates the architecture for violence detection using transfer learning

4.2.2 LSTM features

Violent action between people is scattered across a sequence of frames, LSTM is utilized to extract temporal features as the second feature set to detect fluctuation over time. The LSTM setup implemented consists of $(rnn\ size)m$ units as the output space dimensionality and the tanh function as an activation function.

$$input\_shape = (40, (transfer\_value\_size)_m) \tag{5}$$

The output size of after features extracted from LSTM- $((rnn\_size)_m, input\_shape)$.

For the categorization purpose, a sequence of fully connected layers is used. The details of classification layers and output shape are represented in Table 2. The last layer contains only 2 neurons with softmax activation function for violence and non-violence classes classification.

4.3 Individual identification model

Following the positive detection of violence, the classified violent video proceeds through the face recognition model which recognizes the individual in the violent frame using the prominent technique of One-shot learning van der Spoel et al. (2015); Wang and Deng (2021). One-shot learning is a classification problem in which a very small number of examples or one example is supplied for every class and utilized to construct a model, that must predict numerous unknown examples in the future. Facial identification includes three basic steps face detection, feature extraction, and face recognition.

The performance of the face identification algorithm is strongly dependent on the accuracy performance of the extraction of features and classification stage, which is directly related to both the input and source/reference images quality in the face evaluation procedure. The initial step face extraction includes face detection and features grabbing from the image. The proposed model as shown in Fig. 6 takes input either from a live camera or from an offline video sequence then it takes each frame from input and runs a face detection system on it (68-point or 5-point features) King (2009); Kazemi and Sullivan (2014); Amato et al. (2018).

Table 2 Classification layers details

Layer Name		Output Shape	Activation Function
Single Model	Combine Model		
lstm_1	lstm_1	$(input\_shape, 512)$	tanh
–	dense_1	$(input\_shape, 2048)$	relu
dense_1	dense_2	$(input\_shape, 1024)$	relu
–	dense_3	$(input\_shape, 512)$	sigmoid
dense_2	dense_4	$(input\_shape, 50)$	sigmoid
dense_3	dense_5	$(input\_shape, 2)$	softmax

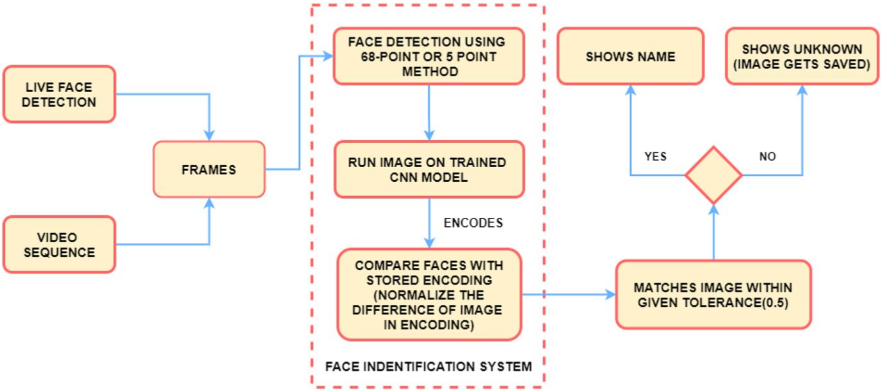
LIVE FACE DETECTION
VIDEO SEQUENCE
FRAMES
FACE DETECTION USING 68-POINT OR 5 POINT METHOD
RUN IMAGE ON TRAINED CNN MODEL
ENCODES
COMPARE FACES WITH STORED ENCODING (NORMALIZE THE DIFFERENCE OF IMAGE IN ENCODING)
FACE INDENTIFICATION SYSTEM
MATCHES IMAGE WITHIN GIVEN TOLERANCE(0.5)
YES
NO
SHOWS NAME
SHOWS UNKNOWN (IMAGE GETS SAVED)

Fig. 6 The figure illustrates the model architecture for individual identification in violence frame

The 5-point feature is used wherever the 68-point feature cannot be used due to the face not being clearly positioned. The 5-point feature uses five specific feature points which are the Left eye centroid, Right eye centroid, the bottom of the Nose, Left and Right parts of the mouth. These 5-point features are linked by five nodal distances: one between the left and right eye centroid, and the other between the rest of the four-point features and the nose.

For face recognition, the model stores these facial features in a feature vector and passes them to trained CNN models. Now, it compares the facial encoding with that stored in the database (as given in Eq. 6). If these images match with each other within a given tolerance, then it displays the label related to that.

$$Tolerance(\tau) = \|I_1 - I_2\|_2 = \sqrt{\sum_{(i,j)} (I_{1(i,j)} - I_{2(i,j)})^2} \tag{6}$$

$$where\|v\|_2 = L2 - Norm$$

where I_1 is the encoding of the image stored in a database and I_2 is the encoding of the image generated by the CNN model on the image that is captured from a frame of video. The above operation is called L2-Norm. It is the square root of the sum of squares of all the values of the vector.

