# Anomalous behavior detection using Spatio temporal Feature and 3D CNN model for Surveillance

by

Jannatun Nahar
18101291
Zarin Tasnim Promi
18101589
Jannatul Ferdous
18101565
Fatin Ishrak
21301716
Ridah Khurshid
18101683

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2022

# Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Jannatun Nahar
18101291

Zarin Tasnim Promi
18101589

Jannatul Ferdous
18101565

Fatin Ishrak
21301716

Ridah Khurshid
18101683

# Approval

The thesis/project titled "Anomalous behavior detection using Spatio temporal Feature and 3D CNN model for Surveillance " submitted by

1. Jannatun Nahar (18101291)
2. Zarin Tasnim Promi (18101589)
3. Jannatul Ferdous (18101565)
4. Fatin Ishrak (21301716)
5. Ridah Khurshid (18101683)

Of Spring , 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January, 2022.

**Examining Committee:**

Supervisor: (Member)

Dr. Amitabha Chakrabarty, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator: (Member)

Dr. Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department: (Chair)

Dr. Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

# Ethics Statement

We thus declare that this thesis is based on the findings of our research. All other sources of information have been acknowledged in the text. This thesis has not been previously submitted, in whole or in part, to any other university or institute for the granting of any degree.

# Abstract

Anomalous and violent action detection has become an increasingly relevant topic and active research domain of computer vision and video processing, within the past few years. It has many proposed solutions by the researchers and this field attracted new researchers to contribute in this domain. Furthermore , the widespread use of cameras used for security purposes in big modern cities has also allowed researchers to research and examine a vast amount of information so that autonomous monitoring can be executed. Adding effective automated violence unearthing to videotape security or multimedia content watching technologies (CCTV) would make the task of carpoolers, walk organizations, and those who are in control of social media activity monitoring much easier. We present a new deep scholarship skeleton for determining whether a videotape is violent or not, based on a suited version of DenseNet , and a bidirectional convolutional LSTM module that allows unscrambling pointed Spatio-temporal features in this paper. In addition, ablation research of the input frames was carried out, comparing thick optic outpouring and touching frames. Throughout the paper, we analyze various strategies to detect violence and their classification in use. Furthermore, in this paper, we detect violence using the Spatio-temporal feature with 3D CNN which is a DL violence detection framework, specially better for crowded places. Finally, we used embedded devices like Jetson Nano to feed with dataset and test our model and evaluate. We want a warning sent to the local police station or security agency as soon as a violent activity is detected so that urgent preventive measures can be taken. We have worked with various benchmark datasets where in one dataset, multiple models achieved a test accuracy of 100 percent, making them invincible. Furthermore, for a different dataset our models have shown 99.50% and 97.50% accuracy rates. We also did a cross dataset experiment in models which also showed pretty good results of higher than 60%. The overall results we got suggests that our system has a viable solution to anomalous behavior detection.



**Keywords:** Human Activity Recognition; Deep learning; DenseNet; 3D bi-LSTM; Spatio-temporal; Violence; 3D CNN; TensorFlow; Keras; Jetson Nano ;

# Dedication

We would like to dedicate this thesis to our loving parents and all of the wonderful academics we met and learned from while obtaining our Bachelor's degree and specially our beloved supervisor Dr.Amitabha Chakrabarty.

# Acknowledgement

First and foremost, we would like to express our gratitude to our Almighty for allowing us to conduct our research, put out our best efforts, and finish it. Second, we would like to express our gratitude to our supervisor, Dr. Amitabha Chakrabarty sir, for his input, support, advice, and participation in the project. We are grateful for his great supervision, which enabled us to complete our research effectively. Furthermore, we would like to express our gratitude to our faculty colleagues, family, and friends who guided us with kindness, inspiration, and advice. Last but not least, we are grateful to BRAC University for allowing us to do this research and for allowing us to complete our Bachelor's degree with it.

# Table of Contents

# List of Figures

# List of Tables

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$AUC$ Area Under Curve

$CNN$ convolutional neural network

$COV - Net$ Covarience Network

$DL$ Deep Learning

$HAR$ Human Activity Recognition

$HFs$ Hockey Fights

$HHMM$ Hierarchical hidden Markov model

$KNN$ K-Nearest Neighbour

$LSTM$ Long short-term memory

$MFs$ Movie Fights

$ReLU$ Rectified Linear Unit

$RWF$ Real Life Violence Situations

$sHOT$ Shot Learning Algorithm

$SVM$ Support Vector Machine

$VDT$ Violence Detection Techniques

$VFs$ Violent Flows

# Chapter 1

# Introduction

At present era, the issue of identifying anybody's movement through video has brought a significant role in the field of computer vision. However, the identification in detecting violence that conducts, received fewer attention in compared to any other human activities. Increase in threats lead the world to use CCTVs to monitor people everywhere in cities and towns. To ensure safety of citizens, human attacks and fights are examples of situations where crime monitoring systems are required. Before all else, we concentrate on the strategies for recognising and detecting violence in surveillance recordings. One of our goals is to determine whether and then when violence happens in a video. Various strategies for recognising and detecting violence have been proposed in the last decade. For individual fight detection, A. Datta, M. Shah et al. al.[3]has used motion trajectory data with limb orientation information. Nevertheless, one of the disadvantages of this method is that it necessitates exact segmentation, which is hard to achieve in real-world videos. Over recent years, with the evolution of technology in the firing line of computer vision, a large number of new techniques have arisen, attracting the interest of researchers due to its variety of applications[19][40]. In South Korea, for example, about 954,261 CCTVs were installed in public spaces in 2017, up 12.9 percent over the previous year [51].

Thus, this is the purpose of focusing on video based violence detection methods to ensure security in public places. Lack of people and ignored dangers are issues that develop when operators fail to discover objects or actions after 20 minutes of watching a CCTV system. This necessitates advancements in automated systems for the identification of violent acts or the detection of guns. Safety has always been a concern in all aspects of daily life. Currently, there are socioeconomic differences., as well as the global economic crisis, have resulted in a rise in violence, as well as the recording and dissemination of such acts. As a result, it's critical to build automated systems to detect these acts and increase security teams' reactivity, as well as the control of social media content and the examination of multimedia data for an age restriction.

Violence detection has been a well-liked topic in recent human activity recognition research, particularly in video surveillance. One amongst the difficulties with human activity recognition normally is the classification of human action in real time, almost instantaneously after the action has taken place. This difficulty escalates when dealing with surveillance video for a variety of things including the standard 1 of surveillance footage is diminished, lighting isn't always guaranteed, and there's

generally no contextual information which will be accustomed to ease detection of actions and classification of violent versus non-violent. Furthermore, for violent scene recognition to be helpful in real-world surveillance applications, the identification of violence must be swift in order to allow for prompt intervention and backbone.

In addition to poor video quality, violence can occur in any given setting at any time of day, therefore, an answer must be robust to detect violence irrespective of the conditions. Some settings for video surveillance where violence detection are often applied include the inside and exterior of buildings, in traffic, or on police body cameras.

Among other things, intentional violence, particularly in person-to-person violence, is one of the subjects of this research. This eliminates unintentional acts such as road accidents or rough play in sports. Strokes and fists are related with abrupt movements, so they should be avoided and also a Spatio-temporal analysis is required to visually discern violence in recordings. Violent behavior might be mistaken for other sorts of behavior, resulting in false positives. A video with hardly any activity, including quick gestures seen during cardiac cpr, Can be comprehended in a wrong way as striking.To prevent this, it's critical to look at the entire video's temporal context before and after the action.

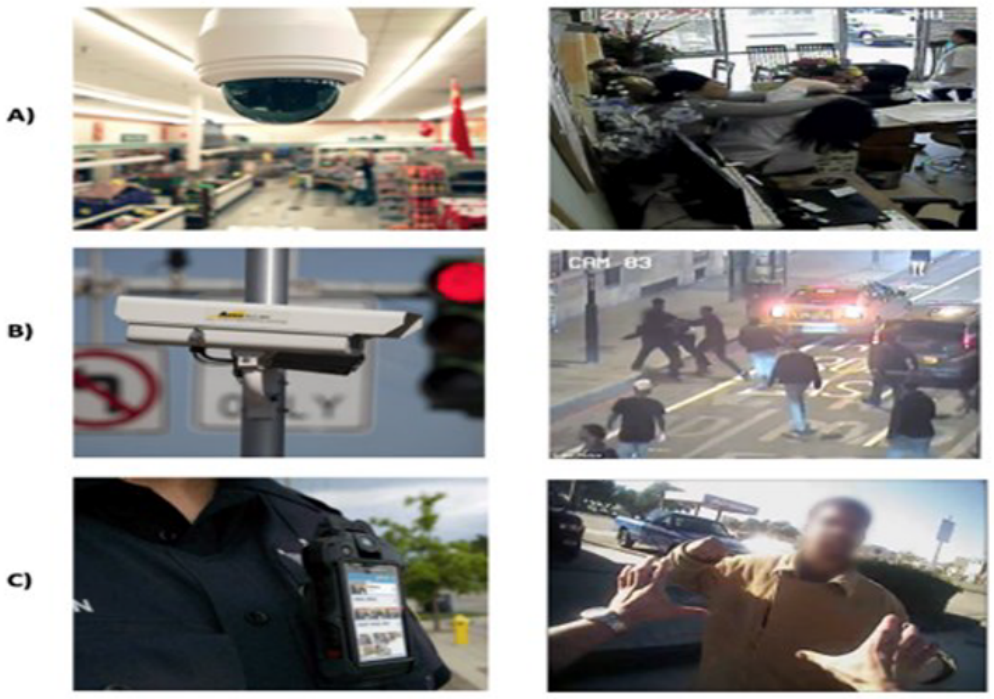


Figure 1.1: Different scenarios where real time violence detection will be applicable and corresponding scenes with violence that should be detected (A) Interior video surveillance (B) Traffic video surveillance (C) police body cameras These use cases provide the motivation for this thesis: the flexibility to rapidly and accurately detect violence in real-time in multiple settings.

These use cases provide the motivation for this thesis: the flexibility to rapidly and accurately detect violence in real-time in multiple settings.

## 1.1 Motivation

Despite the fact that there is a large corpus of study on the subject, there has been absolutely no commercial method currently for detecting violence that combines AI technology and human operators. In terms of any job quality, this is undoubtedly necessary. As a result, the operators who must view this type of video will be less stressed, allowing them to focus on more productive activities in some circumstances. Beyond everything, it's a matter of not being able to accomplish the task accurately and effectively owing to a fundamental limitation in the videos which are observed concurrently and along with requisite focus. The number of false positives that cause the system to be deactivated, as well as the occurrence of false negatives that indicate functionality failures, are the key roadblocks to developing a video surveillance system capable of automatically identifying violence. This is why a more accurate approach is presented, which has been tested on multiple datasets. Over and above that, because violence can take many forms, drawing general inferences from an individual and single dataset is problematic. In this sense, we believe that conducting a cross dataset analysis to see if a training is performed on one dataset can yield good results on another is crucial, and whether it is viable to put the model into production or whether it has to be refined for each scenario.

## 1.2 Problem Statement

Act recognition from visual data is one of the most difficult fields of research in the development of advanced and smart cities, particularly in surveillance applications. For law enforcement authorities to prevent crime, anomalous activity recognition is critical. Unusual activities include those that are harmful for human life or property, such as accidents, destruction of property, breaking the law, or criminal actions such as fighting or theft. To evaluate the algorithms developed in the early testing of activity recognition, different data sets containing activity conducted by one actor under controlled circumstances are used. The focus of this research has shifted to uncontrolled, realistic video data sets, which present more challenges to event detection problems, including image noise, inter and intra class variances, opacity, postures change, lens motions, and so on [41].

Human activity is collected in a series of video frames; consequently, human behavior is recognised and detected based on visual aspect and movement dynamics in a frame sequence[36]. CNNs have recently achieved outstanding results in the matter of image classification and object detection. CNN, on the other hand, only processes one image at a time. As a result, they can't be used to identify visual input in a time frame directly. As a result, the focus of the study is on 3D CNN, which can also absorb the spatiotemporal information of visual data. The next difficult task is to model the video's temporal fluctuation. If the recognition is done online, like in an actual surveillance system, it becomes extremely challenging. Traditional techniques based on trajectories, for example, are optical flow based models that are dependent [11] [37]. Moreover, End-to-End multi-stream methods[32][14] interacted with multi - 2D networks with addition of optical flow. The performance

is good but computational 3 complexity is very high due to extracting the optical flow. Therefore, these methods face challenges over large scale datasets and real-time monitoring. Alternately, other architectures of 3D CNNs, such as two stream 3D ConvNet [35], pseudo-3D CNNs [30], and MiCT-Net [43]. are used to tackle the problem of expensive calculation of spatio-temporal characteristics. Such 3D CNNs can extract spatio-temporal features directly, which improves classification performance while significantly increasing time complexity.

