A DEEP CONVNET FOR VIOLENCE DETECTION USING DYNAMIC MOTION IMAGES

A Dissertation (Major Project-II)

Submitted In Partial Fulfillment Of The Requirements For The Award Of The Degree Of

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS

Submitted by:

AAYUSH JAIN

2K18/ISY/01

Under the supervision of

Dr. Dinesh K. Vishwakarma

Associate Professor, Department of Information Technology

Department of Information Technology

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Bawana Road, Delhi-110042

Augut-2020

DECLARATION

I, Aayush Jain, hereby declare that the work submitted in the Dissertation "**A DEEP CONVNET FOR VIOLENCE DETECTION USING DYNAMIC MOTION IMAGES**" by me in partial fulfillment of the requirements for the award of the degree of Master of Technology in Information Systems from Delhi Technological University, is an original and authentic work under the supervision of **Dr. Dinesh Kumar Vishwakarma**, Associate Professor, Department of Information Technology.

This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: New Delhi

Date :

Aayush Jain

2K18/ISY/01

CERTIFICATE

This is to certify that the Major Project Report-2 titled "**A DEEP CONVNET FOR VIOLENCE DETECTION USING DYNAMIC MOTION IMAGES**" submitted by **Aayush Jain (Roll No. 2K18/ISY/01)** for partial fulfillment of the requirements for the award of the degree of Master of Technology (Information Systems) is a record of the project work carried out by the student under my supervision.

Place: New Delhi
Date :

Dr. Dinesh Kumar Vishwakarma
SUPERVISOR
Associate Professor
Department of Information Technology
Delhi Technological University

ACKNOWLEDGMENT

First and foremost, I would like to express my deep sense of respect and gratitude to my project supervisor **Dr. Dinesh Kumar Vishwakarma,** Associate Professor, Department of Information Technology, Delhi Technological University, Delhi for providing me the opportunity of carrying out this project and for his continuous support during this thesis. This is my heartfelt thanks for his motivational advice, and encouragement without which the project would not have shaped as it has.

Secondly, I am grateful to **Prof. Kapil Sharma**, HOD, Information Technology Department, DTU for his immense support. I would also like to acknowledge the Delhi Technological University faculty for providing the right academic resources and environment for this work to be carried out.

Last but not least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Aayush Jain
2K18/ISY/01

ABSTRACT

Over the last decade, there has been a surge in developing automated intelligent video analysis systems that can monitor human activities in the public environment and recognize abnormalities like violent and suspicious events. Violence Detection is an emerging topic in monitoring human activities. Deploying such violence detection automated systems at highways, shopping malls, sports complexes, market places, public places like airports, railways stations, and bus stands can help us with the preparedness of repercussion of unusual or violent crowd behaviors. Violence detection aims at identifying whether a violent action has occurred and has evolved as a popular theme in the department of image processing and computer vision. Improved highly effective methods for intelligent analysis are highly demanded.

Various methodologies can detect such activities based on Deep Learning algorithms, SVM, and Machine learning algorithms. Deep neural nets and Transfer learning have proven highly successful in the detection of violent activities. The motive of this dissertation is to propose a novel deep ConvNet system for the task of detecting violence by extraction of motion features from RGB Dynamic Motion Images (DMI). Motion feature extraction and prediction of violent content using a stream of RGB DMI is done effectively by pre-trained CNN model – Inception-Resnet-V2 followed by fine-tuning layers. The advantages and limitations of existing state-of-the-art CNN based architectures for violence detection suggested by various researchers and popular datasets used for violence detection are also discussed. For performance validation of the proposed novel model, tests are performed on three popular and publically available benchmarks – Hockey Fight dataset, Real Life

Violence Dataset, and movie dataset. The performance is also checked against the other widely used pre-trained models – Resent50 and Inception V3.

Keywords: Violence Detection, Image Processing, Computer Vision, Deep ConvNet, Transfer Learning, Dynamic Motion Images, Inception-Resnet-V2, CNN, Resnet50, Inception V3

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CCTV	Closed Circuit Television
CV	Computer Vision
VD	Violent Detection
ML	Machine Learning
SVM	Support Vector Machine
DL	Deep Learning
HOMO	Histogram of Optical flow Magnitude and Orientation
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
AUC	Area Under Curve
ROC	Receiver Operating Characteristics
CNN	Convolution Neural Network
RGB	Red Green Blue
ViF	Violent Flows
OViF	Oriented Violent Flows
LSTM	Long Short-Term Memory
2D	2 Dimensional
3D	3 Dimensional
ANN	Artificial Neural Network
NN	Neural Network
RF	Random Forest
DMI	Dynamic Motion Images
ARP	Approximate Rank Pooling
RP	Rank Pooling

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Modern technical advancement and development in the field of image processing and video processing have been extraordinary because of their application in identifying important content across multiple domains, including searching, analyzing, and identifying actions. Identifying actions from videos has grown in prominence in the last decade because of the rise of violent activities involving acts of terrorism. Over the last decade, highly congested public places such as shopping centers, educational institutions, airports, hospitals, banks, markets, streets, etc. are being monitored by surveillance systems having CCTV camera and other equipment to monitor activities of people[1] for ensuring public safety and better crowd management due to rapid increase in global population. But the effectiveness of such systems is questionable since they require adequate manpower and human attention skills are limited[2]. Some assessments tell us that 99% of the generated footage is never watched[3] and hence, detection of suspicious activity is a very difficult task[4][5]. True effectiveness is reduced to only 1% and hence the automated analysis of crowded scenes has received quite some attention from the research community. Researchers are now using computer vision techniques to recognize human activity[6]–[8] and monitor the crowd automatically. Various techniques were introduced for identifying crowd anomalies and helping to detect suspicious activity in surveillance footage. The tracking of violence from the surveillance footage[9], [10] is another kind of detection of activity.

The technology for detecting movements and objects has progressed considerably in terms of advancement and has helped us to combine the various technologies to create an integrated system that can accurately monitor violent and potentially violent activities. Various methods and techniques to detect violent events and other harmful patterns in videos are surveyed by[11], [12]. In such methods, different approaches work with several parameters or features of videos such as optical flow[13], [14], acceleration, appearance, human pose, time, etc. Various researches talk about different techniques explored based on machine learning,

support vector machines, deep learning for the detection process to boost the accuracy, and effectiveness. The research domain toggles between violence detection and abnormal behavior detection, the activities which contain actions like fighting, beating, pushing, etc. are surveyed under violence detection. The detection of aggression in a video stream involves multiple sections or steps, such as target recognition, movement identification, and video classification. A great deal of advancement in developing systems that will help us in detecting violence using machine learning algorithms, deep learning algorithms, etc.

1.2 VIOLENCE DETECTION

Violence in ordinary life is suspicious occasions or exercises. In the domain of activity detection, identification of these events in surveillance footage via CV is currently an active area. In the particular zone of violence detection, the major work is centered around low-level attributes. As a rule, highlights, for example, gradients, intensities, optical flow, and nearby attributes are extracted. Because of human torso diversities, catching viable and biased characteristics in the violent scenes or actions is tough. The primary driver of the varieties is perspective, shared impediment, scale, and dynamic scenes. By perceiving fierce parts, for example, fire, blood, weapons, and sound, most techniques concentrated on identifying savage scenes. Such strategies, be that as it may, may not be proper for checking recordings with poor picture quality. The inconveniences, for example, low detection rate and high bogus alert, limit this kind of technique. Besides, these highlights are not reasonable for general observation frameworks that are continually ailing in sound data.

Countless researchers have developed various approaches and methods for the discovery of brutal or unusual events, due to a rapid increase in the crime, for gradually accurate identification. Different ways of detecting violence are being developed, and have been discussed in recent years. Depending on the classifier used, such methods can be divided into three groups:

1. VD using ML

2. VD Using SVM

3. VD Using DL

1.3 APPROACH OVERVIEW

Different researchers suggested various methodologies for improving the effectiveness, and consistency of the Violence Detection process. Generally, the approach to detect violence follows a common procedure that being extracting data from video, pre-processing the data and converting into fragments or batches before feeding it to the model/classifier and feature extraction like motion features, Spatio-temporal features, etc. using various methods like HOG, HOF, HOMO, optical flow, etc. Then once features are extracted they are used to predict/classify video content and depending on the prediction values, accuracy and AUC/ROC plots are obtained. Once the model is established, it is ready to be deployed to detect violent activity in videos/footage. The basic steps of the VD are shown in fig. 1.1.

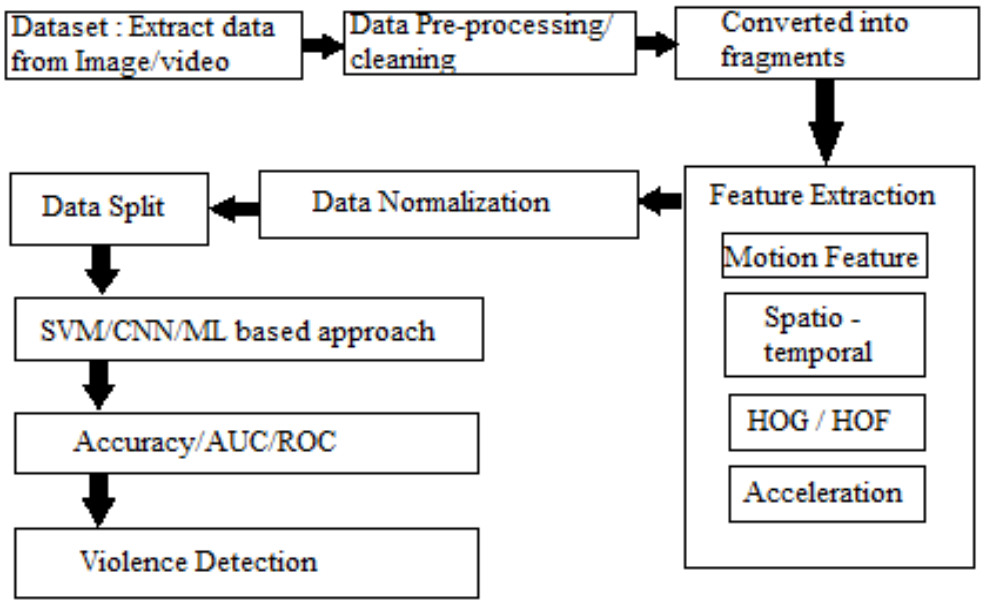

Figure 1.1 Basic steps for a typical violence detection technique

1.4 MOTIVATION

With the spike in violent activities like the recent Delhi riots [15] or the recent mass protests in the US resulting in violent events [16], it is becoming more and more necessary to develop efficient and highly accurate automated intelligent systems which can analyze the public premise for detecting and avoiding suspicious events and violent activities for ensuring better public safety and enabling agencies in deploying better security measures. Recent state-of-the-art has been quite successful