5 Result

This section describes the details of the experiments and the classification model's performance used to detect violent individuals. The experimental setup describes the flow chart that how the data flows when using the ImageNet model for Transfer Learning. The model input and process 40 labeled video frames sequence in batch with the selected ImageNet model. Instead of selecting the final layer, immediately before the final layer of classification of the model is selected. A cache file is used to save the so-called Transfer Values. A cache file is used for the reason that the time taken by an image to process with the ImageNet model is time-consuming. A lot of time is saved by caching the transfer values if, each image is processed many times. Following storing the transfer values to a cache file and processing every one of the videos through the model, the transfer values are now used as the input to the LSTM neural network. The LSTM network is trained using the classes from the datasets (Violence, No-Violence). Using the transfer-values as a roadmap from the ImageNet model, the LSTM network learns to classify images taking into account the 40 frames of the video. If any of them detects violence, the video will be classified as violent. The efficient algorithm of face detection is proposed and applied to violent video which gives better performance accuracy in face recognition. Hence, the algorithm gives better and a much more verified output for violent individual recognition. The training has been done on Nvidia DGX V-100 with the following specifications: 8X NVIDIA Tesla V100 16 GB/GPU 40,960, 5,120 Tensor Cores, 4X 1.92 TB SSDs, 512 GB RAM, and 20Core Intel Xeon E5-2698 v4 2.2 GHz. The same training dataset was used to train all of the models. The accuracy metrics of the training phase are reported and observed that accuracy is not converging to any extend after 100 epochs for most models. Therefore the 100 epochs are taken as a standard for the model. Features extraction plays an important role in prediction models.

5.1 Model analysis

This section describes the details of the experiments and the classification model's performance used for detection of violence and face identification of violent individuals.

5.1.1 Violence detection model

The ImageNet model is provided with a resized video frames pipeline. The shape of the tensors expected as input by the respective pre-trained model is observed and accordingly, the frames are resized as img_size x img_size x num_channels. The number of channels(num_channels) is defined as 3(RGB). The dimension of the transfer values is taken into account while defining the architecture of the LSTM network. From each frame, the ImageNet model network obtains as output a vector of (transfer_value_size) transfer value. Since 40 frames for each video are processed, 40 x (transfer_value_size) values per video are obtained. The first input dimension of the LSTM network is the number of frames per video, which is set to 40. The second is the size of the features vector (transfer values). The fine-tuned LSTM network on transfer values with batch size 500. The default learning rate(0.01) and the default value of momentum are used. To obtain the optimal result, a 100 epochs training scheme is adopted based on the experiment. The loss function and optimizer used in this model are mean_squared_error and adam respectively.

Figure 7 demonstrates predictions based on violent and non-violent video frames. We have used the NASNetLarge model to obtain frame-based predictions. The

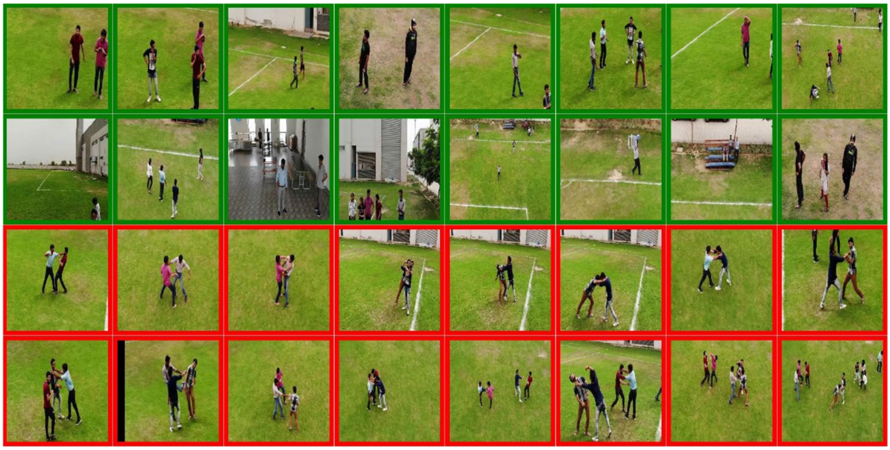

Fig. 7 Prediction on non-violent (Row 1, Row 2) and violent video (Row 3, Row 4) frames

Table 3 Performance evaluation after violence video detection from different proposed model architecture

Models	Precision	Recall/Sensitivity	Specificity	AUC	G-Mean	F1 Score
VGG16+LSTM	0.9625	0.7938	0.9434	0.9323	0.8654	0.8743
VGG19+LSTM	0.8543	0.8947	0.8378	0.9494	0.8658	0.8718
InceptionV3+LSTM	0.8752	0.8536	0.8529	0.9283	0.8532	0.8642
DenseNet201+LSTM	0.8875	0.8875	0.8714	0.9366	0.8794	0.8875
ResNet101V2+LSTM	0.9543	0.8444	0.9333	0.9162	0.8778	0.8941
MobileNet+LSTM	0.9625	0.8556	0.9562	0.9375	0.9015	0.9058
NASNetLarge+LSTM	0.9254	0.8506	0.9047	0.9373	0.8772	0.8862
(VGG16+VGG19)+LSTM	0.8375	0.8272	0.8116	0.8885	0.8193	0.8323
(ResNet50+ResNet152V2)+LSTM	0.9575	0.9872	0.9452	0.9794	0.9658	0.9681
(InceptionV3+ResNet101V2)+ LSTM	0.9875	0.9634	0.9852	0.9842	0.9742	0.9753

proposed architecture has performed admirably on predicting the non-violent and violent videos correctly. To represent the performance evaluation for every model architecture the represented metrics are chosen: Precision, Recall/Sensitivity, Specificity, G-Mean, AUC, and F1 Score shown in Table 3.