Convolutional neural networks (CNNs) have recently emerged as a superior tool for action recognition and security [49] [27] [50], entity monitoring and behavior identification [34] [38], video analysis[45], and disaster risk management [46] .among many computer vision techniques. We solve the difficulties by demonstrating 3D CNN-based violence detection in real-time surveillance, which is based on CNN performance in the previously indicated domain. The following bullet points outline the major difficulty and key contribution of our suggested method:

- We will discuss some classification of state-of-the-art models for better understanding and critically review their novelty using real-world datasets.
- Because of the difficulty of identifying consecutive visual patterns, detecting violence from video analysis and data is a difficult job to accomplish. Other commonly used strategies emerged from classic low-level and limited modeling characteristics that aren't particularly good at spotting highly complex patterns or dealing with time complexity concerns; they're also challenging to employ real- time monitoring. Given there are restriction in case of the present methods,we made the decision to use a deep-learning (DL) based 3D CNN model to understand complicated sequential patterns and effectively predict violence in smart cities and other places.
- The difficulty of processing large amounts of meaningless frames plagues several violence detection systems. As a result, they take up a lot of memory and take a long time. In order to overcome this constraint,in recognize anybody in video we will be utilizing a model which is a pre-trained MobileNet CNN. To achieve efficient processing, only key feeds relating to the event will be transmitted to the 3D CNN model for final prediction once the frames have been cleansed.
- Due to insufficient data from benchmarked datasets, various popular approaches for detecting violence are unable to develop effective patterns and produce low accuracy results. For identifying aggression in both indoor and outdoor situations, the 3D CNN is frequently fine-tuned utilizing publicly available benchmark datasets and the idea of transfer learning. The rate of accuracy is likewise extremely high.
- Lastly, obtaining the trained deep learning model, we will optimize the model using Deep-stream SDK, the toolkit of Nvidia. To deploy these types of devices using the HAR model, we are using development boards such as Jetson Nano. The trained model will be translated into an alternative manner based on trained parameters and topologies using these toolkits.

Moreover, we will implement emerging techniques which are lightweight in recognizing activity that can be easily integrated into image sensors and IoT systems for cost-effective surveillance.

## 1.3 Research Objective

From the past decade, different approaches have been given for HAR design methodology. HAR systems are complex and follows some subsystems shown below:

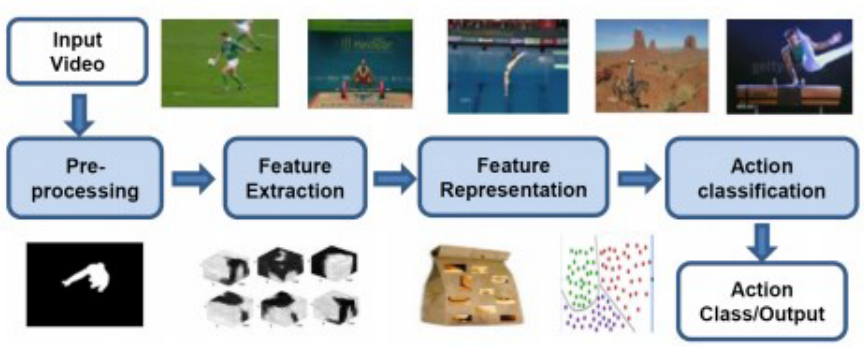


Figure 1.2: Human Action Recognition General Model

The following contributions are presented in this study:

- A system that uses DenseNet, a process that uses self-attention system with multi-head, and Bi-Conv LSTM to recognise violent occurrences in real time.
- Implementation of the given signal as well as practical use of the technique.
- An execution of data-sets that improves the detection of violence.
- To see if we could attain a high accuracy rate, we ran a cross dataset study.
- Implementing in Jetson Nano to detect real time violence.

## 1.4 Thesis outline

This report aims on constructing a surveillance system which will detect anomalous behaviour using spatio-temporal features and a 3D CNN model. The selected datasets were utilized to evaluate the model in this study. The report contains the experimental outcomes, highlights the significance and relevance of the findings and discusses the suggested model's strengths and flaws.

chapter 1 - In the introduction part states the importance and the application of violence detection and human action recognition specially in video surveillance system.

chapter 2 - In the Literature Review section , we have gone through other researchers paperwork and found their approaches. In literature review, we focused on the methodologies and the models they have used to perform their research. we have also acknowledged the outcomes of previous researchers paperwork.

chapter 3 - Proposed Method , states the models we have used to get our desired result.It also describes the architecture of the model.

chapter 4 - In the Data phase, we included the datasets that we have used throught our work.

chapter 5 - In Implementation and Result it describes the implementation of the proposed model for Anomalous-behavior detection in surveillance.

# Chapter 2

# Literature Review

After the development of deep learning, Human Action Recognition in videos got the popularity in the computer vision field to classify images and detect objects. Many assessments were concluded by researchers on deep learning approaches [24] [50]. Here, we will discuss some major approaches for video sequence data used by deep learning and analyze them with both temporal and spatial features. First of all, we will see the Human action recognition (HAR) in action.

## 2.1 Human Action Recognition(HAR)

Human activity detection for video surveillance systems is an automated method of analyzing video sequences and generating intelligent decisions about the behaviors depicted in the footage. It is one of the burgeoning fields of artificial intelligence and computer vision. Since 1980's many researchers have been working in this field. Gavrilla in 1999 made the individual research field of 2D and 3D approaches [2]. Another group of researchers JK Aggarwal and Q Cai came up with a new taxonomy focused on analysis of human motion, tracking from all types of camera view and human activity detection was provided [1]. HAR can be detected in two ways, one is in still image and another is in the video. It is also classified into two types; still image and videos. The algorithm based on videos is better than small images as it has a good amount of information. Therefore it carries temporal information along with space information. In this paper we will focus on video based algorithms.

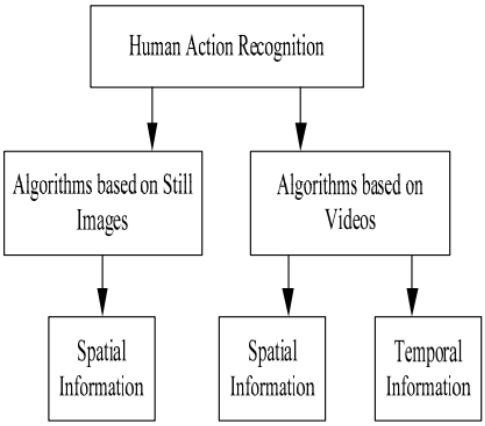


Figure 2.1: Types of Human Action Recognition

When we get a video, the preprocessing part starts to clean the noise of the video. For human action recognition, a moving object is to extract the shape of a human from the background image serially of video frames by observing when they change the place. All objects are detected when the item detection algorithm is done and the classification of objects should be used for extraction. Various types of object classification; the movement characteristics classification and shape classification. Also various types of tracking methods; background subtraction method, optical flow method, block matching, the time difference method, active contour models method, etc.

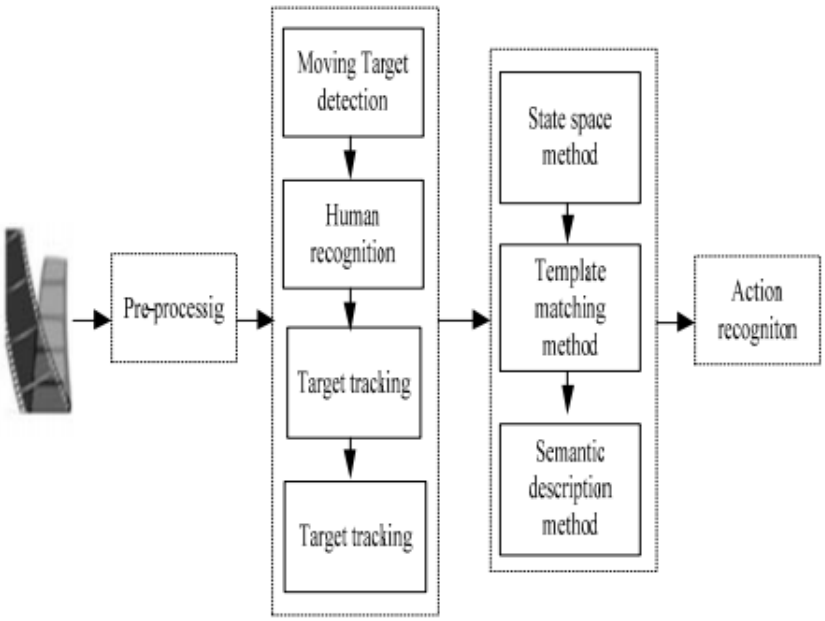


Figure 2.2: Methods of Human Action Recognition

## 2.2 Different State-of-the-art Methods

To detect crucial events and dangerous activities in videos many types of techniques or methods are flourished. Particular approaches are offered in these strategies, each of which works with a different set of input parameters. Different qualities of videos such as look, movement, stream, and so on are called parameters.

### 2.2.1 Basic concepts

Many researchers put their attention towards the computer vision field as it has a wide range of applications for analyzing images and videos. Object detection and activity recognition became top choices for its necessity. The basic architecture is shown here :

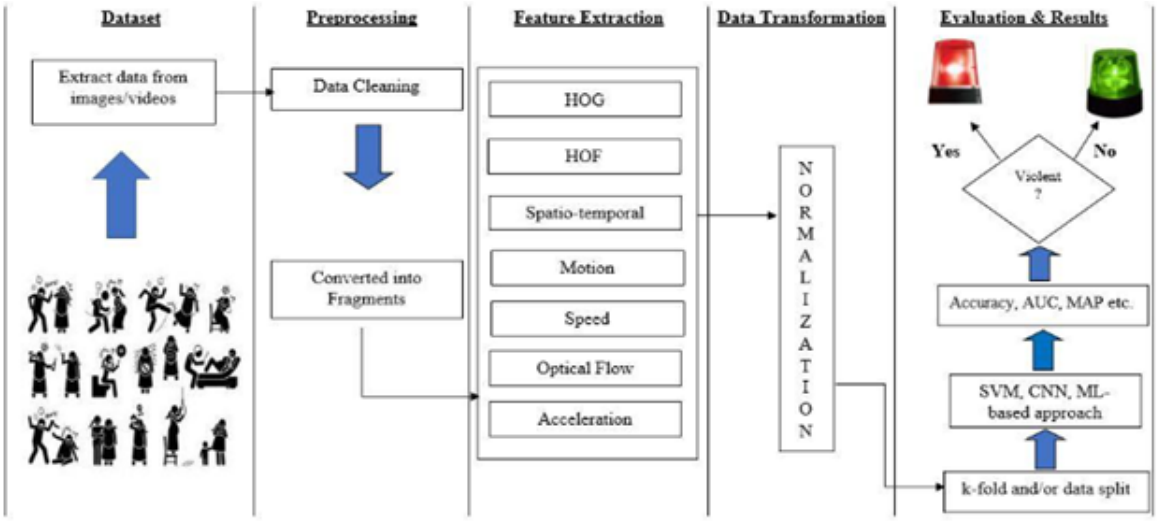


Figure 2.3: Basic Architecture

### 2.2.2 Spatio-temporal Texture Model

This is a high performing method for crowd violence detection [22]. It is a sensitive model and detects sudden changes in crowd motion. STT is composed of spatiotemporal volumes (STV) which transform the frame of the video from 2D based mechanism to 3D model analysis and slide window along the time axis. HRF is then used to compose the STT feature space. Some features of STT are as follows:

- Elementary low-level features: Distribution of grayscale from every low pass band and down sample of image.
- Coefficient features: Local auto correlations of the wavelet subbands.
- Magnitude Features: Large magnitudes that represent borders, areas, and rods are in the sub bands of the images.