and effective in creating surveillance systems and detection methods which can enable enforcement agencies to avoid such incidents [17]–[20]. In the last decade, CNN has achieved remarkable progress in the field of video and image classification with incomparable accuracy and modeling capacity [21]–[23]. Prior studies in the area of violence detection [14], [24]–[29] have ascertained that temporal, spatial, and motion information act as an important trait to classify and detect violent activities in surveillance footage and real-life situations in an effective manner. Researchers have also done extensive work in detecting violent activities using a single-stream architecture extracting spatial information or temporal information or motion information and also by multi-stream architectures to extract spatial-temporal or spatial-temporal-motion data that has improved CNN based systems accuracy by examining RBG and depth based appearance and motion details.

Yet most of the approaches like [30] do not necessarily model the dynamics of the video. RGB images, extracted as a stack of frames from the video, are used to learn discriminative attributes that allow the model to obtain the highest accuracy in tasks of recognition and classification since the primary aim is to represent the action class rather than the motion. Whereas motion features contain a decent amount of evidence for distinguishing the movement and human action. In this regard, optical flow [13], [14], [31], [32], and dense trajectories [33] are used extensively for illustrating the movement of the entity in videos. Various methods proposed descriptors, based on optical flow orientation and magnitude changes for example ViF [34], [35], OViF [36], HOF and HOG [37] , HOMO [24], etc.

However, these approaches regardless of being beneficial also have some disadvantages like high computational time limiting the possibility of real-time implementation of such systems. Similarly, dense trajectories can be extremely sensitive to camera viewpoint.

1.5 ORGANIZATION OF DISSERTATION

The dissertation is organized as follows:

- Chapter 2 deals with a short and crisp overview of the topic, dynamic motion images, CNN and its variants, transfer learning, and literature review related to Violence Detection.

- Chapter 3 showcases the proposed architecture used for violence detection.
- Chapter 4 presents the results achieved by the proposed model on the 3 publically available datasets.
- Chapter 5 concludes along with future work.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW

The series of RGB images called frames constitute an RGB video stream and the deployed systems built must discern people's positions to understand how human activity varies over time in the consecutive video frame. Action recognition constitutes identifying diverse actions in a video clip(a 2D frame sequence)[38]. This can be visualized as a part of the image classification task for multiple frames and then combining the predictions from all frames. Various approaches were proposed for the feature extractions from a frame in a video like [39] used 3D CNN for the spatiotemporal feature extraction. Over a few years, work presented by researchers on action recognition can be distinguished into 2 main classes:

1. Hand-crafted
2. Deep-net

Before deep learning algorithms came to solve the problem, various traditions Computer Vision algorithms for recognizing actions were broken into broadly 3 parts:

1. Local high-dimension visual characteristics which describe the stretch of the video to be sparsely dispersed set of points of interests[40] or densely extracted[33]
2. The extracted characteristics are encoded in video level descriptions. Bag of visual words is one of the most common and popular versions of this process.
3. The final prediction is done using a trained classifier such as SVM or RF.

2.2 DYNAMIC MOTION IMAGES

Considering all such approaches based on motion feature extraction and their shortcomings [41] proposed a new approach which is an extremely powerful, simple,

and yet efficient depiction of the video called Dynamic Images or Dynamic Motion Images, summarizing the motion dynamics present in the video into a single image. Dynamic Motion Images (DMIs) can encrypt information in a conventional content-agonistic manner which results in highly effective long-term, stable representation of motion for classification and recognition tasks and also for other tasks like human pose estimation [42], [43], as well. Examples of DMI are shown in fig. 2.1

Figure 2.1 Examples of Dynamic Motion Images

Key benefits of using Dynamic Motion Images which are as follows:

1. DMIs can be attributed to various forms of sequences.
2. DMIs are very efficient, fast, and simple.
3. Extraction of DMIs reduces computational time, also reduces video analysis to single RBG image analysis.
4. Compact Representation of video.
5. DMIs can be useful for large scale indexing
6. DMIs can be processed by any CNNs

2.3 CONVOLUTIONAL NEURAL NETWORK

ANN is a buildup of several nodes called neurons, which are fully connected and simple architecture producing an arrangement of real-valued activations. Neurons influence learning by forwarding the information. Neurons get activated through weighted links from the previous layer's active neurons. Learning is about obtaining weights that make neural networks exhibit the required behavior. Depending on the behavior and orientation of neurons, such behavior may need many computations stages where each stage enables the network to learn by adjusting weights called the training phase. Deep learning emerged as an amazing technology that gave us the potential to enhance learning in every field of human life[21], [23], [44]. CNN is very similar to other neural nets and is a well-known algorithm in deep learning. In 1980, K. Fukushima proposed the neocognitron [45], which is also regarded as the predecessor of CNN. This discovery was based on the findings of Hubel & Wiesel[46] describing that mammals visually perceive the world using a layered structure of neurons. [47], [48] established the modern framework of CNN.

There are many variants of CNN as I shall discuss in the following section but the basic architecture of CNN remains the same, broadly consisting –

1. Convolutional
2. Pooling, and
3. Fully-connected Layer.

Convolution layer consists of many convolutions learned kernels that convolves the input to compute feature maps by using the kernel. Further, the activation function is applied to the convolved output. The pooling layer aims at shift-invariance by dropping the resolution of the feature map, usually sandwiched between two convolution layers. A fully connected layer (one or more) aims to perform high-level reasoning and generate global semantic data.

2.3.1 COMPONENTS OF CNN

Components of CNN are summarized in table 2.1. The simplified representation of the CNN Architecture is represented by fig. 2.2.

Table 2.1 Components of CNN

Components of CNN	**Types**
Convolution	Network in network
	Dilated
	Tiled
	Transposed module
	Inception module
Pooling	Mixed
	L_p pooling
	Stochastic
	Multi-Scale orderless
	Spectral
Activation Function	Rectified Linear Unit (ReLU)
	Parametric ReLU
	Randomized ReLU
	Leaky ReLU
	Maxout
	Probout
	Exponential linear unit (ELU)

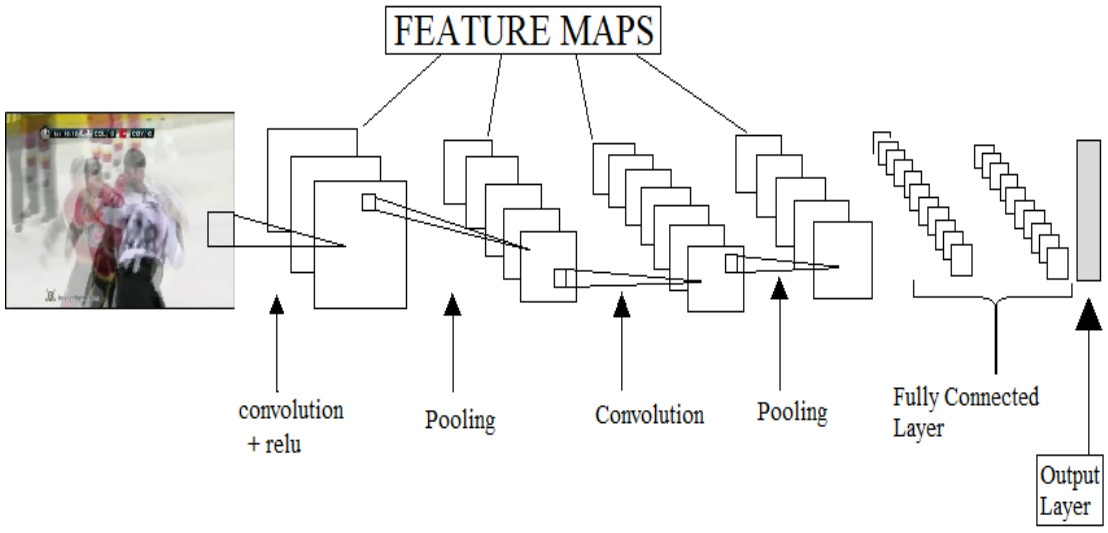

Figure 2.2 Simplified Representation of CNN Architecture

2.3.2 POPULAR VARIANTS OF CNN

In the last decade, we have seen the birth of numerous CNNs. Even the visualization of these models has become extremely challenging because these networks have gotten very deep so lately we stopped keeping track of them and started treating them as black-box models. A plethora of CNN architectures which we see today is a result of various things like improved hardware, ImageNet competition, and new ideas, solving specialized tasks, larger datasets, and new algorithms, and so on. There are various modifications done in existing CNN architecture which improved the network and their performance over time. With numerous variants of CNN, neural networks are slowly becoming for a data scientist or machine learning practitioners what linear regression was one for statisticians.

Table 2.2 summarizes 10 common variants of CNN used by researchers to optimize the solution for their problem statements. Although there are other variants as well but exploring each variant is not done as of now.