$$G - Mean = \sqrt{Sensitivity \times Specificity} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The performance of the proposed models is evaluated using the standard metrics described above. True positive and true negative values are represented by TP and TN, whereas false positive and false negative values are represented by FP and FN respectively.

Table 4 Comparison of overall accuracy report of CNN models

Models	Training(%)	Testing(%)
VGG16	84.28	82.67
VGG19	88.06	76.45
InceptionV3	86.35	77.33
DenseNet201	93.83	82.98
ResNet101V2	85.58	80.32
MobileNet	82.52	78.54
NASNetLarge	85.05	72.67

Table 5 Comparison of overall accuracy report of models

Models	Training(%)	Testing(%)
VGG16+LSTM	92.43	85.33
VGG19+LSTM	93.97	88.67
InceptionV3+LSTM	99.42	85.33
DenseNet201+LSTM	99.77	88.42
ResNet101V2+LSTM	96.58	90.67
MobileNet+LSTM	99.88	89.33
NASNetLarge+LSTM	98.17	91.33
(VGG16+VGG19)+LSTM	90.75	82.64
(ResNet50+ResNet152V2)+LSTM	99.33	96.67
(InceptionV3+ResNet101V2)+LSTM	98.47	97.33

The experiment of features extraction from the frames is carried using seven ImageNet models VGG16, VGG19, ResNet101V2, DenseNet201, InceptionV3, MobileNet, NASNetLarge. Along with the seven models, three combinations of two similar types of models are also tried for extracting the features (Table 4).

The extracted features are provided as input into an LSTM/Dense Layer network. On the retrieved features, ten distinct models were trained and evaluated. Table 5 outlines the accuracy of the experimented models training and testing dataset using CNN + LSTM model. Combining features extracted from the models to LSTM, an architecture is enabled where previous frames information can remember. Therefore, more informed decisions about each frame can be taken by the model and ultimately, about the video. In order to showcase the comparative study for CNN and CNN + LSTM models the accuracy of the experimented models training and testing dataset using CNN models is presented in Table 4.

Additionally, the ROC curve which helps in the analysis of the performance of the proposed technique is shown in Figs. 8, 9, 10, 11 and 12. The AUC value of VGG16, VGG19, ResNet101V2, DenseNet201, InceptionV3, MobileNet, NASNetLarge and (VGG16+VGG19) combination model with LSTM ranged from 0.88 to 0.94, whereas the AUC value of (InceptionV3+ResNet101V2) and (ResNet50+ResNet152V2) combination model with LSTM are 0.98 and 0.97 respectively with illustrates the efficient performance of proposed technique.

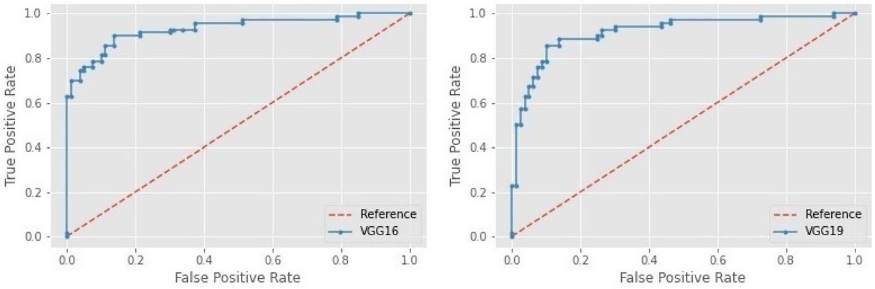

Fig. 8 Violent class ROC curve using VGG16 and VGG19 models with LSTM architecture respectively

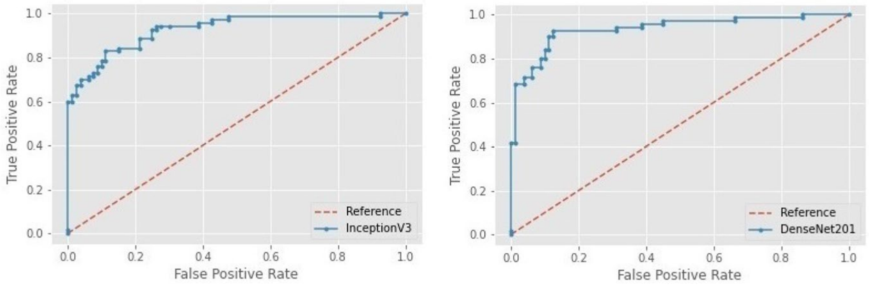

Fig. 9 Violent class ROC curve using InceptionV3 and DenseNet201 models with LSTM architecture respectively