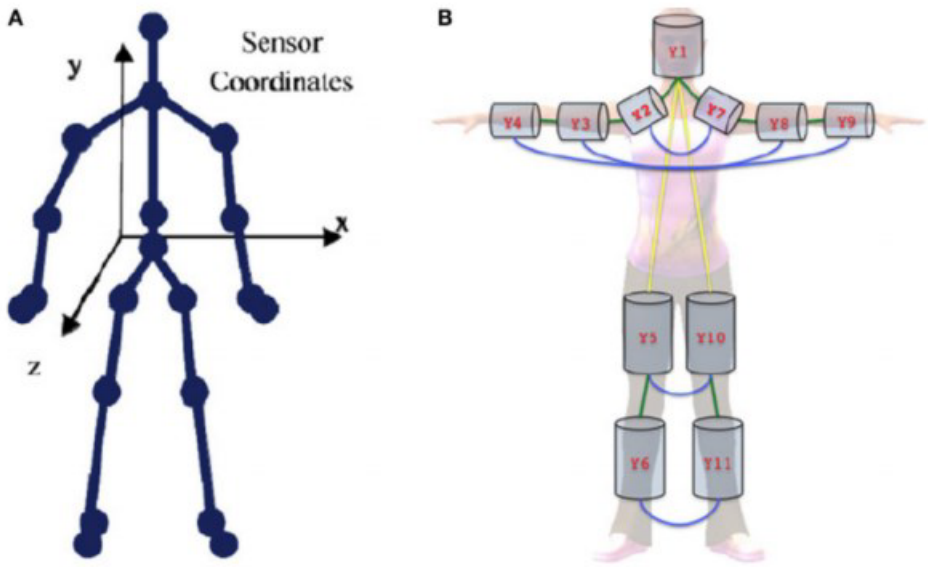


Figure 2.4: Spatio-Temporal Texture Model

### 2.2.3 Classification of Violence Detection Techniques

The use of computer vision to recognize aggression actions in surveillance has become a renowned topic in the field of action detection [39]. VDT is divided into three groups; using machine learning, using SVM and deep learning. Among them SVM and deep learning is widely used due to their success rate and accuracy over benchmark datasets. A short review on each categories as follows:

#### VDT using ML

Different traditional algorithms such as K-Nearest Neighbour, Adaboost are used as a classifier. Some profound techniques are Fast Fight Detection (FFD), Rotation-Invariant feature modeling, Motion Coherence (RIMOC), Fast Face Detection, Automatic Fight Detection, Crowd Violence Detection etc. FFD has been a novel method proposed by [16]. It is seen that motion blobs have specific shape and position in fight scenes. In terms of accuracy, BoW (MoSIFT), BoW (SIFT), ViF, LMP, variant v-1 and variant v-2 that used SVM, KNN and Ada boost are used as a classifier. It offers a significantly faster processing time, making it suitable for real-time applications. RIMOC used HOF vectors to form temporal computation embedded spheric Riemannian manifolds [21].

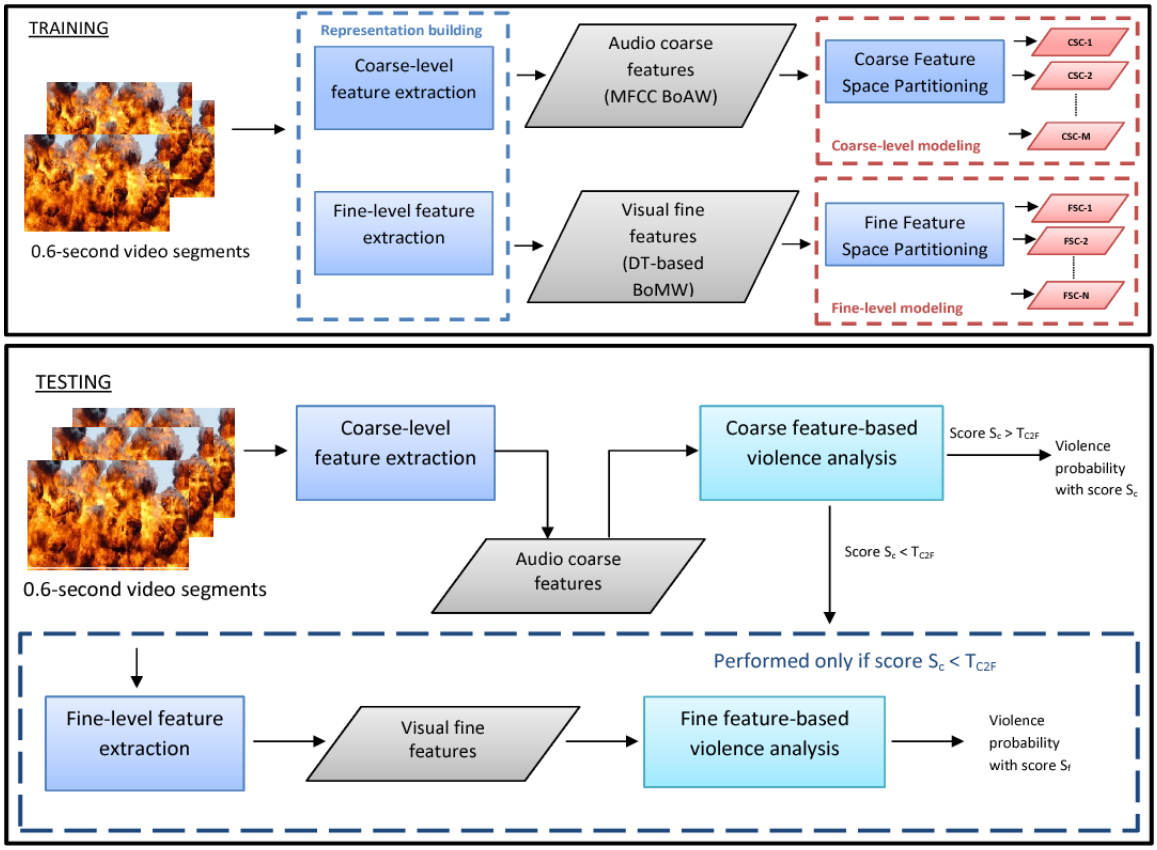


Figure 2.5: The general overview of an approach illustrating the two main phases of the system. The upper part of the figure gives the main steps performed during training (i.e., coarse and fine-level model generation), while the lower part shows the main steps of testing (i.e., execution). (DT: Dense Trajectory, BoAW: Bag-of-Audio-Words, BoMW: Bag-of-Motion-Words, CSC: Coarse Sub-Concept, FSC: Fine Sub-Concept)[17]

A set of apparatus is given by Langarian theory to analyze the long term, non local data of movement in the computer vision field. A particular lagrangian technique [31] is presented on the basis of this theory of Crowd Violence Detection because of auto recognition in video sequence data. Spatio-temporal model based Lagrangian direction areas are used for original features and use the data of foundation movement remuneration, appearances and long term motion. Comes about that the expansion of Lagrangian theory may be an important sign to detect violence and the classification execution executed over the state-of-the-art methods like ViF, HOG + BoW, two stream CNN etc. in terms of AUC and accuracy. Here are some techniques which are mostly used in Table 2.1

| Method | Object Detection Method | Feature Extraction method | Classification Method | Scene Type | Accuracy % |
|---|---|---|---|---|---|
| Motion blob (AMV) acceleration measure vector method for fighting from video [19] | Ellipse Detection method | As an algorithm to find the acceleration | Spatio-temporal feature use for classification | Both crowded and less crowded | Near 90% |
| RIMOC method focuses on speed and direction of an object on the base of HOF(Histogram Optical Flow)[34] | Covariance Matrix method STV method | Spatio-temporal vector method (STV) | STV uses supervised learning | Both crowded and uncrowded | Results for normal situation 97% Dataset of train station 82% |
| The method includes two | Vif Objection | Horn Shrunk | Interpolation | Less | Lower Frame rate |
| Step detection of violence and faces in video by using VIF descriptor and normalization algorithms | Recognition CUDA method and KLT face detector | Method for Histogram | Classification | crowded | 14% too high rate of 35 fs/s 97% |
| SVM method for recognition based on statistical theory without decoding of video frames [45] | Vector normalization method | Macro block technique for features extractions | Region motion and descriptor for video classification | crowded | 96.1% |
| Detecting fights with motion blobs [46] | Binarization of images | Spatio Temporal method to extract blobs | Classify on the base of blob length. Largest consider fighting | crowded | Depend upon dataset 70% to 98% |
| Kinetic framework by analyzing the posture for recognition abnormal activities of ATM 3D cameras [26] | Posture recognition using logistic regression | Joint angle for acquiring posture | Gradient descent Method for classification | Less crowded | 85% to 91% |
| Lagrangian fields of direction and begs of word framework to recognize the violence in videos[53] | Global compensation of object motion | Lagrangian Theory and STP method for extract Motion features | Late fusion for classification | crowded | 91% to 94% |
| A simple approach to form a video to preprocessing then feature extraction and recognition of normal and abnormal events | Gaussian Mixture Model | Apply different formulas on the consecutive frame to extract required feature | Rule based classification using a default threshold | Less crowded | Up to 90% |

Table 2.1: Violence detection techniques using ML

### Violence Detection Techniques using SVM

To resolve classification issues utilizing directed learning is known as the SVM algorithm. In SVM, information plots assume measurement space and separate inside

two classes.

It is a robust technique and kernel based. Kernel is a function that turns data into a high dimensional space in which the problem can be solved. Lack of clarity in result is one of the drawbacks of SVM[5]. Fast Violence Detection (FVT), Detecting Violence in Videos using Subclasses, Human Violence Recognition and Detection (HVRD), Violence Detection using Oriented Violent Flow, Robust Abnormal Human Activity Recognition, Framework for High-Level Activity Analysis, Real Time Violence Detection, Automated Detection of Fighting Styles are some of the widely used methods using SVM.

In the FVT model, it used the BoW framework which is specifically used for recognizing fight Spatio temporal features have been removed from the video frame and sent to ensure 90% of accuracy. Using this method infused with SIFT and MoSIFT it became 15 times faster and an increase of 12% in accuracy [7]. In terms of AUC and accuracy, the suggested descriptor outperforms state-of-the-art descriptors such as HOG, Histogram of Optical Flow (HOF), Combination of HOG and HOF (HNF), MoSIFT, and SIFT.

GMOF for surveillance hasn't gotten nearly as much attention as action recognition. The framework is robust and fast according to [23]. HVRD used Improved Fisher Vectors (IFV) using spatio-temporal position. The IFV formulas are reformulated and a summed area table data structure is employed to speed up the method. Normal spatio temporal features are taken out from videos by the help of Improved Dense Trajectories (IDT). After that, HOG represents the video using IFV. It used a linear SVM model as a classifier. According to [40], the martial arts are classified by Automated Detection of Fighting Styles methods. Mainly, it used KNN and SVM combined models to outperform existing methods in terms of accuracy (Table 2.2)

### Violence Detection Techniques using Deep Learning

This is the technique which uses CNN based categorization [18]. DL is based on neural networks which is the following method in this paper as well. It is mainly based on datasets and extracts attributes using more convolutional layers. Some of the methods of this techniques are discussed below:

**Violence Detection using 3D CNN** is a complex hand-crafted method which relies on datasets. However, huge models can directly act and extract attributes automatically. Ding et al. [12] proposed a novel 3D CNN approach. This method was used without using any dataset to see how it works and later on, it computed the foldings on the set of the frames. After that, it is trained with supervised learning and gradients with back propagation methods and found reliable in terms of accuracy.

**Detecting Violent Videos using Convolutional Long Short-Term Memory (ConvLSTM)** is a DL oriented method utilized by [33] to detect violence. CNN helps to extract the frames. The attributes are accumulated with a LSTM variant which uses convolutional gates.The use of CNN and ConvLSTM together can cap-

ture local spatio-temporal data, allowing for local motion analysis in video whose performance is better than other state-of-the-art methods such as ViF+OVif, three stream+LSTM in terms of accuracy.

**Fight Recognition Method** tasks like aggressive action which are studied less than other methods. The main feature of the detectors is efficiency, which means that these methods should be computationally quick. To achieve high accuracy, it employs a 3D CNN and a hand-crafted spatio-temporal feature.

**Violence Detection using Spatiotemporal Features with 3D CNN** Violence Detection using Spatiotemporal Features with 3D CNN is the structure of triple organized end to end profound learning, violence detection is proposed by Ullah et al. At first within the video streams in the system, persons are detected with the help of light-weight CNN models to overcome and reduce the large processing of unusable frames. Secondly, an order of 16 frames containing recognized individuals is sent to 3D CNN, which extracts spatiotemporal properties from the sequences and feeds them to the Softmax classifier. The 3D CNN model is then advanced utilizing Intel's open visual induction and a neural organizations streamlining tool kit. Prepared model is changed over into a moderate presentation and changes it for implementation at the last stage so that a definitive discovery can be made for brutality. Following the recognizable proof of viciousness, an alarm is sent to a nearby office or police headquarters, which will make a move. The Violent Crowd, Hockey, and Violence in Movies datasets are utilized in the tests. The trial discoveries show that the proposed strategy outflanks cutting edge calculations like ViF, AdaBoost, SVM, Hough Forest, 2D CNN, sHOT, and others as far as exactness, accuracy, review, and AUC.