Table 2.2 10 Common Variants of CNN

Name	Year	Parameters	Key Features
LeNet-5	1998	60,000	- Standard template. - Stacking convolutional and pooling layers, followed with fully-connected layers
AlexNet	2012	60 Million	- First to implement ReLU as an activation function
VGG-16	2014	138 Million	- Contributed to the designing of deeper networks
Inception-v1	2014	5 Million	- Rather than stacking convolutional layers, modules/blocks were stacked within which are convolutional layers
Inception-v3	2015	24 Million	- Among the first to use batch normalization

ResNet-50	2015	26 Million	- Though this was not the first to introduce skip connections but popularized skip connections
Xception	2016	23 Million	- CNN based entirely upon depthwise separable convolutional layers.
Inception-v4	2016	43 Million	- Increased model size due to which it worked better than previous versions.
Inception-ResNets	2016	56 Million	- Inception modules converted into Residual Inception Blocks - More Inception modules were added
ResNeXt-50	2017	25 Million	- Scaling up the number of parallel towers within a module

2.4 TRANSFER LEARNING

Previously, with no development of transfer learning, the training of independent models was done using conventional learning. When transitioning from one network configuration to another, no knowledge was kept. Transfer learning attempts to solve related tasks in the same source domain and is used to improve a model's efficiency and speed up the teaching. In other words, in a completely new model, it can reuse the weights from a pre-trained model for multiple layers. Transfer learning, with the aid of previously trained network, can take advantage of expertise such as characteristics, weights, and so on. The advantages of Transfer learning is listed down below:

• Re-use of the trained model

• Performance improvement of the model in many cases like multiclass classification.

• Accelerated training process through feature extraction or weight initializing strategy

A network first trained on a dataset then the learned features get remodeled or passed to another target network for training purposes onto the dataset. One of the drawbacks of Transfer learning is Overfitting if we have small dataset and a large value of parameters but is not a problem in the large dataset with less parameters that in-fact helps in fine-tuning and improves the overall performance.

2.5 STATE-OF-THE-ARTS FOR VIOLENCE DETECTION

Many variations are proposed to already existing Convolution Neural Network architectures, as well as additional deep learning architectures, which are added to make the classification of violent actions much more accurate. Deep learning architectures, are also based on neural networks, classify violent actions based on the dataset and extracted features using more convolution layers. If I deconstruct a video it is compiled of various frames of images. Video is nothing but an illusion of continuous moving images. If I can classify the image in the frame into violent or not, I can classify the scene of a video as well. Here, the advancement in the violence detection techniques is exhaustively explored which uses a CNN algorithm in the proposed frameworks. The tally of such methods that use existing Convolution Neural Network-based classification is summarized in Table 2.3.

Violence Detection systems dependent on deep learning like [29]–[31] are burgeoning with ever-increasing accuracies and advantages over the other. The evolutions of single-stream architectures into multi-stream state-of-the-art have also shown a potential rise in performance since these architectures incorporate the amalgamation of different information cues – motion, spatial, and temporal. Previous works based on extracting spatiotemporal features from video or pyramid of RGB image frames extracted from videos have shown tremendous potential in detecting violence with accuracies nearing almost 100% mark.

Table 2.3 Violence Detection Techniques Using CNN

Ref. No.	Key Features	Advantage	Short Comings	Accuracy/ Dataset
[17]	- SELayer-3D CNN	-higher detection accuracy and efficiency - fast real-time processing - 500 FPS processed in real-time	- Reduced accuracy due to low ratio of violent samples - require big enough training dataset for violence videos	- accuracy of 98.08% on the Crowd Violence dataset - 99.0% accuracy on the Hockey Fight dataset
[18]	- Deep CNN (ResNet50 / VGG19 / Xception) for spatial feature information - Bi-LSTM to learn a video-level classifier	- very effective model to detect violence	- Bigger dataset required	Resnet50 + LSTM : - 75.73% accuracy on Real-violent dataset - 88.74% accuracy on movie dataset - 83.19% accuracy on Hockey Fight dataset
[20]	Spatio-temporal Encoder, built on BiConvLSTM	- BiConvLSTM performs better than ConvLSTM on more heterogeneous and complex datasets like Violent Flows	- the method is not consistently best, require more training data for better and accurate results	- Hockey dataset: 96.96±1.08 % accuracy - Movies dataset: 100±0% accuracy - Violent Flows

				dataset: 92.18±3.29 %. accuracy
[27]	ScatterNet hybrid DL network	- needs fewer learning examples - real-time processing is done using cloud	- vast amounts of unlimited computational power and resources are required	- 87.6% accuracy. on the Aerial Violent Individual (AVI) dataset
[28]	CNN + LSTM	- speed is 4x faster than the fastest known model, 131 FPS	- The model doesn't perform to the fullest due to lesser availability of data	- accuracy of 98 % on the Hockey Fight Dataset
[39]	3D ConvNets Total 9 layers. Supervised learning and back-propagation - no prior knowledge required	- efficient and accurate - directly operate on the image pixels -Automatically learn the video features	- less accuracy - Possibility to detect mid-level concepts	91.00% accuracy rate on hockey dataset
[50]	Deep residual Architecture-Resnet Crowd for crowd density level classification, violent behavior detection, and simultaneous crowd counting	- individual task performance boosted using multi-task approach most notably for violent behavior detection	- lack of an appropriately labeled multi-task dataset	- 9% boost in ROC curve AUC on [50] dataset
[51]	2 approaches to use CNNs for classification 1. use in the end-to-end fashion, take raw features as input and at its last layer produce the classification result 2. use CNNs for feature	- explicit violent relevance Video clips consisting of shot, scream and heavy metal music, have high scores	- audio-level violent identified as non-violent - non-violent videos with elements like cheer, shutter sound, clapping, etc. get high	- Average Precision(AP) of 0.485 and 0.291 on MediaEval 2015 dataset on the validation

	representations.		scores.	and testing sets
[52]	- extracting frame-level characteristics from a video using AlexNet model as CNN LSTM variant aggregates the frame-level characteristics.	Localized Spatio-temporal features can be captured using CNN + convLSTM helping in the analysis of local motion - violent video classification by an end-to-end trainable deep NeuralNet	- unable to distinguish violent actions, marks certain videos as non-violent, and unable to outperform the previous technique on Violent-Flows dataset,	97.1±0.55% accuracy on hockey dataset 100±0% accuracy on movies dataset 94.57±2.34 % accuracy on Violent-Flows dataset
[53]	- extracting the convolutional feature maps using 2 Deep CNN models as an extractor	- capture long-term action information by integrating Deep NN and improved trajectory - can be expanded to crowd scene videos	- detection of violence regions are yet to be taken into consideration	- approx. 98% on Hockey Fights dataset - approx. 92.5% accuracy on Crowd Violence dataset
[54]	- pre-trained light-weight MobileNet CNN - 3D CNN for feature extraction	- nearest security department or a police station is alerted if violence detected - unnecessary processing of useless frames is reduced - achieved better accuracy	- require good hardware - resource constraint devices might not be able to deploy the proposed model	- violent crowd dataset: 98% acc - movies dataset: 99% acc - hockey fight dataset: 96% acc
[55]	representative image classified and the final decision for the sequence obtained using 2D CNN	- a good tradeoff between accuracy and computational time. - real-time application - fast enough	- better results are achieved only because the camera position was static	- 99% accuracy on movies dataset -94.6% accuracy on

				hockey dataset - 91.42% accuracy on Behave dataset
[56]	Convolutional Neural Networks classifier	- existing information or pre-processing not required -Large scale sports video classification	- low accuracy level - High computation	65.4% accuracy achieved, by Fine-tuning top 3 layers, on UCF-101 dataset
[57]	AlexNet based CNN with a subset of ImageNet classes - 2-stream CNN for extracting characteristics on motion optical flows and static frames - longer-term temporal dynamics captured using LSTM applied on top of the 2-stream CNN characteristics	- very effective model to detect violence	- additional computational complexity	- Mean AP 0.296 in the violence detection Subtask - In the induced effect detection subtask, accuracy of 0.418 and 0.488 for arousal and valence respectively
[58]	Convolutional Neural Networks classifier	- needs lesser computational resources - the model can be trained on CPU only	- model not tested on the popular benchmark datasets	- 95% accuracy on [58] dataset
[59]	2 DNNs frameworks - 3D-based convolutional neural network + CNN-LSTM	- tries to incorporate more specialized concepts on what can be classified as a	- lesser accuracy than other accurate and highly effective models	-MediaEval -2013-VSD data set: 63% acc

		violent scene in the foreground of violence detection		
[60]	pre-trained deep learning model VGG-Net	---	- Needs more accuracy improvements	- accuracy of 75.00% on hockey fight dataset

Motion information is a rich tool for gathering knowledge dependent on human action. Techniques such as optical flow, optical flow based ViF, OViF are used to capture motion information. Optical flow-based methodologies record the optical-flow within consecutive image frames encapsulating the motion using key components. [34] used ViF descriptors, object detection method, for representing stats collected for short frame sequences, and bags of feature method for feature extraction. The fundamental aim of the strategy is to identify the transition of violent to non-violent action having the shortest delay since the change occurred. [36] used OViF for feature extraction and captures the essence of information about the motion magnitude change. However, only the local dynamics are provided by these methodologies, and local motion analysis is performed using simple summarization techniques. CNN can extract meaningful information from the input provided hence it is highly important to decide how we provide the video information to CNN. Recent advancements in capturing the motion information using Dynamic Images for the task of action recognition [61] have shown significant improvement in the action recognition task using motion information

2.6 DATASETS

Due to the novelty and subjectivity, the field of violence detection uses a variety of datasets available and does not contain a single main dataset to study. The datasets used for studying the performance of the models proposed by researchers that were explored are summarized in table 3.2 with their brief overview and characteristics. Most of the datasets use ACC(accuracy), which is the ratio of true results over total cases examined, and AUC(area under the curve), which means that a classifier will assign a higher probability to a positive example than a negative

example, for evaluation purpose. Recall, Precision, etc. measures usually are calculated together.

Dataset had to be accurately chosen since the huge availability of datasets. Having lots of options to use for training our model thorough description and the compatibility with the system was needed. Hence, the Hockey Fight dataset, movie, and real-life violence dataset were chosen to train and evaluate the proposed implemented model.