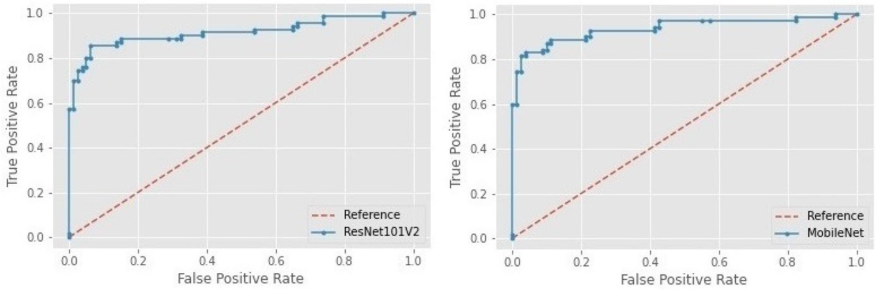

Fig. 10 Violent class ROC curve using ResNet101V2 and MobileNet models with LSTM architecture respectively

5.1.2 Face identification model

The image frames are obtained from the drone-captured high definition camera from the height of 3–6 meters having run time between 5 seconds to 34 seconds videos (aerial images) and unconstrained environment for the noticeable task of face recognition. With the proposed method, the face detection and face

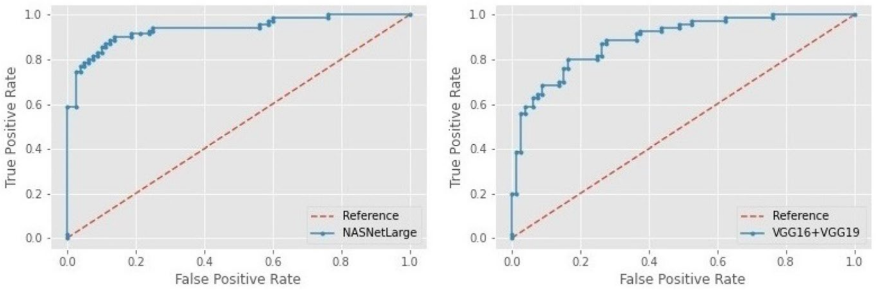

Fig. 11 Violent class ROC curve using NASNetLarge and (VGG16+VGG19) models with LSTM architecture respectively

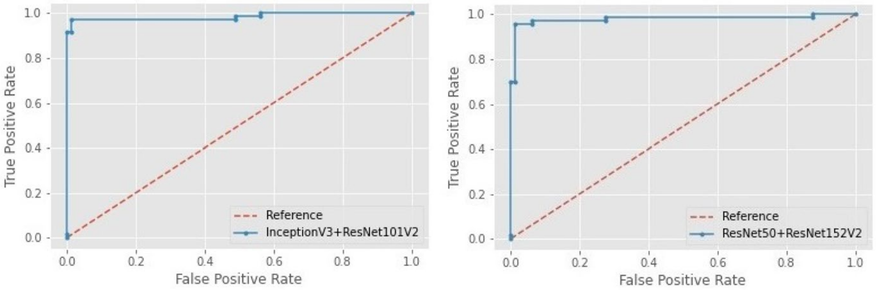

Fig. 12 Violent class ROC curve using (InceptionV3+ResNet101V2) and (ResNet50+ResNet152V2) models with LSTM architecture respectively

recognition model is performed to obtain more accurate and observable facial features to make the comparison more precise and obtain accurate results. The performance assessment of the face recognition rank calculation in the paper for individuals performing various violent activities is presented in Fig. 13.

The face identification experiment is performed on different videos of resolution 1980 × 1020 and 30 frames per second. In these videos, the subjects are generally depicted performing different violent activities. For anchor images, 2–13 different people's face images are taken having resolution ranging from 440 × 500 to 2400 × 3000 (ultra high definition) JPEG images. The different tolerance values ranging from 0.6 to 0.8 are experimented with and observed that best results on 0.6 tolerance levels are achieved. The graph in Fig. 13 refers to the accuracy of the model at different ranks. Rank-1 accuracy means the face detected by the model is that of the true face label. Rank-2 accuracy means the true face label must be contained in the face label with top-2 most probabilities and so on. Some improvements are observed in accuracy from Rank-1 to Rank-2, but minor improvements after that.