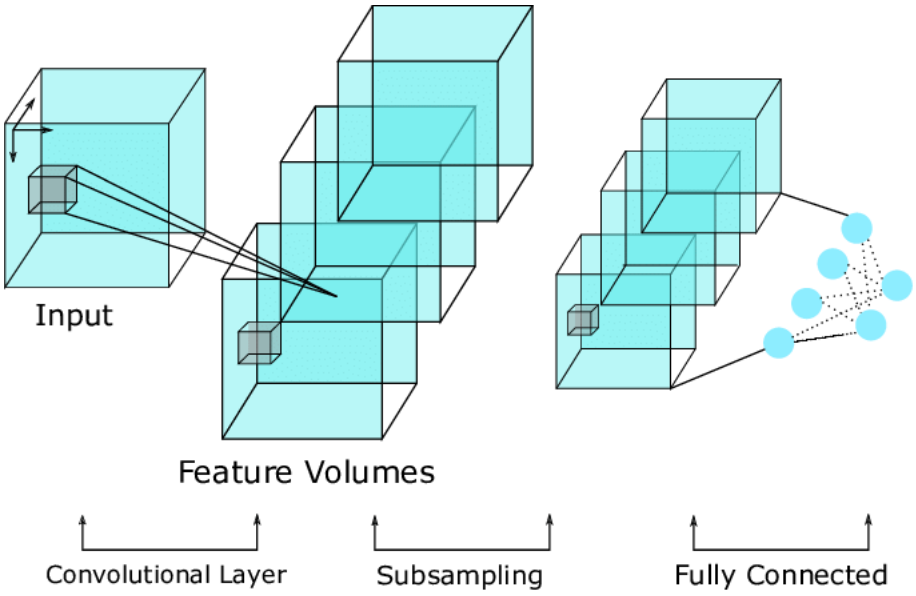


Figure 2.6: Basic 3D CNN Architecture

#### VDT using Keras , Convolutional 3D, Convolutional 2D LSTM and Convolutional 3D Transpose

Keras is an open source python toolkit which helps to create and examine DL models that are easy to use. It is compressed in Theano and Tensor- Flow, two efficient numerical computing frameworks, and allows us to design and train neural network models with just a few lines of code. The true neural network model is represented by the Keras model [47]. Keras offers two ways to build models: a basic and straightforward Sequential API and a more versatile and sophisticated Functional API. On the other Hand, A 3D Convolution is a form of convolution in which the kernel slides in three dimensions rather than two as in 2D convolutions. Medical imaging is an example of a use case in which a model is built utilizing 3D picture slices. When extracting features in three dimensions or establishing a connection between three dimensions, 3D CNNs are employed. LSTM layer with 2D convolutional convolutions. The input and recurring transformations are indeed convolutional, comparable to a conventional LSTM. The CNN LSTM, is an LSTM architecture built particularly for sequential prediction problems using spatial inputs such as pictures or videos. In Keras, we might make a NN LSTM model by first indicating the CNN layer or layers, then, at that point, encompassing them in a TimeDistributed layer, and afterward determining the LSTM and result layers. There are two ways to deal with depict the model, the two of which are indistinguishable and just contrast as far as inclination. We might characterize the CNN model first, then, at that point, encase the total series of CNN layers in a TimeDistributed layer to join it to the LSTM model. Over a multi plane input picture, a 3D transposed convolution operator is applied. The transposed convolution applies the results across all input plane outputs by multiplying each input number element wise by a learnable kernel. This module may be thought of as Conv3d's gradients with regard to its input. It's also referred to as a deconvolution or a fractionally stride convolution. Convolution is utilized to extricate applicable attributes from an info stream. It might be accomplished involving a wide range of channels in picture handling. Moreover, Convolutions in three dimensions exist. They are a 2D convolution's generalization. The filter depth in 3D convolution is less than the depth of the input layer (kernel size channel size). As a result, it is capable of moving in all three directions (height, width, channel of the image). The element-wise multiplication and addition produce one number at each location. The output numbers are also organized in 3D space since the filter glides across it. The result is a 3D data set (Table 2.3).

### 2.2.4 Related works

#### Hand-crafted Feature Based Approaches

In this approach, few methods have been developed by researchers around the world. Datta et al. [3], ffor example, employed a person's limb orientation and trajectory motion information to detect irregularities.HHMM has been suggested by Nguyen et al. [4]. Utilization of HHMM and its structure was the main contribution of them. Numerous analysts attempted to incorporate sound and Video technique,for

the identification of HAR for savagery. For example, Mahadevan et al. [8] fostered a framework which recognized blood and blazed what's more with level of movement and sound to identify violence. Some other techniques were adopted by Hassner et al. [10] to work with flow vector magnitude in other words violent flow descriptor (ViF). By the help of a SVM, the ViF descriptor has been classified for crowded scenes determining the violent or non-violent acts. Moreover, Huang et al.[13] showed a method taking as it were the factual properties of the optical stream field in video information to decide violence swarmed information which is grouped 15 into ordinary and irregular conduct classes utilizing SVM. A Gaussian model of optical stream for district extraction and utilization of an direction histogram of optical stream (OHOF) has been utilized by Zhang et al. [23] to decide savagery from a video transfer of an observation framework classed by direct SVM. Also, Gao et al. [20] proposed a technique which portrays movement greatness what's more direction data, both utilizing focused fierce stream descriptors (OViF).

**Deep Learning-Based Approaches**

Many methods have been developed since last decade. Chen et al.[6], in addition to Harris corner detector, (STIP) [6], and (MoSIFT) [9], [15], have used spatio-temporal interest points to detect violence. To identify violent and aberrant crowds, Lloyd et al. [29]created latest labels known as gray level co-occurrence texture measures (GLCM), in which alternate swarm surfaces are recorded by transient outlines. Additionally, Fu et al. [25] fostered a model to distinguish a battle scene whose capacity is to take a gander at a progression of qualities upheld by movement investigation utilizing three ascribes, including movement speed increase, movement size, and furthermore the movement district. These properties are known as motion signals that are acquired by the total of the motion region. Sudhakaran er al. [33] used LSTM and nearby structure differentiation to put in the model with the help of encoding the alternatives in the videos. Histogram of optical flow magnitude and orientation (HOMO) proposed by Mahmoodi et al. [48].An analyst named Fenil et al. [44] gave a structure in light of histogram of arranged angle (HoG) highlights from each casing. They utilized the element to prepare bidirectional LSTM or (BD-LSTM) which guarantees forward and in reverse data admittance to contain data about fierce scenes. The methods described above attempted to address a variety of issues in violence detection, such as camera viewpoints, complicated crowd patterns, and intensity changes. When variation occurs within the physique for violence detection, for example, they did not capture the discriminative and useful traits by extracting them. Viewpoint, considerable mutual occlusion, and scale all contribute to these differences [42]. Furthermore, ViF is unable to detect the difference between the two flow vectors, limiting the accuracy when the movement of a vector for same magnitude and different direction is present in the two frames having 1 pixel.

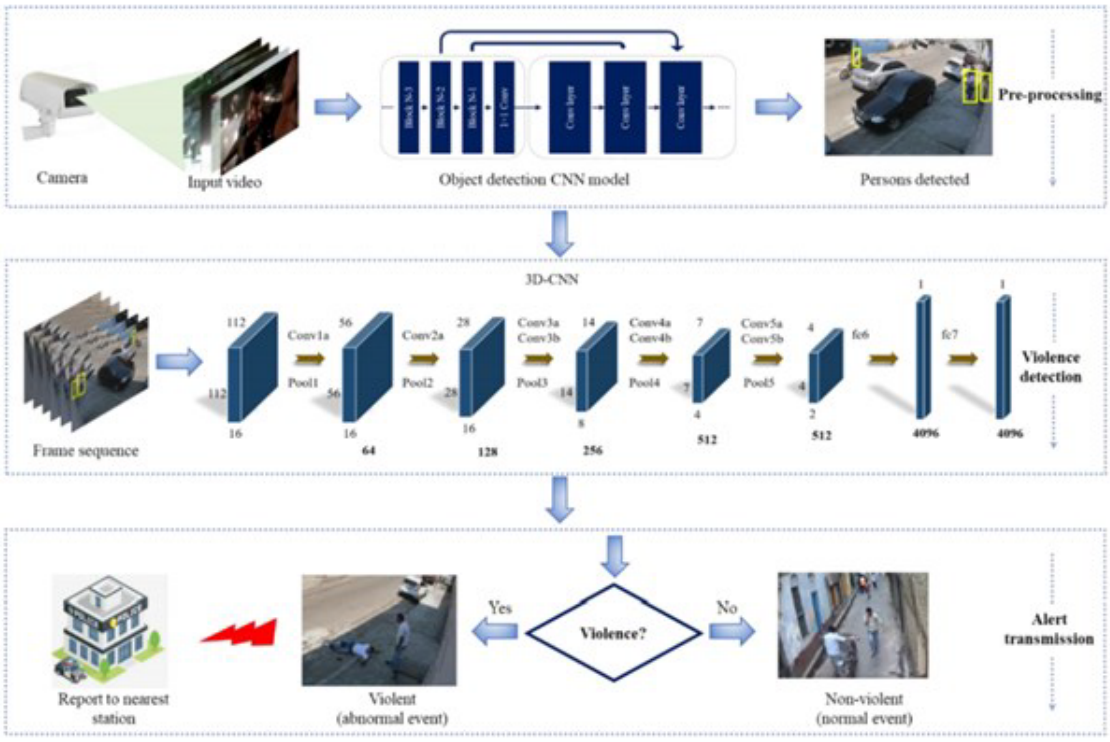


Figure 2.7: The Framework of the proposed violent detection method

| Method | Object Detection Method | Feature Extraction method | Scene Type | Accuracy % |
|---|---|---|---|---|
| Real time detection of violence in crowded scenes [4] | ViF descriptor | Bag of features | crowded | 88% |
| Bag of words Framework using acceleration | Background | Ellipse estimation method | Less | Approx 90% |
| Multi model features framework on the base of the subclass[10] Image CNN and ImageNET | Google net for feature extraction | Less | Crowded | 98% |
| To determine the occurrence of violent purpose extended form of FV(Improved Fisher vector) and sliding windows [13] | Spatial pyramids and grids for Object Detection | Spatio temporal grid technique for feature extraction | Crowded | 96%-99% using different data sets |
| Violence detection using oriented violent Flow [23] | Optical flow method | Combination of ViF and OViF descriptor | Crowded | 90% |
| AEI and HOG combined framework to recognize the abnormal event in visual motions[6] | AEI technique for background subtraction | HOG and Spatio temporal methods to extract features | Both crowded and Less Crowded | 94%-95% |
| The framework includes preprocessing, detection of activity and image retrieval. This work identifies the abnormal event and image from data-based images[20] | Optical flow and temporal difference for object detection CBIR method for retrieving images | Gaussian function for video analysis | Less Crowded | 97% |
| Late fusion method for temporal perception layers to detect high level activities. Use multiple cameras from 1 to N [15] | A motion Vector method to identify from multiple cameras in 2D | SGT MtPL method | Less Crowded | 98% |
| Bi-channel | A motion vector method to identify from multiple cameras in two dimensions | SGT MtPL method | Less Crowded | 98% |
| Convolutional neural network for real time detection[29] | Image Net method of object detection | VGG-f model for feature extraction | Crowded | 91%-94% |
| Solve detecting problem by dividing the objective in depth and clear format using COV-Net[25] | Movement detection And TR of Model | BoW Approach | Less Crowded | 96% |

Table 2.2: Violence detection techniques using SVM

| Method | Object Detection Method | Feature Extraction method | Scene Type | Accuracy % |
|---|---|---|---|---|
| Violence Detection Using 3D CNN[44] | 3D convolution is used to get spatial information | Back propagation method | crowded | 91% |
| Deep architecture for place recognition [42] | VGG VLAD method for image retrieval | Back propagation method for feature extraction | crowded | 87% -96% |
| Tracking violence sights using CNN and deep audio feature[39] | MFB | CNN is used | crowded | Approx 90% |
| Detect violent using convLSTM[16] | CNN along with the ConvLSTM | CNN model | crowded | Approx 97% |
| Detecting Human violent behavior by integrating Trajectory and deep CNN | Deep CNN | Optical flow method | crowded | 98% |
| Hough forest Methodology for recognition | Object detection using Spatio- temporal features | MOSIFT method to extract video features | Less crowded | 84% -98% |
| Violence detection using Spatiotemporal Features with 3D CNN [5] | Pre-train MobileNet CNN model | 3D CNN | crowded | Approx 97% |

Table 2.3: Violence detection techniques using 3D

# Chapter 3

# Proposed Method

## 3.1 Model Architecture

Each video is transformed to optical flow from RGB for achieving the later classification using a fully connected network from robust generation for video encoding, to mainly organized violence in videos. Optical flow is encoded by a dense network as a segment of featuring maps. At first the featured maps go through a multi head self-attention layer. After that it goes through a bidirectional ConvLSTM layer. After all these, attention mechanisms are applied in the forward temporal direction as well as in the backward temporal direction respectively. This is basically a spatio-temporal encoder which actually extracts the necessary features regarding the spatio as well as temporal from each and every video. Finally, after doing all these the features which are encoded are inserted to a four-layer classifier which mainly signifies whether it is a violent video or not.