Table 2.4 Summary of popular datasets used for Violence Detection

Dataset Name	**Year of release**	**Description**
Caviar	2004	- 28 video sequences grouped into 6 different activity scenario having a total of 26500 labeled frames
Behave	2006	- Surveillance cameras - Only clip 1 annotated (52:11 minutes) - A fixed point of view Simulated actions - 200,000 frames
Souza	2010	- 400 videos (50% fight scenes)
Giannakopoulos	2010	- 50 clips from 10 movies (2.5 hours total)
Movies	2011	- 200 video clips from action movies
Hockey Fight	2011	- 1000 clips - 50 frames per clip
Perperis	2011	- 25 movie segments (approx. 1 min each)
UFC 101	2012	- 13,000 clips
Violent Flows	2012	- 246 videos from YouTube (1-7 seconds)
SBU Kinect Interactions	2012	- 21 sets of 8 violent interactions - total of 300 interactions approx.

Violent Scene Dataset	2014	-Train: 25 Hollywood movies, 32678 shots -Test: 7 Hollywood movies, 11245 shots - 86 web videos
LIRIS-ACCEDE (media eval)	2015	- 10900 short video clips (8-12 seconds) -Train:100 movies, 6144 shots - Test:99 movies, 4756 shots
RE-DID	2015	- Real Life Scenarios - urban fights - 30 videos Length: 0:20 to 4:20
KARD	2017	- 18 activities, grouped into 10 gestures and 8 actions - total of 1hr of videos - 540 sequences having FPS = 30 frames/s and resolution of 640 × 480 pixels
Crowd-11	2017	- 6272 videos - 3005 scenes - 621,196 frames

CHAPTER 3

THE PROPOSED WORK

3.1 THE PROPOSED ARCHITECTURE

The proposed deep convolutional neural net architecture, fine-tuned pre-trained Inception Resnet V2 based CNN model for Violence Detection is shown in Fig.3.1. The architecture is designed to learn motion features and use motion information for the task of Violence Detection. The motion content or the motion information present in the video is captured by transforming RGB video into Dynamic Motion Image (DMI). The transfer learning approach is used for recognizing violent and non-violent acts with the assistance of fine-tuned pre-trained Inception-Resnet-V2. DMIs typically focuses on the salient object's movement by combining and averaging the background pixels as well as the movement patterns while retaining the long-term kinetics. The shape of the obtained DMI is identical to the original frame. The constructed DMI is transferred through the implemented architecture to compute the motion characteristics.

The architecture is a combination of Inception ResnetV2, pre-trained on the ImageNet Dataset, followed by a set of Dense and Dropout Layers. ReLU is used as the activation function on the Dense Layers expects the end Dense Layer. Last Dense Layer used the Softmax Activation. The architecture of our proposed model is as follows:

Input DMI (224x224x3) – Inception-Resnet-V2 () –Intermediate Vector Shape (5x5x1536) – Dense Layer (1024) – DropOut Layer (0.5) – ReLU – Dense Layer (512) – DropOut Layer (0.5) – ReLU – Dense Layer (256) – DropOut Layer (0.5) – ReLU – Dense Layer (128) – DropOut Layer (0.5) – ReLU – Dense Layer (2) – Softmax ()