The graph in Fig. 14 represents the Recall vs Tolerance graph, where recall is defined as the percentage of recognized faces (correctly/incorrectly) to the overall

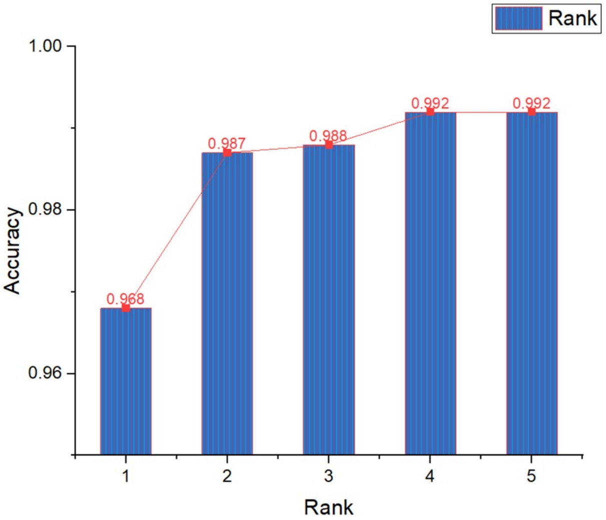

Fig. 13 Rank versus face identification accuracy graph

count of people's faces (unknown + recognized). As the tolerance is increased, the acceptable range of difference between the vector of the known face from the database and the detected face also increases. This leads many unknown faces to be matched with a known face, which contributed to an increase in recall.

The graph in Fig. 15 represents the Tolerance vs Accuracy graph, where accuracy is defined as the percentage of correctly classified faces to the total recognized faces. The downward trend of accuracy wrt tolerance can be explained by the fact as the

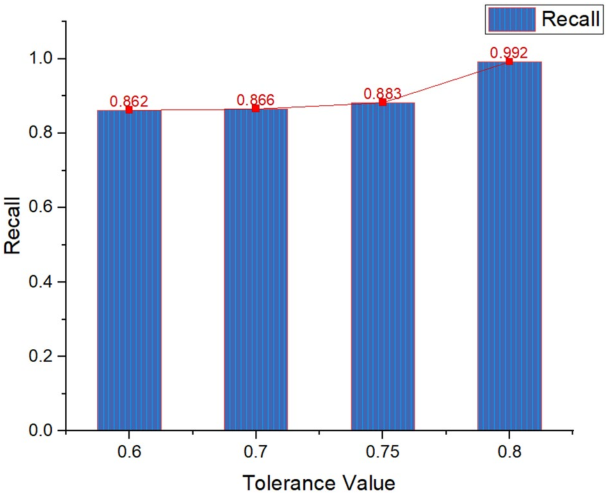

Fig. 14 Recall versus tolerance graph

tolerance is increased, the acceptable range of difference between the vector of the known face from the database and detected face also increases. This leads the model to misclassify the detected face which in turn decreases accuracy.

The comparative study for the face identification model with the existing state-of-the-art methods on the proposed dataset is represented in Table 6. The result of recognized faces is shown in Fig. 16. The results in Fig. 16 show that the proposed violent individual identification system in a violent frame is very stable to be used in a regulated drone surveillance environment in real life as compared to other methods and makes an incremental contribution to the improvement of face recognition algorithm.

5.2 Runtime analysis

The experimented pre-trained models' execution times are presented in milliseconds in Table 7. It takes into account the time required for feature extraction and classification. The transfer value is extracted first, followed by LSTM features. These features then are fed into dense neural network models for classification. As a result, the suggested violence recognition model's runtime involves two tasks: (1) The time required for extracting transfer values and computing LSTM features for each frame is included in the feature extraction time. (2) The time required to classify per frame for classification models is included in the classification time. This runtime analysis is not universal, though. These results are based on a particular environmental structure. The Nvidia Tesla T4 GPU in Google Colab is used for the runtime evaluation. It cannot be determined which model takes the least amount of time based on the table, but it can be deduced that hybrid models have taken more time than comparable single models of a given technique.