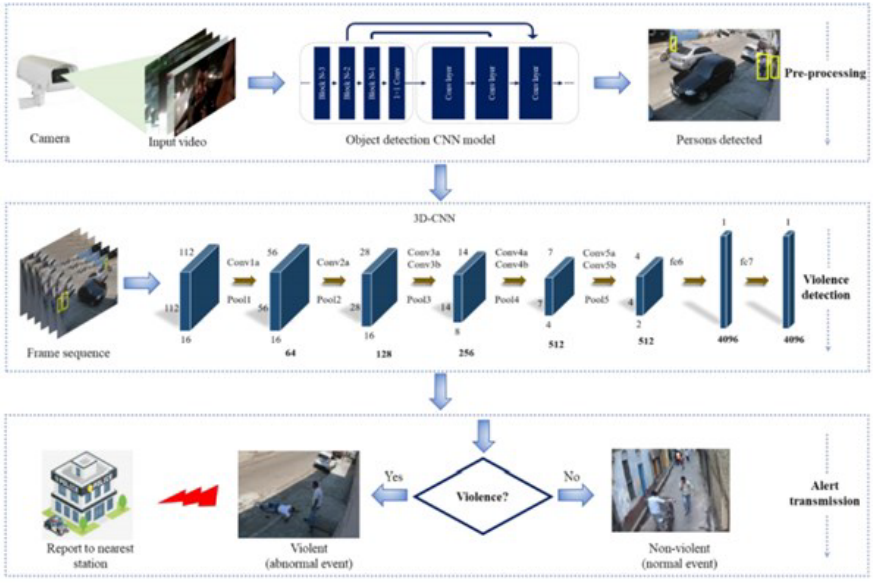


Figure 3.1: Framework of our proposed method

## 3.2 Model Justification

As shown in the figure below we can see a flourishing outcome of the test result of the block composing the model architecture in the sector of human action recognition as well as violent actions. Besides, for video classification the 3D DenseNet variant has been used. The bidirectional recurrent convolutional block allows the feature analysis in the forward and backward temporal direction and thus the block improves the efficiency while recognizing the violent actions. The attention mechanism mainly recognizes three things. Those are: human action, convolutional network combination and bidirectional convolutional recurrent blocks. The model really helped us in developing our proposal as it is based on blocks which recognize human actions. The detailed description of ViolentNet architecture in given below:

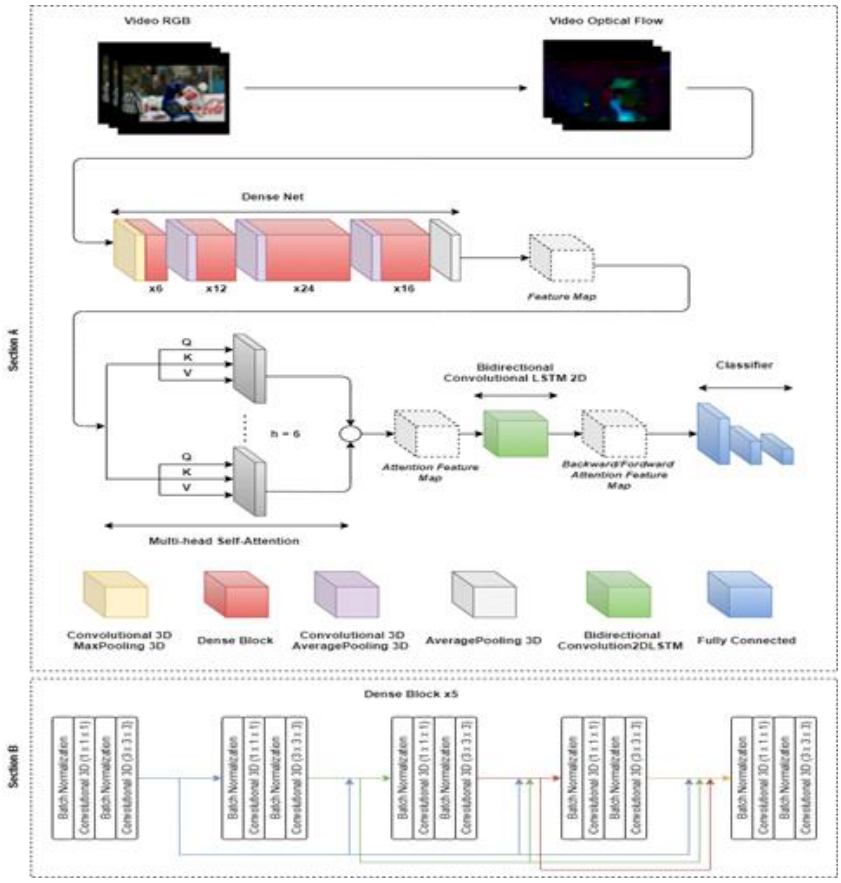


Figure 3.2: Violence Net Architecture Model

### 3.2.1 Optical Flow

The pattern of apparent mobility of a visual object between two consecutive frames created by the movement of an object or camera is known as optical flow.It's a two dimensional vector field in which each vector is a displacement vector indicating the movement of points from one frame to the next. Structure from Motion, Video Compression, and Video Stabilization are just a few of the uses for optical flow.

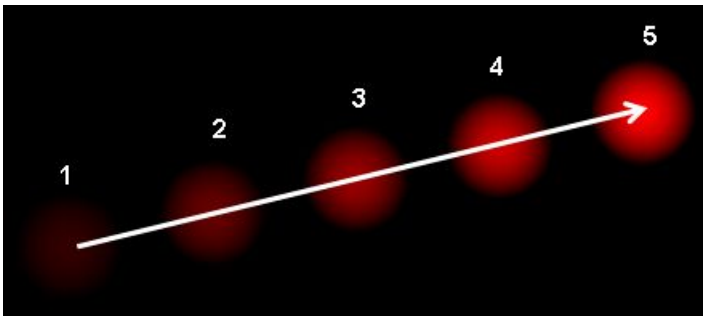


Figure 3.3: optical Flow

It shows a ball moving in 5 frames in a row. Its displacement vector is indicated by the arrow. The following assumptions underpin optical flow:

- The pixel intensities of an item do not change between frames.
- The mobility of neighboring pixels is similar.

Here, we consider the first frame's pixel ***I(x,y,t)***where a new dimension, time, has been added here. We used to simply work with photographs, thus there was no need for time. In the next frame after dt time, it moves by distance ***(dx,dy)*** . We may say

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{3.1}$$

because those pixels are the same and the intensity does not vary. Remove common terms and divide by ***dt*** using the right-hand side taylor series approximation to produce the following equation:

$$f_x u + f_x v + f_t = 0 \tag{3.2}$$

$$f_x = \frac{\partial f}{\partial x} \tag{3.3}$$

$$f_y = \frac{\partial f}{\partial y} \tag{3.4}$$

$$u = \frac{dx}{dt} \tag{3.5}$$

$$v = \frac{dy}{dt} \tag{3.6}$$

The equation above is known as the Optical Flow equation.We can locate image gradients $f_x$ and $f_y$ in there. Similarly, $f_t$ represents the gradient over time.

### Farneback Method

The Farneback algorithm produces an image pyramid, with each level having a lower resolution than the one before it. When we select a pyramid level greater than 1, the system can track points at various resolution levels, starting with the lowest. It has many functions and parameters when we compute a dense optical flow using Gunnar Farneback's algorithm. Optflow Farneback gaussian gauges optical stream employing a Gaussian winsizewinsize channel instead of a box channel of the same measure; regularly, this alternative gives z more exact stream than a box channel at the cost of slower speed; regularly, winsize for a Gaussian window should be set to a bigger esteem to attain the same level of vigor. Using the algorithm, the function finds an optical flow for each previous pixel.

$$prev(y, x)\ next(y + flow(y, x)[1], x + flow(y, x)[0]) \tag{3.7}$$

On the other hand, The minEig Threshold algorithm divides the number of pixels in a window by the minimum eigenvalue of a 2x2 normal matrix of optical flow equations (this matrix is called a spatial gradient matrix); if this value is less than minEig Threshold, a corresponding feature is filtered out and its flow is not processed, allowing for the removal of bad points and a performance boost.

### Dense optical flow

This is one of the inputs of our network. The generation of frame sequence is done in this algorithm where the most moved pixels between the consecutive frames are represented with greater intensity. It has been the most vital component in violent clips . Besides that the main components are contact and speed. The pixels have a trait of moving much at the time of a particular segment in comparison to the other segments of the video and also they have a tendency to make a cluster in a particular portion of the video. Along with the optical flow we obtained a 2 channel matrix after the application of the algorithm , the magnitude as well as the direction are also included. The hue value of a picture is corresponded mainly through the direction. That value is used for visualization purposes only and the magnitude corresponds to the value plane. We chose dense optical flow over discrete optical flow because dense optical flow generates flow vectors for the entire frame, up to one flow vector per screen size, whereas sparse optical flow only generates flow vectors for certain features, such as some pixels that portray the edges/seams of an object within the frame.

Dense optical flow is used as an input in a model because. In deep learning models, just like our proposed one we can see an unsupervised feature here and thus we can see a wide range of features is way better actually.

### 3.2.2 DenseNet Convolutional 3D

The structure of DenseNet was built in order to process images. It has a 2D convolutional layer. Other than that the DenseNet can be modified to work with videos. The modifications are:

- Replacing the 2D convolutional layer with 3D
- Replacing the 2D reduction layer with 3D.

The DenseNet has a system of working in layer by layer and the layers are connected in feed-forward fashion[27]. The reduction layers of DenseNet are MaxPool2D and AveragePool2D the pool sizes are (2,2) and (7,7). Instead of them the MaxPool3D and AveragePool3D were used which have the size of (2,2,2) and (7,7,7). The basis of DenseNet structure is the dense blocks. The blocks are made of the feature maps of a layer with all of its product.

In our suggested system we have used four sense blocks, all of them have different sizes. The dense blocks consist of a course of layers. The layers follow the manner of batch normalization convolutional 3D . The main reason for using DenseNet is its simplicity in using feature maps.The DenseNet works in a more efficient way than ResNet or Inception. The DenseNet structure can work in such a way that is more prosperous and it generates a lower number of screens and specifications in order to achieve high performance.

### 3.2.3 Multi-Head Self-Attention

The machinery of Multi-Head Self-Attention is built in such a way that it joins different positions in a single string and creates an outcome of it which concentrates the most relevant part of the string[52]. It is established on the attention mechanism that was introduced first in 2014. Multi-Head self attention implements multiple self-attention techniques. The architecture of this procedure is showing the input data by applying different linear projections learned from the same data and finally executing the self-attention mechanism in every output. We selected the multihead-self attention mechanism to determine which elements are common in both temporal directions by developing a weighted matrix that consists of more relevant past and future information. The specification of multi-head self attention layer are:

- number of heads h=6
- dimension of questions $d_q = 32$
- dimension of values $d_v = 32$
- dimension of keys $d_k = 32$

The improvements we got from this layer will be discussed in chapter 5. Multi-head self attention mechanism forms new relations among features, determining if the action is violent or not.

### 3.2.4 Bidirectional Convolutional LSTM 3D

This system has two states: forward-state and future-state. The generated output can get data from both states. This module is well-known for its ability to look back in video. It divides the components of the periodic layers in positive and negative time [28]. Usually in neural networks the temporal features are selected but spatial features sometimes disappear. In order to avoid such a situation we proposed a model where we generated convolutional layers instead of entirely connected layers, here the convLSTM is able to observe both spatial and temporal features and let us get data from both features. Bidirectional convolutional system is an advanced convLSTM which has the access to look backward and forward in a video,which gives the system an overall better outcome.

### 3.2.5 Classifier

Classifier is made up of connected layers. Each layer has nodes that are ordered in a definite manner. These are 1024,128,16,2. So, we can see there are four full layers which are actually connected. However, the ReLu activation function is used by a layer which is hidden. The Sigmoid activation function is engaged verifying whether an action category is violent or not and the output is a binary predictor which is of the last layer. Self-Attention mechanism achieved a very high success rate on determining the relevance of words in natural language processing and text analyzing.