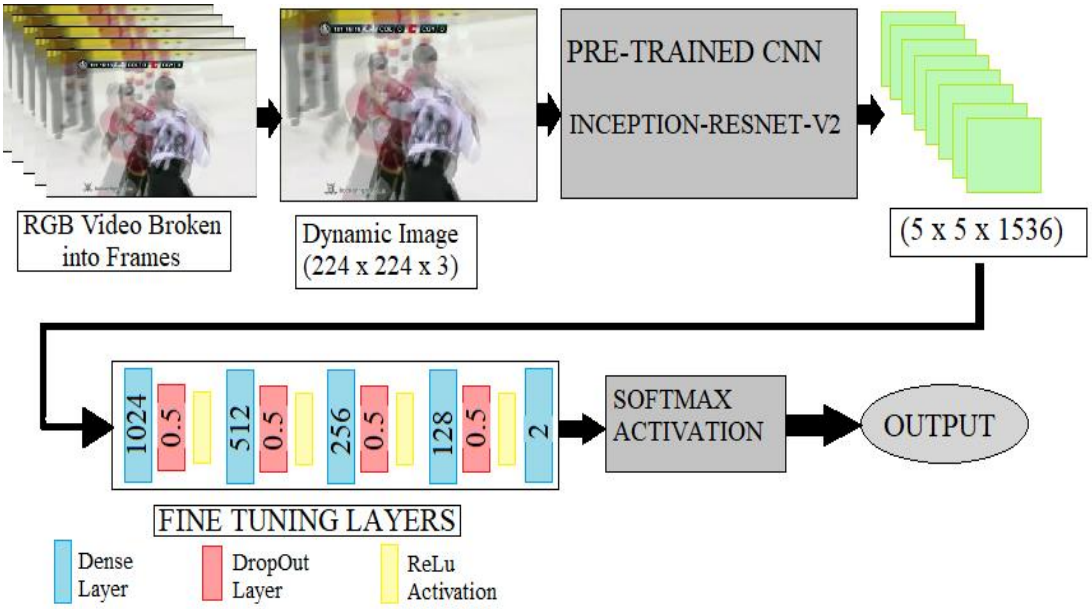

Figure 3.1 Block Diagram of Proposed Deep ConvNet

The pre-trained Inception-Resnet-V2 layers are not trained and kept frozen during the whole training duration. The layers for fine-tuning the model are trained end-to-end for updating weights according to the training sample. Since the data size is small, overfitting will be an issue. To counter it, the Dropout Layer helps to handle the overfitting phenomena as well as the training data size was also doubled by capturing the 2 dynamic images from a single video, created using distinct mutually exclusive frames meaning the frame used to create first DMI was not used for creating the second DMI. Early stopping was used so that model doesn't over-train. Without early stopping the validation loss increases and accuracy degrades. Model Checkpoint is used to obtain the best-trained model weights during training having the lowest validation loss. For testing purposes, the same saved model checkpoint having the best weights is used to achieve the desired results. The same network architecture was used to test the results by replacing Inception-Resent-V2 with Resent-50 and Inception-V3 as the pre-trained CNN. The model summary for the fine-tuned layers can be seen in fig. 3.2

```
model.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 1024)              39322624
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_4 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_4 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 258
=================================================================
Total params: 40,011,906
Trainable params: 40,011,906
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.2 Fine-Tune Layer Model Summary

The sole objective of this research is to establish the importance of the distinct motion features embedded and encoded as RGB-DMIs and how a deep learning convolutional neural net can be used to capture such motion characteristics and improve the task of violence detection. The fine-tuned pre-trained Inception-Resnet-V2, Inception-V3, and Resnet50 models are used to evaluate how well the proposed framework performs and establish the novelty of the work on the three publically available datasets or benchmarks – hockey fight, real-life violence, and movie dataset.

3.2 PRE-TRAINED MODELS

Due to limited data available in the datasets as well as the computational limitations, it was not possible to create and train a new convolutional neural network. The pre-trained models like Inception-Resnets, Resnets, Inception series

CNN were a good fit since they are already trained on the ImageNet dataset making them an efficient choice for feature extraction. Pre-trained models incorporate a transfer learning mechanism wherein the trained model on a different dataset is utilized for either feature extraction or object identification or classification since these pre-trained models have already learned a lot during their training phase. Also, they prove their high effectiveness in the cases where the data is limited and over-fitting needs to be avoided. During the conduct of this research, all of the pre-trained models were analyzed and upon careful analysis and recent state-of-the-art it was concluded that our proposed architecture be evaluated using the Inception-Resnet-V2, Resent-50 and Inception –V3 due to their top 1 and top 5 percent accuracy. Each of which had an edge over another with Inception-Resnet-V2 being the best among the lot in terms of accuracy, performance, and recent state-of-the-art.

3.3 INPUT FRAMES PROCESSING

The RGB video from the datasets – Hockey fight, Real-Life Violence, and Movie datasets are broken down into RGB frames so that later these RGB frames can be used for the creation of dynamic motion images. The libraries used in python for the fetching and processing of the RGB frames are OpenCV and pandas. A pandas data frame was created to store the video names and paths which in turn were used to capture the video using the VideoCapture method of CV2. A frame rate of each video was captured using the inbuilt function of the OpenCV library and accordingly for each frame number which was required for dynamic motion image creation was processed and stored using the imwrite method of the OpenCV library.

3.4 DYNAMIC MOTION IMAGE CREATION

Deep Neural Nets can automatically learn powerful features but only operate within the confines of a specific hand-crafted architecture. While designing architecture it is required to visualize how the video must be presented to CNN. Standard solutions used to date included sub-videos of fixed length or duration as arrays or using recurrent architecture but DMI proposed an efficient approach that could summarize the video into a single still image and used this image for the task

of action recognition. RP and ARP techniques were devised and latter being much faster in computation.

A modified approach of approximate rank pooling was used in our proposed work. From RGB video, we compute 2 dynamic motion images by adopting the methodology proposed by [61]. RGB video was broken into a stack of frames and 2 sets of frames were produced, one starting with 1st frame having stride of frameRate/2 and the other set starting with 1st frame = frameRate/4 and having stride of frameRate/2. Each frame in the 2 sets was taken and split into R, G, and B channels separately. A coefficient was computed for each frame and a list of frames split by channels was multiplied by the coefficients. The weighted aggregate of all 3 channels - R, G, and B in each frame separately is collected to obtain the R, G, B channel of the DMI. Once we achieved the R, G, B channels of DMI, they were then merged to form a single DMI normalized into pixels having a value between 0-255. The size of the generated DMI was the same as the original frame. The DMI creation process is shown in fig 3.3