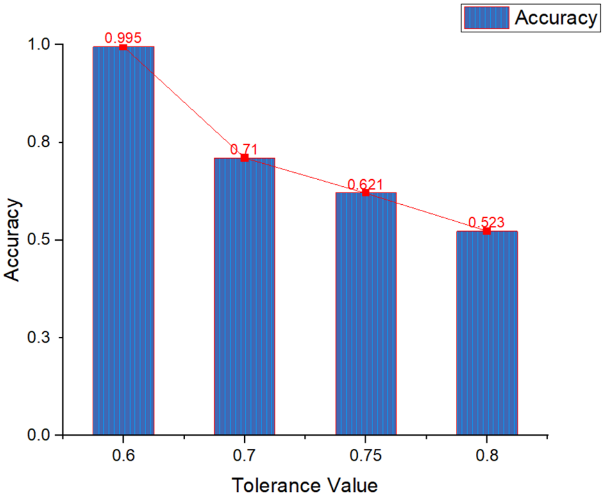

Fig. 15 Face identification accuracy versus tolerance graph

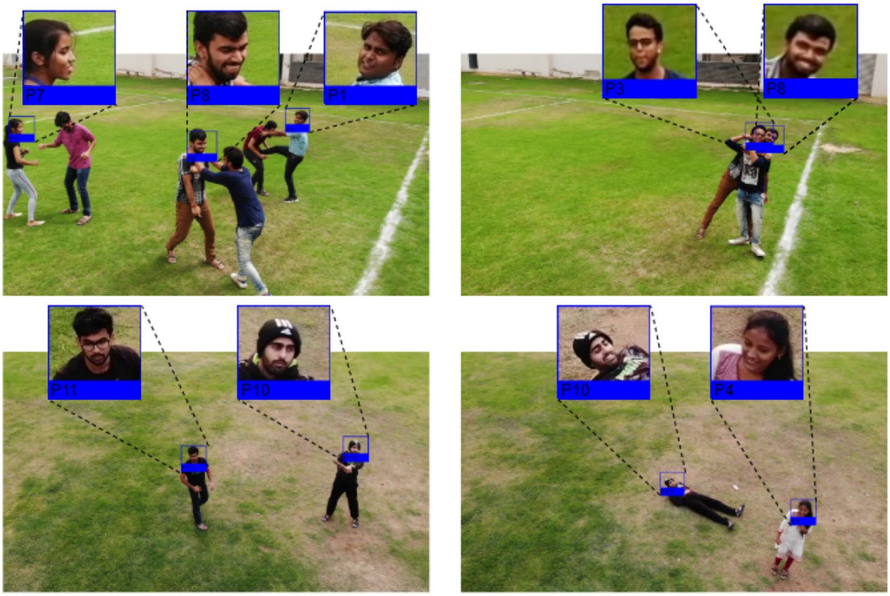

Fig. 16 Identification of individuals involved in violence frames

Table 6 The comparative study with existing state of the art methods for the face identification model

Architecture	Accuracy(%)
HOG	97.85
VGG16	98.37
Inception	98.81
Proposed (ResNet-28)	99.20

Table 7 Runtime evaluation in milli seconds of proposed violence non-violence classification work with different pre-trained models and LSTM architecture

Models	Time per frame
	Feature extraction + classification
VGG16+LSTM	19.82 ± 2.36
VGG19+LSTM	33.34 ± 5.23
InceptionV3+LSTM	22.78 ± 2.34
DenseNet201+LSTM	22.75 ± 2.42
ResNet101V2+LSTM	23.42 ± 2.04
MobileNet+LSTM	16.68 ± 3.44
NASNetLarge+LSTM	49.27 ± 2.14
(VGG16+VGG19)+LSTM	39.58 ± 3.61
(ResNet50+ResNet152V2)+LSTM	67.44 ± 8.04
(InceptionV3+ResNet101V2)+LSTM	68.32 ± 8.16

Table 8 represents the time of execution for the identification of an individual in a violent frame in milliseconds. It takes into account the amount of time for a different level of tolerance (0.6, 0.7, 0.75, 0.8) and recorded time for the complete recognition mechanism (Detect + Recognize + BB creation). From the runtime analysis, the per-frame processing could be around 1311 ms–1372 ms. However, the proposed's runtime analysis achieves its performance in real-time.

6 Conclusion and future work

The research has proposed unique network architecture for violent interaction along with aggressive individual detection in the drone captured surveillance video. This study demonstrates that creating a deep network from the scratch leads to overfitting problems. An alternate transfer learning technique is therefore used, whereby existing models are pre-trained on millions of ImageNet images before being applied to new datasets. The network is combined with long short-term memory networks that outperform straightforward convolutional neural networks. The accuracy also increases to a certain margin as compared to pure transfer learning models. Face recognition using drone-captured surveillance videos was accomplished by preprocessing the input face images for face detection, to have better image features and using the CNN model to improve the face recognition algorithm and experiment results. The datasets used in the research and testing, as well as the methods required in data pre-processing, have been described in depth. The system also incorporates dynamic scene patterns as well as the most discriminating features for abrupt camera motions. The system provides a simple graphical user interface highlighting the border with two separate colors for violence detection along with violent individual identification to interact with the deep learning model.

The proposed system accurately detects violence as well as the violent individual causing the trouble. Considering the scope of the experiments, in the near future explosive weapons, fire, and ammunition detection can also be implemented using the proposed work. The recognition and extraction of the face images from the poses that are not clearly positioned can be improved by exploring the area of 3D pose detection for the faces. Database creation where the timestamp of suspicious and violent activities with violent individuals recognized by the proposed method will

Table 8 Runtime evaluation in milli seconds of indentifying individuals in violence frame

	Frames	Time (0.6)	Time(0.7)	Time(0.75)	Time(0.8)	Average Time	Time per Frame
Set 1	254	348.27	347.21	346.25	345.98	410.43	1615.86
Set 2	333	449.78	447.11	446.34	448.49	531.18	1595.14
Set 3	83	110.35	107.87	107.58	109.39	129.55	1560.81
Set 4	112	150.32	148.63	149.29	150.18	177.61	1585.76
Set 5	157	203.43	199.62	200.71	202.55	240.83	1533.93

be recorded and evidence for legal authorities to take appropriate measures, relevant screenshots were taken during surveillance can be retained in the dataset.