# Chapter 4

# Datasets

## 4.1 Datasets

In our thesis we actually have used the most widely accepted datasets which are basically the benchmarks; those are the hockey fight dataset, movie fight dataset, violent flows dataset, RWF-2000 dataset. These are well balanced, labeled and they were having a (80-20)% ratio while splitting for training purposes as well as for testing purposes respectively. Besides, these datasets cover mainly indoor scenes, outdoor scenes and few weather conditions.

Figure 4.1: training and testing videos in different datasets.

**The Hockey fight dataset (HF)**

This dataset contains an equal amount of violent and nonviolent actions during professional hockey games, with two players often involved in close body contact and contains 1000 videos extracted from hockey games of NHL which is USA's National Hockey League.This dataset contains indoor scenarios.


Figure 4.2: Indoor scenes

### The Movie Fight dataset (MF)

This dataset includes violent and non-violent events of some action movies which have around 200 clips of collection. It has scenes that have both indoor and outdoor scenes but not any weather conditions scenes.


Figure 4.3: Outdoor scenes

### The violent Flows dataset (VF)

This is a collection of 246 videos captured in places that include violence in crowds. This dataset is actually active on mass violence occurring outside and besides that some scenes contain some weather conditions too.

### The RWF dataset

This dataset contains 1000 violent videos which contain real life street fighting events and 1000 non-violent videos which contain scenes related to normal daily life activities. This dataset has a collection of raw surveillance videos from YouTube. To work with it more efficiently these videos have been sliced into 5-second chunks at 30 frames per second and recognize each clip as violent and non-violent action respectively. The duplicate material appearing in both the training and validation sets were removed. Then at last it obtained 2000 clips and 300,000 frames as a dataset for violent action detection.

Figure 4.4: Real world videos captured by surveillance cameras with large diversity

Figure 4.5: Crowd violence (246 videos captured in places), Movies Fight (200 videos extracted from action movies) and Hockey Fight (1k videos extracted from crowded hockey games)

## 4.2 Data Preprocessing

In **RWF-2000** This data preprocessing system is done with python scripts and to convert the videos they used a tensor with shape where the number of frames, the image height and width were declared. Here, its last channel has three layers for the RGB components and two layers for optical flow.

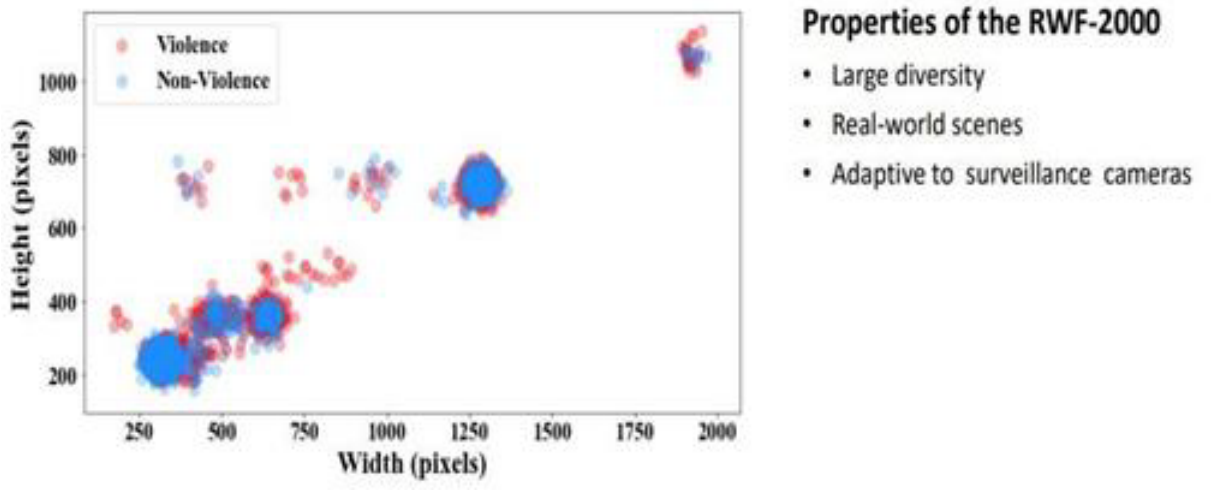


Figure 4.6: Detection of violence and non violence

To extract the features, in images the color changed RGB to Gray through optical flow. The dimension is also used when we convert video to python file which is of Height, width 224*224 and only 1 channel is used as it is gray. After padding two channels have been created: one is a normal channel and another is an empty channel. Lastly, before converting to numpy a channel has been created named empty array to convert gray channel to RGB. After that the color again converted to RGB from BGR and has been reshaped into three channels red, green, blue respectively by intacting 224*224. For final data preprocessing five channels have been created with the same height and width.

The **Hockey fights dataset** Here the frames are isolated from each movie into distinct batches of frames and then applied all available augmentation procedures to each dataset individually, such as removing black borders from the" Hockey" dataset. Images are subtracted pixel by pixel from adjacent frames which are input of the encoder model. Instead of the raw pixels from each frame, this was done to include spatial movement in the input films.

For **Movie Fight dataset** In the case of this dataset the preprocessing method is extraction of the frames from the shot and then this extraction will be inserted into an algorithm which will be trained for violent recognition. In comparison to other videos, such as surveillance, movies can usually be readily sampled into shots due to their hierarchical structure. This hierarchical structure aids in the analysis of the entire film at the shot level and thus a histogram-based method is used to segment the movie into shots. A saliency map is created for all of the frames in a one shot and then they are compared using the maximum number of non-zero pixels which is divided by the summation of all the pixels in the frame.

From this dataset we faced some issues, due to the dark surroundings, fast movement of objects, illumination blur, and other factors, many of the footage obtained by surveillance cameras in public locations may not have high picture quality. Some instances like only a portion of the person is seen in the photograph, crowds and

chaos, detecting objects from far distance, temporary actions, low resolution etc. Therefore, to navigate the model we have saved all the datasets in Audio Video Interleave extension. We tried to avoid the obstacles by removing the blurry pictures and using the RGB components to get a most possible and accurate result and for RWF-2000 and UCF datasets the level of accuracy we got in RWF-2000 and UCF crime 2000 are 86.75% and 99% respectively in case of violence detection.

# Chapter 5

# Implementation and Result

In this chapter the implementation of our proposed model has been described for detecting anomalous behavior for the surveillance system. This model used Python, Tensorflow, Keras and OpenCV for implementation and used for testing also. The implementation part is divided in four sections; Training Methodology,Training Metrics, Ablation Study and Experimentation on Cross Dataset.

## 5.1 Implementation

This section summarized the training process and assessed the ablation research where it would be understood the importance of the self-attention mechanism that we proposed. Moreover, the experimentation of a cross dataset is implemented to assess the generality of anomalous behavior.

### 5.1.1 Training Methodology

In this stage, the weights of all neurons in the model were randomly initialized. Here, A range of 0 to 1 has been chosen to normalize each input pixel value. To calculate the input video sequence from all of the videos from every dataset the average of all sequence was calculated. The extra frame has been removed if the input video consists of more than the average frames. To maintain the average frame, if there was less sequence than the average, the last sequence is being repeated untill it can reach the average. Moreover, the frames were enlarged to the standard size for Keras pre-trained models, which is $224 \times 224 \times 3$. The following parameters were chosen: 104 as the rate of base learning, 12 films as the batch size, and 150 epochs. The weight decay has set to 0.1. In addition to that, the default setup of the Adam optimizer has been employed. For the last layer of the classifier, as the loss function determiner the Binary Cross Entropy has been taken and for activation function the Sigmoid function has been chosen. On the Nvidia GT 730 GPU, the CUDA toolbox was utilized to extract deep features for the tests where the operating

system was Windows 10 and the processor was an Intel Core i7. It is decided to use a three-fold cross validation method and to test the performance of the dataset a random permutation crossvalidator method was used. Two types of input method were used to conduct the implementation: one is with optical flow and the other is with neighboring frames removal which can be named as the pseudo-optical flow. The temporal component was implicitly expressed in both entries, but in various ways. To see what kind of input performs better and generates the best outcome, two input methods were used. By removing two consecutive frames, the pseudo optical flow was created.

On each pair of the adjacent frames, a matrix subtraction has been performed in a sequence of frames

$$\forall n < k \in [0, k] : sn = f(n) - f(n) + 1 \tag{5.1}$$

Using the way, any variation of two successive frames was introduced between the pixels. Three violent situations are transformed into both input formats individually. The input method that uses the neighboring frames removal technique represents pixels that have not moved between two successive frames (f0.....fx)which is the fundamental distinction between the two approaches. The pixel turns black when the value does not change in both frames as both of the frames had been removed and it is not taken into account if that pixel moved or not. The pixels that turn black in the optical flow method represent those pixels that never moved between two successive frames.

### 5.1.2 Metrics

To identify the performance and efficiency measurement, the set of training metrics has been used for this model are as follows:

**Train accuracy**: The number of right classified sections made by the proposed algorithm on the train set on which the training was built was fraction of total number of classification.

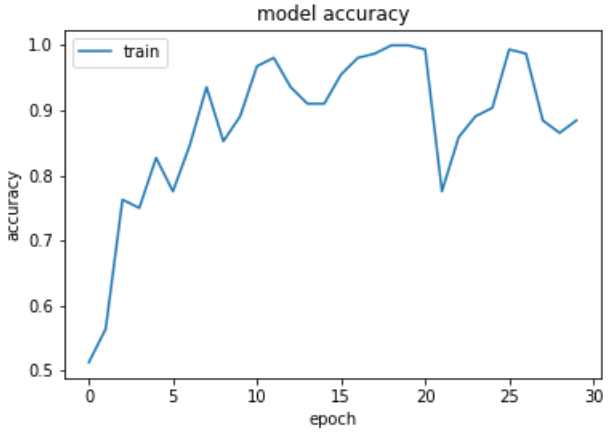


Figure 5.1: Train Accuracy

**Test accuracy**: The number of right classified sections made by the proposed algorithm on the new cases divided by the total number of classification.

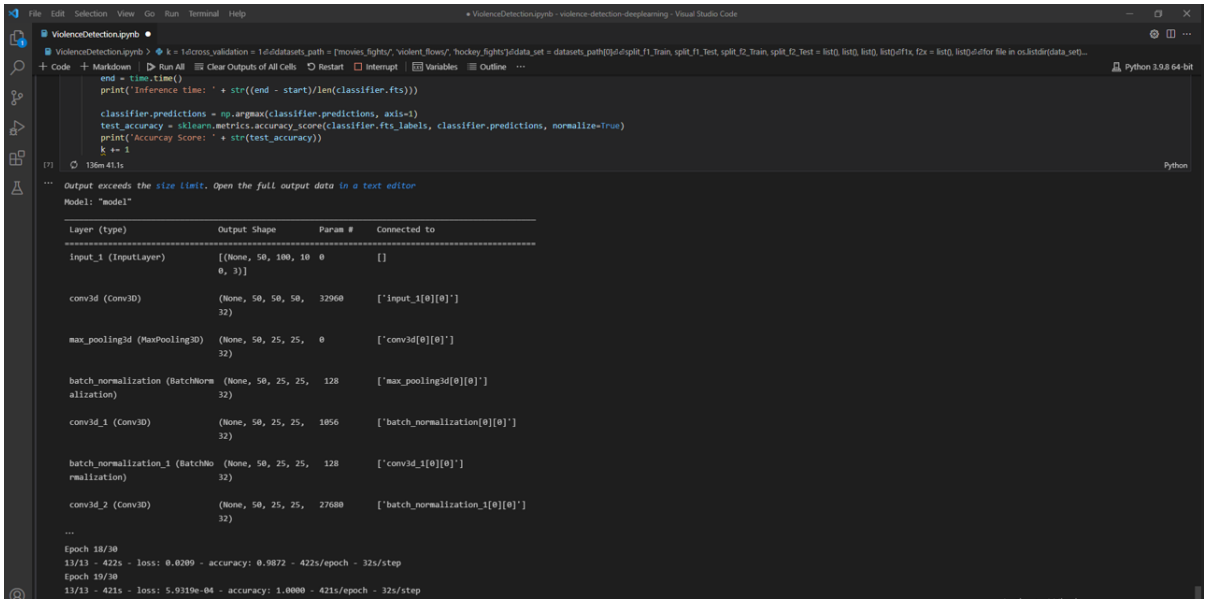


Figure 5.2: Test Accuracy

**Inference time during test**: When making atomic predictions on the test dataset, the average delay time.

### 5.1.3 Ablation Study

To see the usage of both input methods (Optical flow and Neighboring frames removal technique) and to examine the self attention mechanism on the model and the importance of it, a two-fold experiment was developed for the ablation investigation. The direct connection between the Bi-ConvLSTM and the DenseNet avoided the self-attention process.