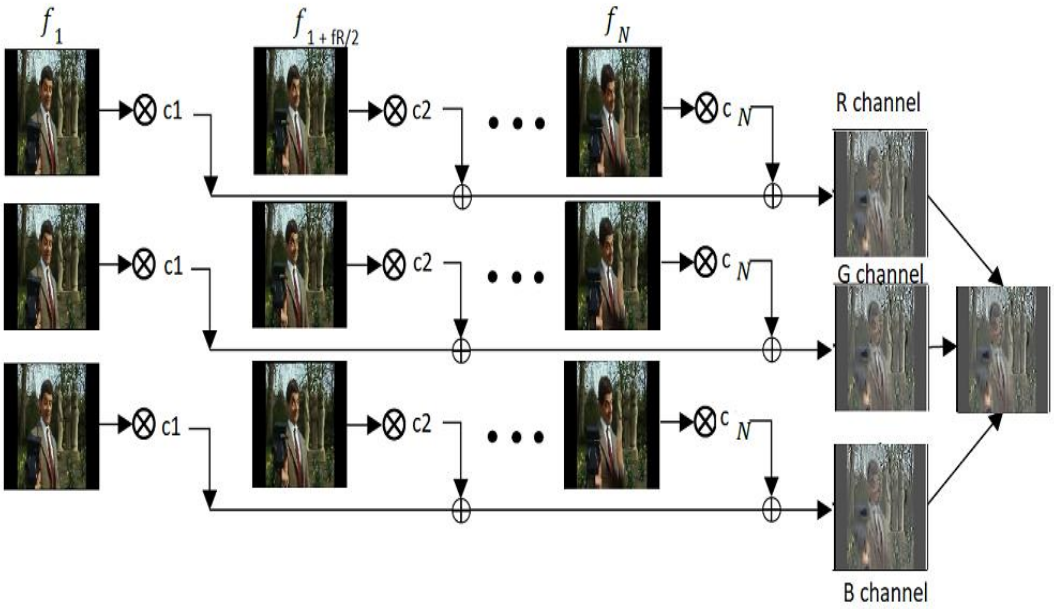

Figure 3.3 Dynamic Motion Image Creation

3.5 CALLBACKS

Callbacks are a very important and highly used feature which can help in performing various actions during various phases in training like at the start or end of an epoch, before or after a single batch, etc. Callbacks can help in visualizing training of model, prevent overfitting and build better models by implementing callback like EarlyStopping or customized learning rate, etc. Various callbacks available in Keras API are:

- Base Callback class
- TensorBoard
- RemoteMonitor
- LearningRateScheduler
- ProgbarLogger
- EarlyStopping
- TerminateOnNaN
- ModelCheckpoint
- ReduceLROnPlateau
- LambdaCallback
- CSVLogger

Apart from in-built callbacks, custom callbacks can also be used to implement simple and powerful implementation to help ease the training and visualization as per need. During the training of the proposed model, 2 callbacks were used namely – Early Stopping and Model Checkpoint. The callbacks used are explained briefly in the following sub-section.

3.5.1 EARLY STOPPING

Early stopping prevents overtraining your model by terminating the training process if it's not learning anything. This is pretty flexible — you can control what metric to monitor, how much it needs to change to be considered "still learning", and how many epochs in a row it can falter before the model stops training. Some relevant parameters:

- monitor: value being monitored
- patience: number of epochs with no improvement after which training will be stopped
- min_delta: minimum change in the monitored value.
- restore_best_weights: set to True if you want to keep the best weights once stopped

3.5.2 MODEL CHECKPOINT

This callback is especially handy for models where epochs take an extremely long time and save model as a checkpoint file (in hdf5 format) to disk after each successful epoch. The file can be dynamically named. Some relevant parameters:

- file-path: OS directory path
- mode: auto, min, or max
- monitor: the value being monitored
- save_best_only: set to True if you want to preserve the latest best model

3.6 MODEL TRAINING AND PREDICTION

Since the dataset contained limited videos that weren't sufficient for training a new model end to end and computational limitation, the pre-trained model on the ImageNet dataset served as an efficient structure for conduction the training. Pre-trained models for motion feature extraction and then passing on the extracted motion features to the sequential model comprised of dense and dropout layers helped in achieving a better framework for violence detection. The pre-trained network's layers were frozen during the training and only the sequential model was supposed to be trained. DMI was supplied to our proposed method and various experiments were done using SGD, Adam, and Nadam optimizer until it was established that Adam optimizer was working best on the dataset. During the training of the proposed model, callbacks were used for better management. The model checkpoint was used to save the best-trained weights so far on the condition that the model had minimum validation loss. Since over-fitting could be an issue so Early

stopping was also used to stop the training if no progress was seen in the value of validation loss.

Once training was completed, for prediction on unseen data, saved model checkpoint was loaded and RGB video was converted into DMI, and the prediction was done. For predicting, the network was recreated as was used during training. Once all the test videos were passed on to the network for prediction, the model evaluation was performed based on the testing phase accuracy, loss, and f1 score values. Although only testing accuracies were later used for establishing the performance results.

3.7 ACTIVATION FUNCTION

Referring to the cerebral cortex, researchers were able to assert that the true cause of universal approximation capacity was the NN structure but they also neglected to identify the side effects of implementation by ignoring the activation function option. This hypothesis motivated researchers to dive deeper and explore more complicated activations to decrease network complexity.

The performance of a NN depends on the quality of the activation function used and is responsible for the precision, the computational efficiency of the model when training our algorithm. The speed of the Converge and Convergence is also regulated by activation functions, and in some situations, it may prevent the network from converging at the initial point. The primary purpose of an activation function is to transform an input signal to a node's output signal in an ANN where this output signal is then used as a reference to the next input layer. NN operates on non-linear activation functions that make the network know more complex features and also helps solve and lean complex mathematical representations and provides correct predictions. The activation functions used in the proposed model are defined in the following sub-section.

3.7.1 ReLU

Rectified Linear Unit or ReLU activation function is widely used due to its wide range in the non-negative axis i.e. (0 to infinite) and its quick convergence property.