References

Amato, G., Falchi, F., Gennaro, C., Vairo, C.: A comparison of face verification with facial landmarks and deep features. In: proceedings of the 10th international conference on advances in multimedia (MMEDIA 2018) (c), 1–6 (2018)

Aydin, B.: Public acceptance of drones: knowledge, attitudes, and practice. Technol. Soc. **59**, 101180 (2019). https://doi.org/10.1016/j.techsoc.2019.101180

Baba, M., Gui, V., Cernazanu, C., Pescaru, D.: A sensor network approach for violence detection in smart cities using deep learning. Sensors (Switzerland) (2019). https://doi.org/10.3390/s19071676

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Lecture notes in computer science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6855 LNCS, pp. 332–339 (2011). https://doi.org/10.1007/978-3-642-23678-5_39

Bindemann, M., Fysh, M.C., Sage, S.S.K., Douglas, K., Tummon, H.M.: Person identification from aerial footage by a remote-controlled drone. Sci. Rep. **7**(1), 1–10 (2017). https://doi.org/10.1038/s41598-017-14026-3

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017, vol. 2017-Janua (2017). https://doi.org/10.1109/CVPR.2017.143

Cheng, W.H., Chu, W.T., Wu, J.L.: Semantic context detection based on hierarchical audio models. In: proceedings of the 5th ACM SIGMM international workshop on multimedia information retrieval, MIR 2003, pp. 109–115 (2003). https://doi.org/10.1145/973264.973282

Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV workshops 2009, pp. 1282–1289 (2009). https://doi.org/10.1109/ICCVW.2009.5457461

Dandagpl Vishwajit, Hiemanshu Gautam, Akshay Ghavale, Radhika Mahore, Sonewar., P.A.: IRJET-review of violence detection system using deep learning. Int. Res. J. Eng. Technol. (IRJET) (2019)

Deeba, F., Ahmed, A., Memon, H., Dharejo, F.A., Ghaffar, A.: LBPH-based enhanced real-time face recognition. Int. J. Adv. Comput. Sci. Appl. 10(5), 274–280 (2019). https://doi.org/10.14569/ijacsa.2019.0100535

Deniz, O., Serrano, I., Bueno, G., Kim, T.K.: Fast violence detection in video. In: VISAPP 2014 - proceedings of the 9th international conference on computer vision theory and applications, vol. 2, pp. 478–485 (2014). https://doi.org/10.5220/0004695104780485

Ding, C., Fan, S., Zhu, M., Feng, W., Jia, B.: Violence detection in video by using 3D convolutional neural networks. In: Lecture notes in computer science (including Subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 8888, pp. 551–558 (2014). https://doi.org/10.1007/978-3-319-14364-4_53

Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017) arXiv:1411.4389. https://doi.org/10.1109/TPAMI.2016.2599174

Fu, E.Y., Leong, H.V., Ngai, G., Chan, S.C.F.: Automatic fight detection in surveillance videos. Int. J. Pervasive Comput. Commun. **13**(2), 130–156 (2017). https://doi.org/10.1108/IJPCC-02-2017-0018

Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Lecture notes in computer science (including Subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 3955 LNAI, pp. 502–507 (2006). https://doi.org/10.1007/11752912_55

Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A multi-class audio classification method with respect to violent content in movies using Bayesian Networks. In: 2007 IEEE 9Th international workshop on multimedia signal processing, MMSP 2007 - proceedings, pp. 90–93 (2007). https://doi.org/10.1109/MMSP.2007.4412825

Goya, K., Zhang, X., Kitayama, K., Nagayama, I.: A method for automatic detection of crimes for public security by using motion analysis. In: IIH-MSP 2009 - 2009 5th international conference on intelligent information hiding and multimedia signal processing, pp. 736–741 (2009). https://doi.org/10.1109/IIH-MSP.2009.264

Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: proceedings of the international joint conference on neural networks, vol. 2016-October, pp. 381–388 (2016). https://doi.org/10.1109/IJCNN.2016.7727224

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 2016-December, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. Adv. Neural Inf. Process. Syst. **2015**, 2017–2025 (2015)

Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013). https://doi.org/10.1109/TPAMI.2012.59

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F.: Large-scale video classification with convolutional neural networks. In: proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 1725–1732 (2014). https://doi.org/10.1109/CVPR.2014.223

Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 1867–1874 (2014). https://doi.org/10.1109/CVPR.2014.241

Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. In: proceedings of the IEEE computer society conference on computer vision and pattern recognition **2016-Decem**, 4873–4882 (2016) arXiv:1512.00596. https://doi.org/10.1109/CVPR.2016.527

King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)

Kohonen, T.: Self-organization and associative memory, vol. 8. Springer, Berlin (2012)

Laptev, I., Lindeberg, T.: On space-time interest points. Int. J. Comput. Vision **64**(2), 107–123 (2005)

Li, X., Chuah, M.C.: SBGAR: Semantics Based Group Activity Recognition. In: proceedings of the IEEE international conference on computer vision, vol. 2017-Octob, pp. 2895–2904 (2017). https://doi.org/10.1109/ICCV.2017.313

Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10), 1728–1740 (2008). https://doi.org/10.1109/TPAMI.2008.73

Mumtaz, A., Sargano, A.B., Habib, Z.: Violence detection in surveillance videos with deep network using transfer learning. In: proceedings - 2018 2nd European conference on electrical engineering and computer science, EECS 2018, pp. 558–563 (2018). https://doi.org/10.1109/EECS.2018.00109