### 5.1.4 Experimentation on Cross Dataset

The goal of the experimentation of a cross dataset is to see how accurately the trained model on one dataset can evaluate the examples from the other dataset. One of the major goals that needed to be observed is if the learning of the model in concept of violence is generic enough or not to appropriately detect violence on different datasets. In this research, two types of cross dataset configuration have been examined. For the first type, the model has been trained with one dataset and then it is tested with the first one; and the second configuration was the training which is combined of three dataset and tested on the remaining one.

## 5.2 Results

Here in this section we have discussed the comparison between the state of the art and the cross dataset experimentation and we have gained the results from the ablation study. In addition to that, this segment provides the implementation result after running our proposed model for violence detection in surveillance. Python, Tensorflow and Keras have been used to run the test on unclassified input data and to generate the result.

### 5.2.1 Ablation Study Results

Despite the fact that a very powerful core network has been used here than in any prior studies. We thought it would be really amazing to watch the improvement of performance through the modification of the input of the network (that is the optical flow and the pseudo optical flow) and also engaging the attention mechanism. Even though we have found mainly two major benefits: improvement in accuracy and shorter inference time after measuring the two versions of the proposed model; those are optical flow and the pseudo-optical flow and those were compared to the ones without having the self-attention module. We can see a constant progress towards betterment in accuracy as well as in the inference time from the Table 5.1 shown below. The reason for the progress is as the operation time required for bidirectional convolutional recurrent layer required longer on application to feature maps generated from CNN compared to attention layer that is concatenated in terms of sequence, it was visible that the inference time required with attention mechanism was less than without it.

| Dataset | Method of Input | precision evaluations (with attention) | precision evaluation (without Attention) | precision reasoning time (with Attention) | precision reasoning time (without Attention) |
|---|---|---|---|---|---|
| HF | feature extracted with Optical flow | 99.10±0.6% | 98.90±1.0% | 0.1398±0.0025s | 0.1627±0.0035s |
| HF | Neighboring frame removed method | 97.40±1.0% | 97.30±1.0% | 0.1398±0.0024s | 0.1627±0.0034s |
| MF | feature extracted with Optical flow | 100.00±0.0% | 100.00±0.0% | 0.1917±0.0093s | 0.2018±0.0045s |
| MF | Neighboring frame removed method | 100.00±0.0% | 100.00±0.0% | 0.1917±0.0093s | 0.2018±0.0045s |
| VF | feature extracted with Optical flow | 96.90±0.5% | 94.00±1.0% | 0.2971±0.0030s | 0.3115±0.0073s |
| VF | Neighboring frame removed method | 94.81±0.5% | 92.51±0.5% | 0.2993±0.0030s | 0.3115±0.0073s |
| RWF-2000 | feature extracted with Optical flow | 95.61±0.6% | 93.50±1.0% | 0.2768±0.0020s | 0.3029±0.0059s |
| RWF-2000 | Neighboring frame removed method | 94.20±0.8% | 92.30±0.8% | 0.2777±0.0020s | 0.3029±0.0059s |

Table 5.1: Ablation Study Result

The variation accuracy was least when the dataset consisted of fifty frames on the average (HF and MF), but when it reached hundred frames on the average (VF and RWF-2000), the model with self attention outperformed the others by two 2 points. The deduction time lessen on each dataset; it lessened from 4% (VF) to 16 %(HF) without attention. Lastly,when we got the result for all dataset and we saw that pseudo-optical flow without attention and optical-flow with attention both were successful in all dataset environments except for the MF dataset. The outcomes were 2 points in HF, 4.4 points in VF, 3.4 points in RWF-2000 respectively.

### 5.2.2 Cross-Dataset Experimentation Results

Two logical facts emerged from the cross-dataset experiments,firstly there was minor connection in different datasets and a better understanding in various environments. Secondly it revealed a better perception of the idea when we compare multiple dataset at a time. We can also see that the results did not exhibit successful generalization for pairings of datasets which were significantly dissimilar in terms of the sort of violence scenario. When comparing MF and VF the results show a low accuracy rate of 52.32% because it has completely different environments. But when we trained the dataset in the opposite direction it showed a slightly higher accuracy of 60.02%. The best result we found for HF,RWF-2000,VF dataset where the test accuracy was 81.51%. Overall the RWF-2000 dataset gave us the best result in every aspect.

The results we got while testing in different environments are shown in Table 5.2

| Training | Testing | precision evaluation Optical Flow | precision evaluation Pseudo-Optical flow |
|---|---|---|---|
| HF | MF | 65.19±0.34% | 64.87±0.41% |
| HF | VF | 62.57±0.33% | 61.23±0.22% |
| HF | RWF-2000 | 58.23±0.24% | 57.37±0.22% |
| MF | HF | 54.93±0.33% | 53.51±0.12% |
| MF | VF | 52.33±0.34% | 51.78±0.30% |
| MF | RWF-2000 | 56.73±0.19% | 55.81±0.20% |
| VF | HF | 65.17±0.59% | 64.77±0.49% |
| VF | MF | 60.03±0.24% | 59.49±0.16% |
| VF | RWF-2000 | 58.77±0.49% | 58.33±027% |
| RWF-2000 | HF | 69.25±0.27% | 68.87±0.14% |
| RWF-2000 | MF | 75.83±0.17% | 74.65±0.22% |
| RWF-2000 | VF | 67.85±0.32% | 66.69±0.22% |
| HF+MF+VF | RWF-2000 | 70.09±0.19% | 69.85±0.14% |
| HF+MF+RWF-2000 | VF | 76.10±0.20% | 75.69±0.14% |
| HF+RWF-2000+VF | MF | 81.52±0.09% | 80.50±0.05% |

Table 5.2: Cross Dataset Experimentation

## 5.3 State of the Art Comparison

When the studies were completed we discovered that the optical flow input produced superior results than the pseudo-optical flow input. In the below table the results are shown for the training and testing process.

Here we can see the optical flow highlighting the spatio-temporal features of the movies and it works better than the pseudo-optical flow because it has a bigger loss function reduction.

While comparing our method with the state of the art Table 5.3 we can see that it shows better results than previous studies,even for those studies which do not include cross-validation and maintain a small number of specifications. Where there was person-to-person violence the best result we got from MF and HF. The MF dataset, in particular, was the most comparable and least demanding. The concept also functioned well in situations where there was a lot of crowd violence.

| Dataset | Object Detection Method | Feature Extraction method | Train Accuracy | Training Loss | Test Accuracy Violence | Test Accuracy Non-Violence | System Usage |
|---|---|---|---|---|---|---|---|
| **HF** | **3D Bi-LSTM (Proposed)** | **Optical flow** | **100%** | **1.30×10-5** | **99.50%** | **100.00%** | **68%** |
| HF | LSTM-AE | NFR | 97% | 99% | 1.35×10-5 | 97.00% | 85% |
| HF | RNN that uses convLSTM cells | FeedForward | 100% | 1.22×10-5 | 98.00% | 100.00% | 93% |
| HF | 2D ConvLSTM | Back propagation method | 93% | 1.53×10-5 | 91.00% | 92.80% | 73% |
| **MF** | **3D Bi-LSTM (Proposed)** | **Optical flow** | **100%** | **1.18×10-5** | **100%** | **100%** | **67%** |
| MF | LSTM-AE | Neighboring frames removal | 99.00% | 1.39×10-5 | 98.40% | 99.00% | 82% |
| MF | RNN that uses convLSTM cells | FeedForward | 100% | 1.19×10-5 | 100% | 100% | 87% |
| MF | 2D ConvLSTM | Back propagation method | 95.00% | 1.62×10-5 | 94.20% | 95.00% | 70% |
| **VF** | **3D Bi-LSTM (Proposed)** | **Optical flow** | **98%** | **1.50×10-4** | **97.00%** | **96.00%** | **71%** |
| VF | LSTM-AE | Neighboring frames removal | 97.00% | 2.94×10-4 | 95.00% | 94.00% | 83% |
| VF | RNN that uses convLSTM cells | FeedForward | 99% | 1.34×10-4 | 98.00% | 98.80% | 91% |
| VF | 2D ConvLSTM | Back propagation method | 90% | 3.10×10-4 | 89.00% | 89.90% | 73% |
| **RWF-2000** | **3D Bi-LSTM (Proposed)** | **Optical flow** | **96%** | **3.10×10-4** | **96.00%** | **95.00%** | **65%** |
| RWF-2000 | LSTM-AE | Neighboring frames removal | 95% | 7.3×10-4 | 94.00% | 93.00% | 80% |
| RWF-2000 | RNN that uses convLSTM cells | FeedForward | 98% | 2.90×10-4 | 96.00% | 97.60% | 93% |
| RWF-2000 | 2D ConvLSTM | Back propagation method | 92% | 8.4×10-4 | 91.00% | 92.00% | 69% |

Table 5.3: State of The Art Comparison

In the hockey game the player constantly moves and hardly makes any physical contact. Our proposed model made a huge rate of success in observing the movements in this dataset. Our model acquired temporal properties based on vigorous movement at the time it occurred.

The VF dataset had a distinct problem generalizing the concept of violence than the hockey fight dataset. The videos in the VF dataset showed a large scale of violent activities such as protests,concerts and other gatherings. During major events, a lot of things happened at the same time. Because the viewpoints were so far away from the action, many people seemed in low resolution. The distance between the event and the observer machine made the object tiny and the motion detection on a single film made it a lot more difficult to tell if the action was violent or not and in fact, it is difficult for humans too. In addition to that, the environment with mass events contains unique events such as a crowd trying to catch a baseball which seems to be the start of a fight. Because the RWF-2000 dataset is so diverse, it is also challenging to normalize the process of violence. The RWF-2000 scenes were not topic-specific, unlike the other three datasets. When it is about non-violence category, the heterogeneity of the dataset becomes more visible. Along with that, the actions were vastly different for every scene from each other. Our model has achieved state-of-the-art test accuracy for various datasets. In comparison to other models, ours was in a decent spot. For the HF, MF, VF, and RWF-2000 datasets, various models were used before our suggestion. There were many models that have come to the field by now that achieved 100% accuracy in terms of test set which made them unbeatable.

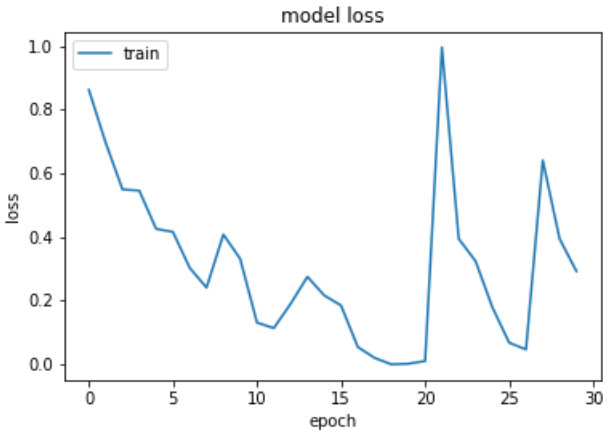


Figure 5.3: Training loss

Our approach surpassed the nearest one by more than 2 points on the VF dataset. Prior to ours, the RWF-2000 dataset had only been evaluated with one model. Using hold-out validation, this model has a test accuracy of 92.00 percent for the RWF2000 dataset. Our model scored 95.60 percent, which is an improvement over the current state of the art.

The optical flow input method's test accuracy was significantly greater than with the pseudo-optical flow input. The MF dataset was an exception to that. It has the same test accuracy value for the both input methods. Apart from that, this happened with

all datasets. Even when measuring our model to others who employed a hold-out validation procedure, it is clear that our strategy outperformed the competition.

In terms of trainable parameters, it is another remarkable achievement our model got, which is numerously less than the other proposed model till now based on the data available. The usage of DenseNet architecture and the underlying model of feature map concatenation of the architecture are to blame for this. The Flow-Gated Network 3D CNN Flow-RGB is the only model that has lower trainable parameters than our proposed model which is the only one standing model. However, this model was not applicable because of its test accuracy which did not exceed 60 percent for any the datasets.

However, RNN based Conv-LSTM has a better score in terms of accuracy but it has greater system usage than our proposed model. We are using Jetson Nano developer board to execute our model and make it use for real time surveillance. In this situation, we need to trade accuracy with the resource. Our model requires 10.70 Giga FLOPs. Compared to high accuracy RNN based Conv-LSTM it is 40.92 Giga FLOPs that requires more time and resources. Therefore, we chose our proposed model that is best suited for implementation in Jetson Nano with satisfactory accuracy rate.

## 5.4 Jetson Nano

After completing implementation of our proposed model trained with benchmark datasets, we used the pre-trained model to detect violence with jetson nano. For our implementation, we used Jetson Nano 2 GB development board and used all 4 cores of Maxwell CPU and used 4 GB data swap for better performance. In the figure we can see the first violent act video feed captured in the camera however not detected yet.