3.7.2 SOFTMAX

This function exhibits the property of cumulative distribution functions. The range of Softmax function is (0 to 1) and generally used for Binary Classification.

CHAPTER 4

EXPERIMENTAL WORK AND RESULT

4.1 OVERVIEW

For substantiating the achievement of our proposed violence detection framework, three publically available benchmarks – Hockey Fight Dataset, Real Life Violence Dataset, and Movie Datasets are used. The RGB videos in the datasets are transformed into Dynamic Motion Images based on the technique discussed in the previous section. These Dynamic Motion Images are pre-computed so that training can begin smoothly with the data prepared beforehand. The dataset individually is split into training and testing sets and the filenames and path of the videos in each training and testing set are stored in different files which can be accessed when the conversion of RGB video into DMI is started.

4.2 SETUP AND DATASET

In the experiment, the end-to-end training of fine-tuned layers of the proposed model takes place. Before the training phase begins, the dataset is subdivided into testing and training samples using the 30-70 splitting strategy, and training sample videos are processed to get dynamic images. For each training, once DMI was obtained for the training sample, motion features were extracted from DMI using pre-trained Inception-Resnet-V2 on ImageNet dataset followed by fine-tuning layers using Adam optimizer with epoch = 50, and batch size = 64. Callbacks were used to save the best model weights based on minimum validation loss and early stopping with patience = 20 was used to stop the model in case model weights were not improving, hence accelerating the training process. Training input was further divided into actual training set and validation set using 80-20 splitting strategy. In Testing, the model checkpoint saved during training is loaded and DMI is computed for each testing video simultaneously.

4.2.1 HOCKEY FIGHT DATASET

The Hockey Fight Dataset was introduced by [62] containing 1000 violent and non-violent action clips each, recorded during a hockey game of the NHL. Each video has a frame rate of 25fps consisting of 50frames of 720x576 pixels. The dataset was labeled into a fight and non-fight category. All the clips have the same background with only Ice Hockey players entering the frames. The illustrations of DMI obtained for violent fights are shown in fig. 4.1 and the DMI obtained for the non-violence are shown in fig. 4.2.

Figure 4.1 DMI for violent scenes obtained from Hockey Fight Dataset

Figure 4.2 DMI for Non-violent scenes obtained from Hockey Fight Dataset

4.2.2 REAL-LIFE VIOLENCE DATASET

[63] Introduced a new and quite challenging benchmark, The Real-Life Violence Dataset, as compared to Hockey Fight and movie dataset with 1000 videos of violent and non-violent genre each, having a wide range of gender, age, and race collected from diverse backgrounds and environment. Violent videos are captured in an environment like a prison, street, and schools whereas the non-violent videos are captured in environments like sports fields and arena featuring events such as swimming, doing archery, playing basketball. The average duration of the clip is 5seconds with maximum duration being 7seconds and a minimum duration of 3 seconds with a frame rate of 25fps. Fig. 4.3 and 4.4 show the example of violent and non-violent DMI created from the Real-Violence Dataset.

Figure 4.3 DMI for violent scenes obtained from Real-Life Violence Dataset

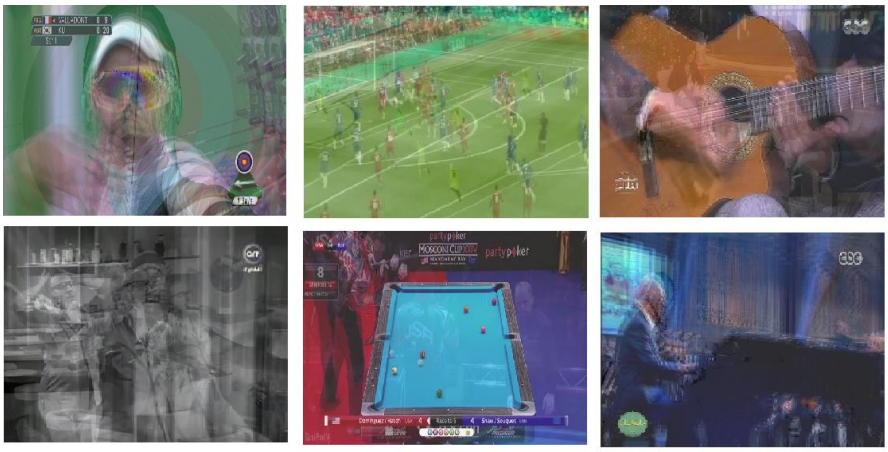

Figure 4.4 DMI for Non-violent scenes obtained from Real-Life Violence Dataset

4.2.3 MOVIE DATASET

The Movie Dataset [62] is a considerably small dataset containing only 200 movie clips from action scenes bifurcated equally into violent and non-violent video genres. The movie dataset has also varying backgrounds. The clips have an average duration of 1 second and some clips have a duration of 2seconds, having a frame rate of either 25fps or 30fps. Fig. 4.5 and 4.6 show the example of violent and non-violent DMI created from Movie Dataset.

Figure 4.5 DMI for violent scenes obtained from Movie Dataset

Figure 4.6 DMI for Non-violent scenes obtained from Movie Dataset

4.3 PERFORMANCE EVALUATION

The performance evaluation of our proposed model – pre-trained Inception-Resnet-V2 with fine-tuning layers was done on the 3 publically available benchmarks mentioned above and the results were recorded for comparative analysis. Along with the Inception-Resnet-V2, the results were also obtained on the Resnet-50 and Inception-V3 pre-trained models to see the performance on these models as well. All the experiments were done using Jupyter Notebook running on top of the Anaconda environment, the computation was performed with the help of the i5 8^{th} gen processor. The programming language used to develop the proposed model was python 3. The plots of model accuracy and model loss per epoch during the training phase for each experiment were also obtained using the matplotlib library. The performance evaluation results of the proposed model tested on Inception-Resnet-V2 architecture and tested on Resnet-50 as well as Inception-V3 are shared in the following sub-section and organized as follows: sub-section 4.3.1 contains the results for Inception-Resnet-V2 based proposed architecture for all 3 publically available benchmarks stated in the previous section. Sub-section 4.3.2 and 4.3.3 contain the result for Resnet-50 and Inception-V3 based architecture respectively.