Naik, A.J., Gopalakrishna, M.T.: Violence detection in surveillance video-a survey. Int. J. Latest Res. Eng. Technol. (IJLRET) **2017**, 11–17 (2017)

Ordóñez, F.J., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors (Switzerland) (2016). https://doi.org/10.3390/s16010115

Penmetsa, S., Minhuj, F., Singh, A., Omkar, S.N.: Autonomous UAV for suspicious action detection using pictorial human pose estimation and classification. Electron. Lett. Comput. Vision Image Anal. (2014). https://doi.org/10.5565/rev/elcvia.582

Ramanathan, M., Yau, W.Y., Teoh, E.K.: Human action recognition with video data: research and evaluation challenges. IEEE Trans. Hum. Mach. Syst. (2014). https://doi.org/10.1109/THMS.2014.2325871

Saypadith, S., Aramvith, S.: Real-time multiple face recognition using deep learning on embedded GPU system. In: 2018 Asia-Pacific signal and information processing association annual summit and conference, APSIPA ASC 2018 - proceedings, pp. 1318–1324 (2019). https://doi.org/10.23919/APSIPA.2018.8659751

Singh, A., Patil, D., Omkar, S.N.: Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. In: IEEE computer society conference on computer vision and pattern recognition workshops, vol. 2018-June, pp. 1710–1718 (2018). https://doi.org/10.1109/CVPRW.2018.00214

Soliman, M.M., Kamal, M.H., Nashed, M.A.E.-M., Mostafa, Y.M., Chawky, B.S., Khattab, D.: Violence recognition from videos using deep learning techniques. (2019). https://doi.org/10.1109/ICICIS46948.2019.9014714

Srivastava, A., et al.: Recognizing human violent action using drone surveillance within real-time proximity. J. Real Time Image Process. (2021). https://doi.org/10.1007/s11554-021-01171-2

Sumon, S.A., Goni, R., Hashem, N.B., Shahria, T., Rahman, R.M.: Violence detection by pretrained modules with different deep learning approaches. Vietnam J. Comput. Sci. **07**(01), 19–40 (2020). https://doi.org/10.1142/s2196888820500013

Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W.: Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors (Switzerland) (2019). https://doi.org/10.3390/s19112472

van der Spoel, E., Rozing, M.P., Houwing-Duistermaat, J.J., Eline Slagboom, P., Beekman, M., de Craen, A.J.M., Westendorp, R.G.J., van Heemst, D.: Siamese neural networks for one-shot image recognition. ICML - deep learning workshop **7**(11), 956–963 (2015) arXiv:arXiv:1011.1669v3

Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021) arXiv:1804.06655. https://doi.org/10.1016/j.neucom.2020.10.081

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: Lecture notes in computer science (including Subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 9912 LNCS, pp. 20–36 (2016). https://doi.org/10.1007/978-3-319-46484-8_2

Wu, B., Ai, H., Huang, C., Lao, S.: Fast rotation invariant multi-View face detection based on real adaboost. In: proceedings - Sixth IEEE international conference on automatic face and gesture recognition, pp. 79–84 (2004). https://doi.org/10.1300/J083v43n02_06

Xu, M.: Robust object detection with real-time fusion of multiview foreground silhouettes. Opt. Eng. **51**(4), 047202 (2012). https://doi.org/10.1117/1.oe.51.4.047202

Zaheer, M.Z., Kim, J.Y., Kim, H.G., Na, S.Y.: A preliminary study on deep-learning based screaming sound detection. In: 2015 5th international conference on IT convergence and security, ICITCS 2015 - proceedings (July) (2015). https://doi.org/10.1109/ICITCS.2015.7292925

Zajdel, W., Krijnders, J.D., Andringa, T., Gavrila, D.M.: CASSANDRA: audio-video sensor fusion for aggression detection. In: 2007 IEEE conference on advanced video and signal based surveillance, AVSS 2007 proceedings (2007). https://doi.org/10.1109/AVSS.2007.4425310

Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H.: Real-Time Action Recognition with Enhanced Motion Vector CNNs. In: proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 2016-December, pp. 2718–2726 (2016). https://doi.org/10.1109/CVPR.2016.297

Zhou, P., Ding, Q., Luo, H., Hou, X.: Violent interaction detection in video based on deep learning. J. Phys. Conf. Ser. (2017). https://doi.org/10.1088/1742-6596/844/1/012044

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

Authors and Affiliations

Anugrah Srivastava[1] · **Tapas Badal**[1] · **Pawan Saxena**[1] · **Ankit Vidyarthi**[2] · **Rishav Singh**[3]

✉ Rishav Singh
rishav.singh@nitdelhi.ac.in

Anugrah Srivastava
AS5271@bennett.edu.in

Tapas Badal
Tapas.Badal@bennett.edu.in

Pawan Saxena
pawan.saxena@bennett.edu.in

Ankit Vidyarthi
dr.ankit.vidyarthi@gmail.com

[1] Computer Science Engineering Department, Bennett University, Greater Noida, India

[2] Department of CSE and IT, Jaypee Institute of Information Technology, Noida, India

[3] Department of Computer Science and Engineering, National Institute of Technology, Delhi, India