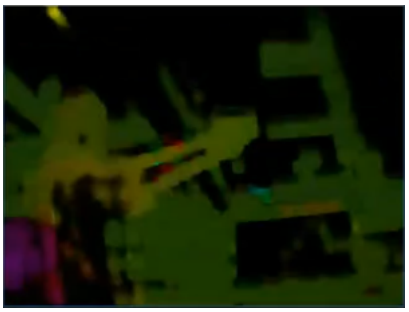

Figure 5.4: Act of violence begin detection

In the following two figures, we can see that Jetson Nano has kept only guns in red and removed all other features of the picture as per our model that increases the violence detection rate and also increases the fps.

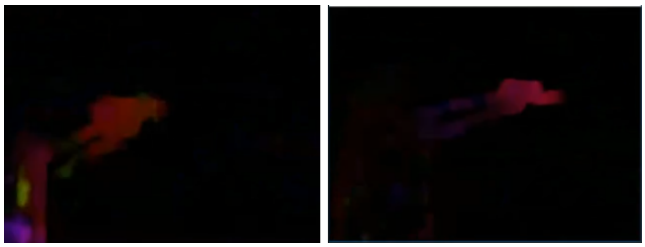

Figure 5.5: Gun detected

Our approach was not limited to detecting violence but also we wanted to implement it in smart cities. Therefore, we created feeds for CCTV so that later we can identify the culprit. The real time CCTV feed version of our model with detection of violence is shown in the figures.

Figure 5.6: Violent Action Detected

For better understanding of our approach, the figure shown below is the side by side comparison of the original CCTV footage which will be monitored by the proper authorities and the underlying process of the Jetson Nano for detecting the violence.

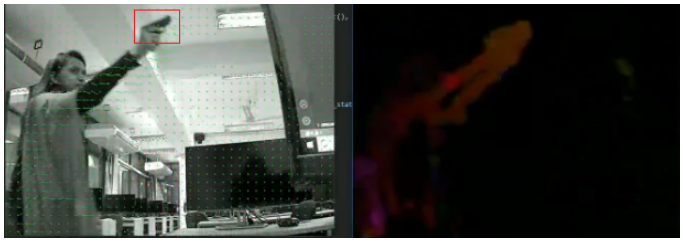

Figure 5.7: side by side comparison

In this section, we will see the comparison of our model with other models after implementing on Jetson Nano in terms of performance. To get a clear idea, the comparison is shown based on the FPS of the video. In the figure, we can see that our model outperformed the rest of the model in terms of the fps.

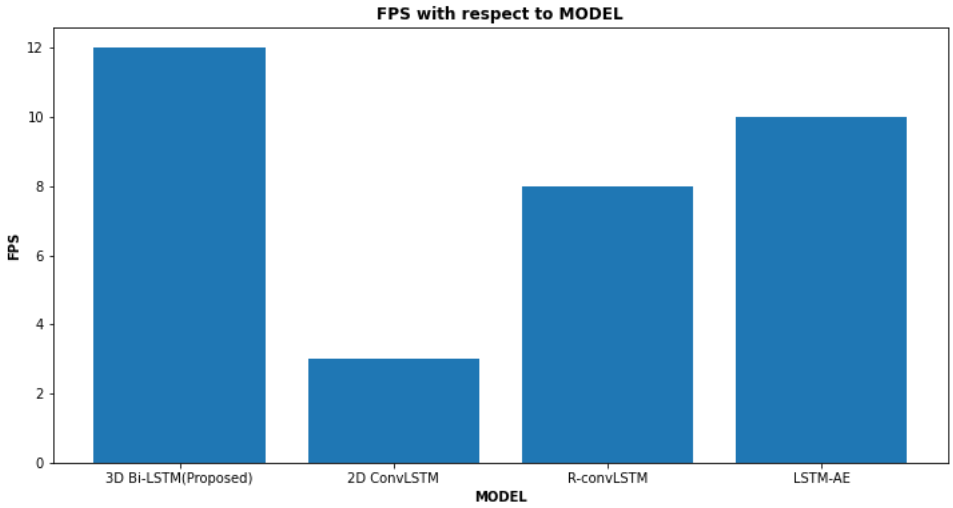


Figure 5.8: Comparison based on FPS

The figure shows our model has achieved 12 fps while detecting the violence which is the highest among others. The second best model with 10 fps is the LSTM AutoEncoder that has a lower accuracy rate. The RNN based approach that has convLSTM cells in between Encoder and Decoder has the fps of 7 to 9 which is lower than the proposed method of ours.

After a few trade offs with accuracy to gain performance, our model outperformed other models to detect real time violence with satisfactory accuracy and performance.

## 5.5 Detection Process in CCTV

The architecture of violenceNet has the ability to classify the videos whether they are violent or nonviolent. Our model worked with CCTV footage of different videos of the same length. The segments of the videos were trained with dense optical flow technique and violenceNet takes input from it and classifies if the segments are violent or not.

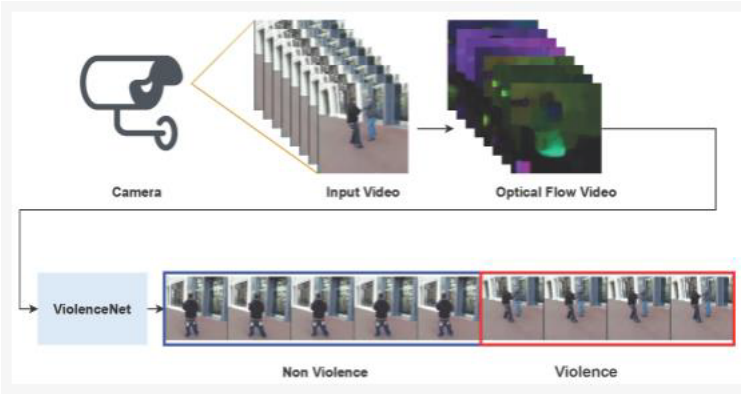


Figure 5.9: Human Action Recognition Subsystems

This system identifies each fragment based on the light flow provided by a camera in a CCTV system. The violent features are represented by the red box and non-violent features are represented by the blue box.

# Conclusion

ViolenceNet is a space-time encoder architecture that advances the state-of-the-art in recognizing violence. Our key contribution is an architecture which is the combination of a modified DenseNet with a multi-head self-attention module and a 3D LSTM bi-directional convolutional module which is an ablation study of the self-attention mechanism as RNN and bi-directional RNN have been engaged in problems which are mainly video centric as it has been shown in a number of studies. Our datasets show that our proposed method exceed the state of the art and the cross dataset experiment to work further based on violent actions. Some inspection on short video datasets show us that the accuracy dropped from 95 percent to 100 percent using the same dataset cross-validation to 70.08 percent to 82 percent in cross-dataset experiments which guides us to believe that future research should focus on abnormality recognition on large video datasets. UCF-Crime, XDViolence, UBI-Fights, or CCTV-Fights are among the datasets that are worth highlighting. In these circumstances, it will be necessary to detect a video whether it is violent or not and also when the violence occurs. As a result, an embedded device which is Jetson Nano described here, can capture temporal information successfully in both directions, is a good solution to cope with increasingly heterogeneous datasets. The application of new deep learning techniques based on transformers is another fascinating area of research to follow. Finally, human features are not included in our model and perform appropriately as per the datasets we used. In future to accomplish a generalization of violence involving people it is necessary to include pose estimation or face identification. Despite the strong solution we provided with 12 FPS there is still room for more improvement in terms of FPS that we hope to see in future. We feel that more research into this topic will yield positive outcomes.

# Bibliography


[1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer vision and image understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[3] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *Object recognition supported by user interaction for service robots*, IEEE, vol. 1, 2002, pp. 433–438.

[4] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov model," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, vol. 2, 2005, pp. 955–960.

[5] L. Auria and R. Moro, "Advantages and disadvantages of support vector machines," *Credit Risk Assessment Revisited: Methodological Issues and Practical Implications*, pp. 49–68, 2007.

[6] D. Chen, H. Wactlar, M.-y. Chen, C. Gao, A. Bharucha, and A. Hauptmann, "Recognition of aggressive human behavior using binary local motion descriptors," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2008, pp. 5238–5241.

[7] M.-y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," 2009.

[8] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 1975–1981.

[9] E. B. Nievas, O. D. Suarez, G. B. Garcıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*, Springer, 2011, pp. 332–339.

[10] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2012, pp. 1–6.

[11] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[12] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3d convolutional neural networks," in *International Symposium on Visual Computing*, Springer, 2014, pp. 551–558.

[13] J.-F. Huang and S.-L. Chen, "Detection of violent crowd behavior based on statistical characteristics of the optical flow," in *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, IEEE, 2014, pp. 565–569.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv preprint arXiv:1406.2199*, 2014.

[15] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on mosift feature and sparse coding," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 3538–3542.

[16] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," *PloS one*, vol. 10, no. 4, e0120448, 2015.

[17] E. Acar, F. Hopfgartner, and S. Albayrak, "Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies," *Neurocomputing*, vol. 208, pp. 225–237, 2016.

[18] T. Agrawal, A. Kumar, and S. K. Saraswat, "Comparative analysis of convolutional codes based on ml decoding," in *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)*, IEEE, 2016, pp. 41–45.

[19] G. Batchuluun, Y. G. Kim, J. H. Kim, H. G. Hong, and K. R. Park, "Robust behavior recognition in intelligent surveillance environments," *Sensors*, vol. 16, no. 7, p. 1010, 2016.

[20] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[21] P. C. Ribeiro, R. Audigier, and Q. C. Pham, "Rimoc, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance," *Computer vision and image understanding*, vol. 144, pp. 121–143, 2016.

[22] J. Wang and Z. Xu, "Spatio-temporal texture modelling for real-time crowd anomaly detection," *Computer Vision and Image Understanding*, vol. 144, pp. 177–187, 2016.

[23] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications*, vol. 75, no. 12, pp. 7327–7349, 2016.

[24] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang, and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Systems with Applications*, vol. 81, pp. 108–133, 2017.

[25] E. Y. Fu, H. V. Leong, G. Ngai, and S. C. Chan, "Automatic fight detection in surveillance videos," *International Journal of Pervasive Computing and Communications*, 2017.

[26] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[27] K. W. Lee, H. G. Hong, and K. R. Park, "Fuzzy system-based fear estimation based on the symmetrical characteristics of face and facial feature points," *Symmetry*, vol. 9, no. 7, p. 102, 2017.

[28] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional lstm based spectral-spatial feature learning for hyperspectral image classification," *Remote Sensing*, vol. 9, no. 12, p. 1330, 2017.

[29] K. Lloyd, P. L. Rosin, D. Marshall, and S. C. Moore, "Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (glcm)-based texture measures," *Machine Vision and Applications*, vol. 28, no. 3-4, pp. 361–371, 2017.

[30] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[31] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE transactions on information forensics and security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[32] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[33] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2017, pp. 1–6.

[34] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[35] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, 2017.

[36] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Kernel fusion based extreme learning machine for cross-location activity recognition," *Information Fusion*, vol. 37, pp. 1–9, 2017.

[37] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6016–6025.

[38] S. Lee and E. Kim, "Multiple object tracking via feature pyramid siamese networks," *IEEE access*, vol. 7, pp. 8181–8194, 2018.

[39] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, 2018.

[40] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer lstm," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.

[41] B. Yousefi and C. K. Loo, "A dual fast and slow feature interaction in biologically inspired visual recognition of human action," *Applied Soft Computing*, vol. 62, pp. 57–72, 2018.

[42] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS one*, vol. 13, no. 10, e0203668, 2018.

[43] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 449–458.

[44] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan, *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[45] I. U. Haq, K. Muhammad, A. Ullah, and S. W. Baik, "Deepstar: Detecting starring characters in movies," *IEEE Access*, vol. 7, pp. 9265–9272, 2019.

[46] S. Khan, K. Muhammad, S. Mumtaz, S. W. Baik, and V. H. C. de Albuquerque, "Energy-efficient deep cnn for smoke detection in foggy iot environment," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9237–9245, 2019.

[47] H. Lee and J. Song, "Introduction to convolutional neural network using keras; an understanding from a statistician," *Communications for Statistical Applications and Methods*, vol. 26, pp. 591–610, 2019.

[48] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert systems with applications*, vol. 127, pp. 121–127, 2019.

[49] M. Sajjad, S. Khan, T. Hussain, *et al.*, "Cnn-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognition Letters*, vol. 126, pp. 123–131, 2019.

[50] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services," *Information Sciences*, vol. 479, pp. 416–431, 2019.

[51] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3d convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, 2019.

[52] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[53] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107 561, 2020.