4.3.1 INCEPTION-RESNET-V2

Table 4.1 shows the training and validation accuracy and loss occurred during the training phase as well as the testing accuracy achieved for all 3 datasets. It is worth noting that even though training accuracy reaches a 100% mark, training loss is yet a non-zero value. The reason behind this behavior is that accuracy and loss often appear to be inversely proportional but there is no concrete mathematical relationship between these two metrics [64]. Accuracy can be seen as a count of correct predictions whereas loss can be seen as a distance between predicted probability and true value. The plot for training accuracy v/s validation accuracy achieved over each epoch for Inception-Resnet-V2 based proposed model, as well as training loss v/s validation loss over each epoch for all the 3 datasets used, are shown in figures 4.7 to 4.12.

Table 4.1 Training, Validation, and Testing Accuracy along-with Training, Validation Loss achieved on the 3 publically available benchmarks for Inception-Resnet-V2 based proposed model

Dataset Name	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Testing Accuracy
Hockey Fight Dataset	100 %	0.0212 %	94.99 %	17.18 %	93.33 %
Real-Life Violence Dataset	100 %	0.0252 %	92.8571 %	21.58 %	86.7892 %
Movie Dataset	100 %	7.30e-03 %	100 %	5.75e-05 %	100 %

1. HOCKEY FIGHT DATASET

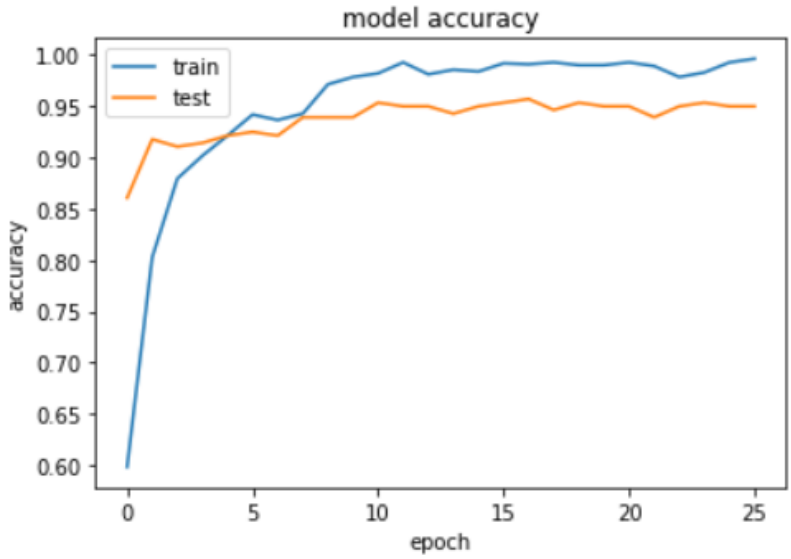

Figure 4.7 A Plot of Training Accuracy v/s Validation Accuracy on Hockey Fight Dataset for Inception-Resnet-V2

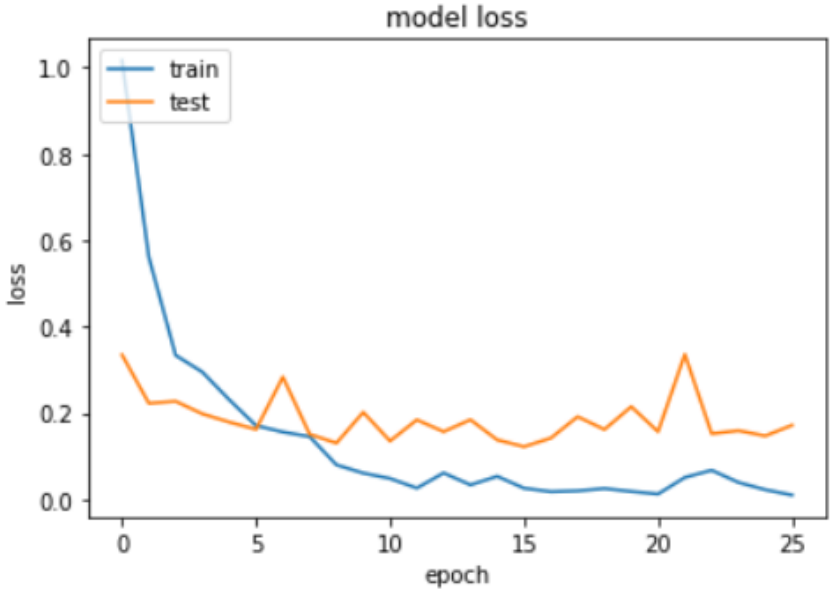

Figure 4.8 A Plot of Training Loss v/s Validation Loss on Hockey Fight Dataset for Inception-Resnet-V2

2. REAL-LIFE VIOLENCE DATASET

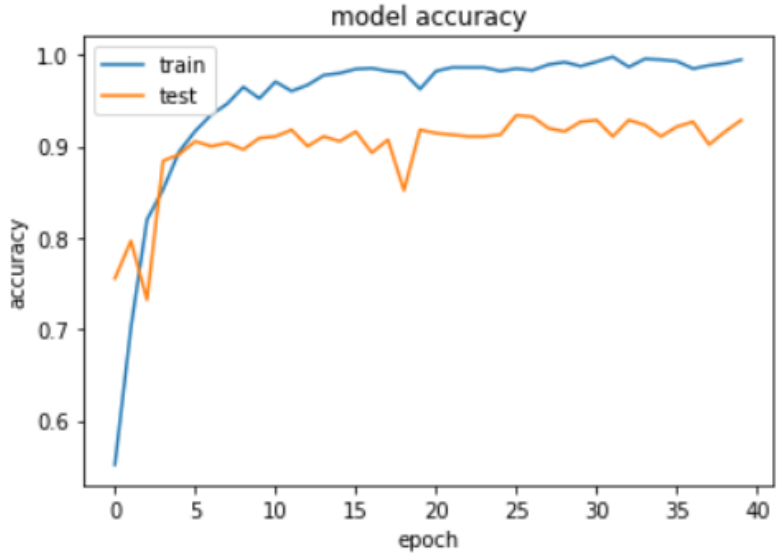

Figure 4.9 A Plot of Training Accuracy v/s Validation Accuracy on Real-Life Violence Dataset for Inception-Resnet-V2

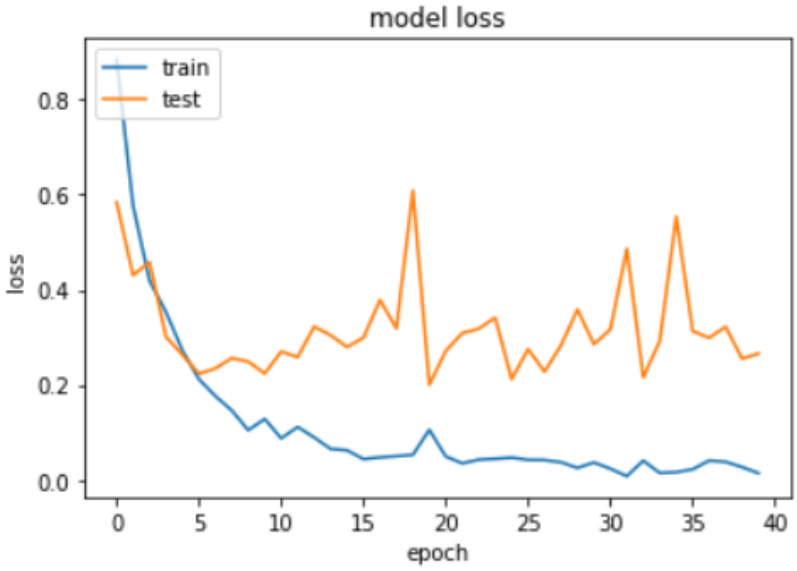

Figure 4.10 A Plot of Training Loss v/s Validation Loss on Real-Life Violence Dataset for Inception-Resnet-V2

3. MOVIE DATASET

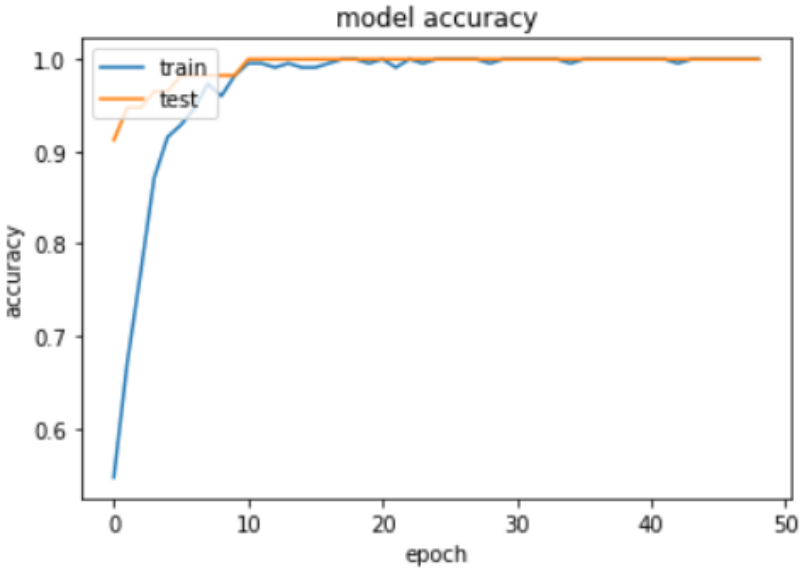

Figure 4.11 A Plot of Training Accuracy v/s Validation Accuracy on Movie Dataset for Inception-Resnet-V2

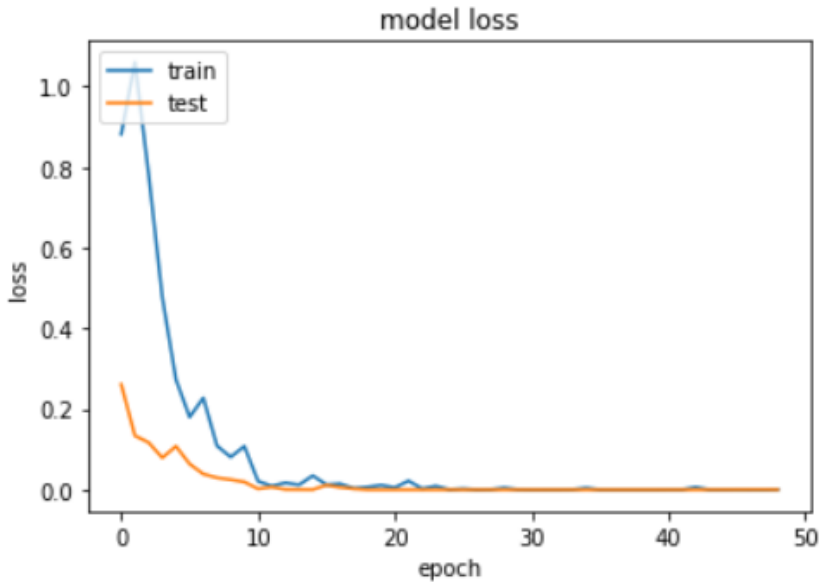

Figure 4.12 A Plot of Training Loss v/s Validation Loss on Movie Dataset for Inception-Resnet-V2

4.3.2 RESENT50

Table 4.2 shows the training and validation accuracy and loss occurred during the training phase as well as the testing accuracy achieved in Resnet-50 based architecture for all 3 datasets. The plot for training accuracy v/s validation accuracy achieved over each epoch for Resnet-50 based proposed model, as well as training loss v/s validation loss over each epoch for all the 3 datasets used, are shown in figures 4.13 to 4.18.

Table 4.2 Training, Validation, and Testing Accuracy along-with Training, Validation Loss achieved on the 3 publically available benchmarks for Resnet-50 model

Dataset Name	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Testing Accuracy
Hockey Fight Dataset	92.85 %	27.29 %	93.57 %	28.03 %	90.66 %
Real-Life Violence Dataset	50 %	69.52 %	50 %	69.5 %	49.28 %
Movie Dataset	99.11 %	4.22 %	99.9 %	1.92 %	98.33%

1. HOCKEY FIGHT DATASET

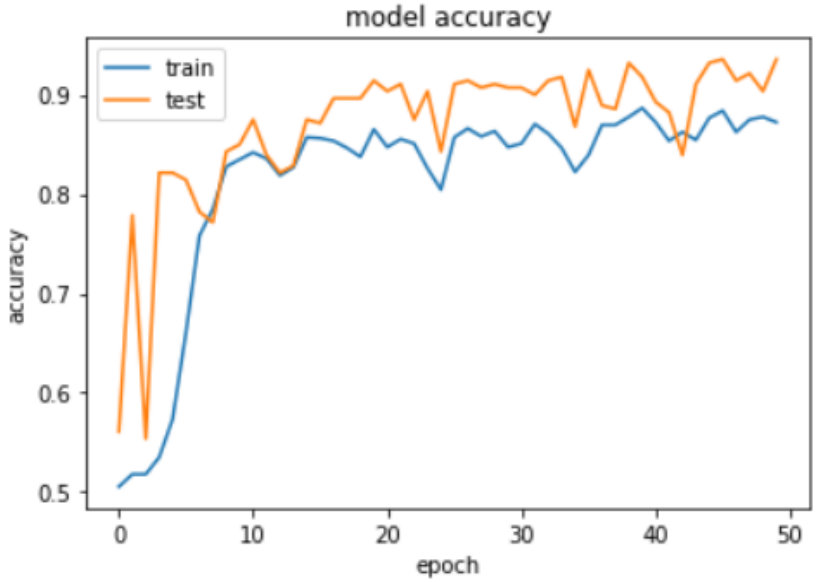

Figure 4.13 A Plot of Training Accuracy v/s Validation Accuracy on Hockey Fight Dataset for Resnet-50

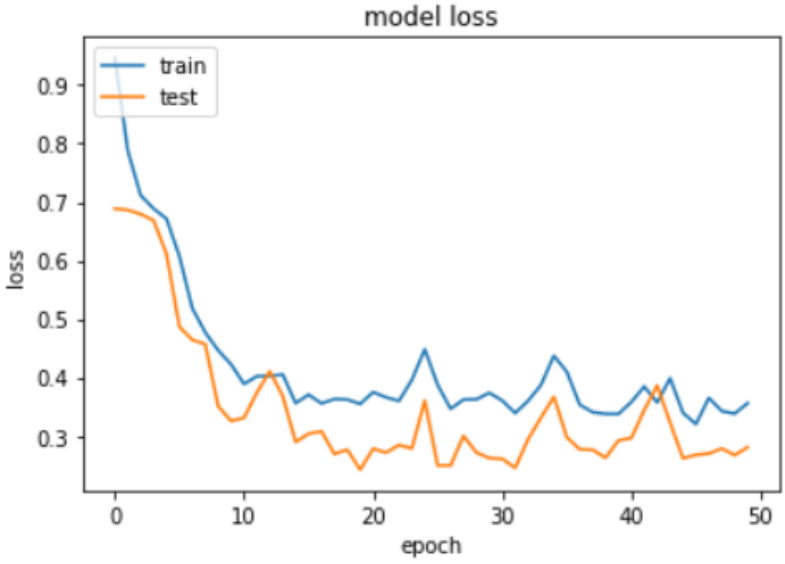

Figure 4.14 A Plot of Training Loss v/s Validation Loss on Hockey Fight Dataset for Resnet-50

2. REAL-LIFE VIOLENCE DATASET

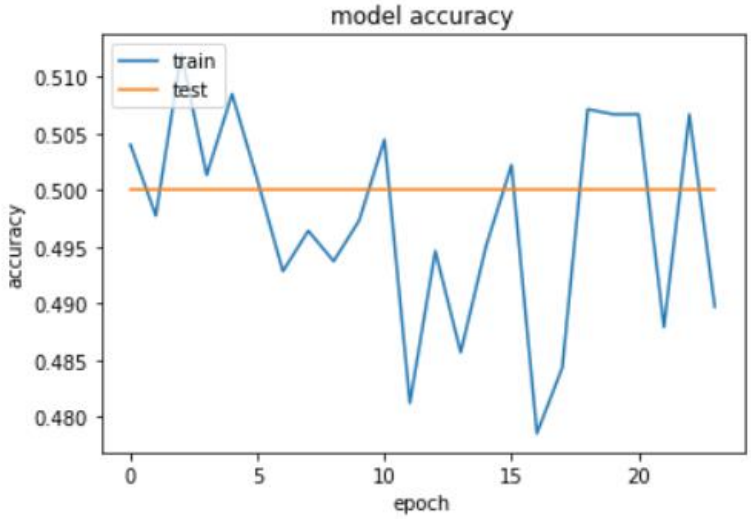

Figure 4.15 A Plot of Training Accuracy v/s Validation Accuracy on Real-Life Violence Dataset for Resnet-50

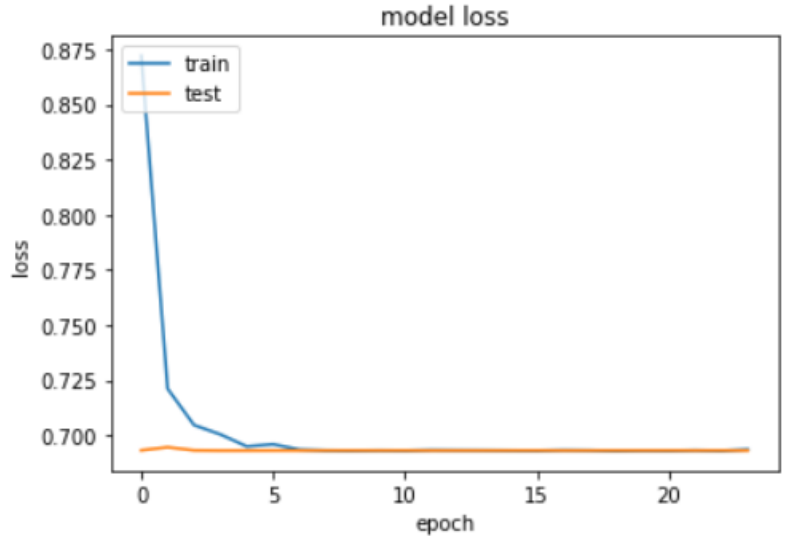

Figure 4.16 A Plot of Training Loss v/s Validation Loss on Real-Life Violence Dataset for Resnet-50

3. MOVIE DATASET

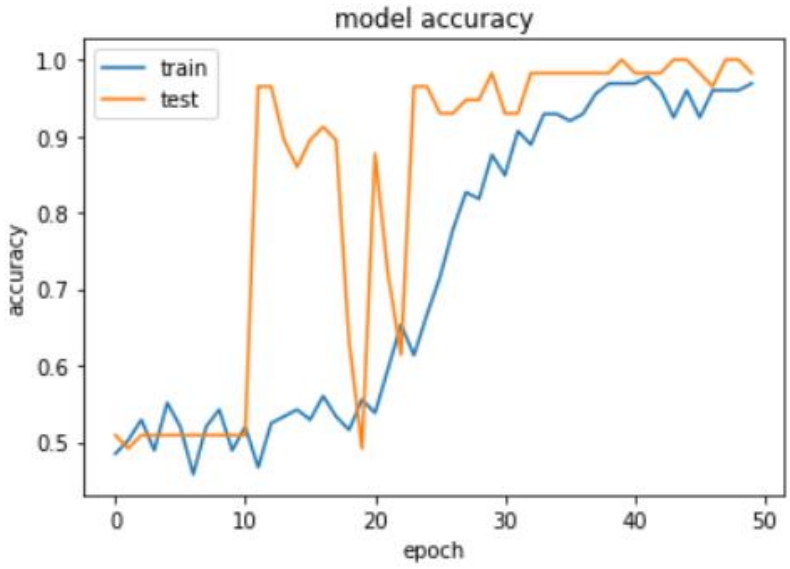

Figure 4.17 A Plot of Training Accuracy v/s Validation Accuracy on Movie Dataset for Resnet-50

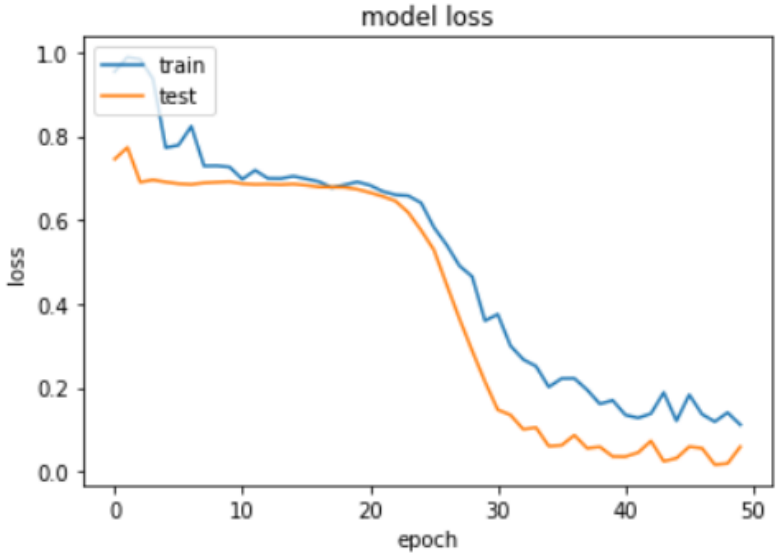

Figure 4.18 A Plot of Training Loss v/s Validation Loss on Movie Dataset for Resnet-50

4.3.3 INCEPTION-V3

Table 4.3 details the training and validation accuracy and loss occurred during the training phase as well as the testing accuracy achieved in Inception-V3 based architecture for all 3 datasets. The plot for training accuracy v/s validation accuracy achieved over each epoch for Inception-V3 based proposed model, as well

as training loss v/s validation loss over each epoch for all the 3 datasets used, are shown in figures 4.19 to 4.24.

Table 4.3 Training, Validation, and Testing Accuracy along-with Training, Validation Loss achieved on the 3 publically available benchmarks for Inception-V3 model

Dataset Name	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Testing Accuracy
Hockey Fight Dataset	99.99 %	0.2426 %	95.714 %	19.13 %	92.66 %
Real-Life Violence Dataset	100 %	0.061 %	88.57 %	26.31 %	83.166 %
Movie Dataset	100 %	1.05e-05 %	100 %	9.81e-02 %	98.33 %

1. HOCKEY FIGHT DATASET

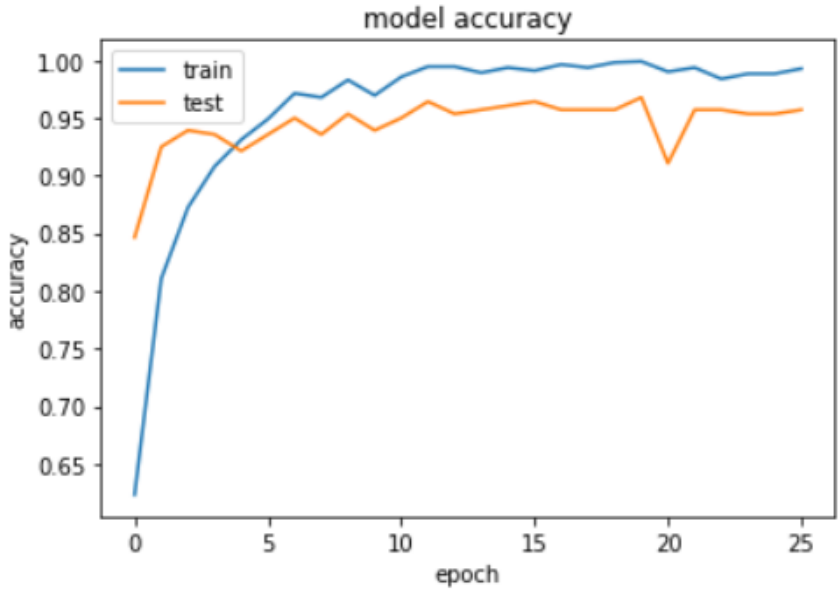

Figure 4.19 A Plot of Training Accuracy v/s Validation Accuracy on Hockey Fight Dataset for Inception-V3

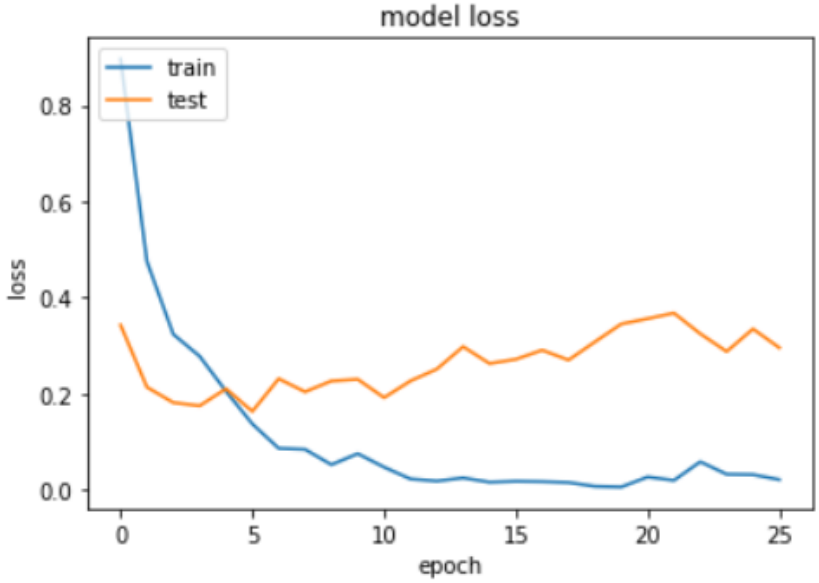

Figure 4.20 A Plot of Training Loss v/s Validation Loss on Hockey Fight Dataset for Inception-V3

2. REAL-LIFE VIOLENCE DATASET

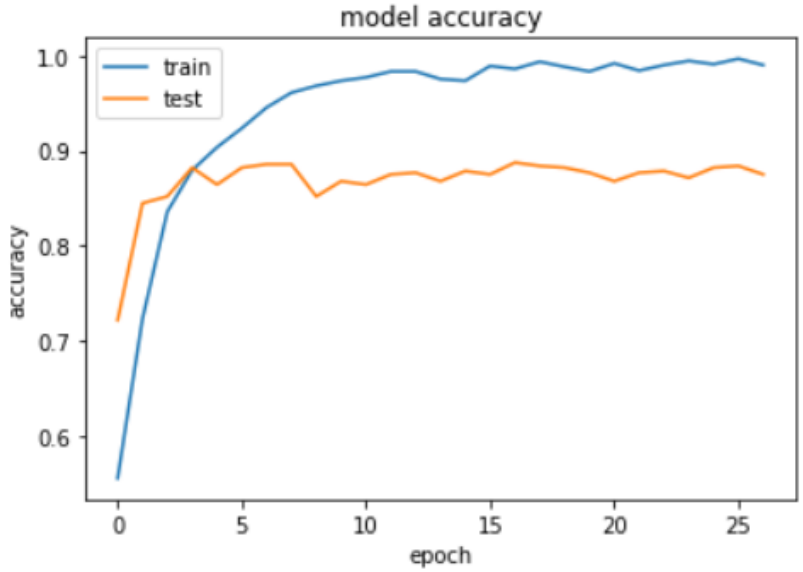

Figure 4.21 A Plot of Training Accuracy v/s Validation Accuracy on Real-Life Violence Dataset for Inception-V3

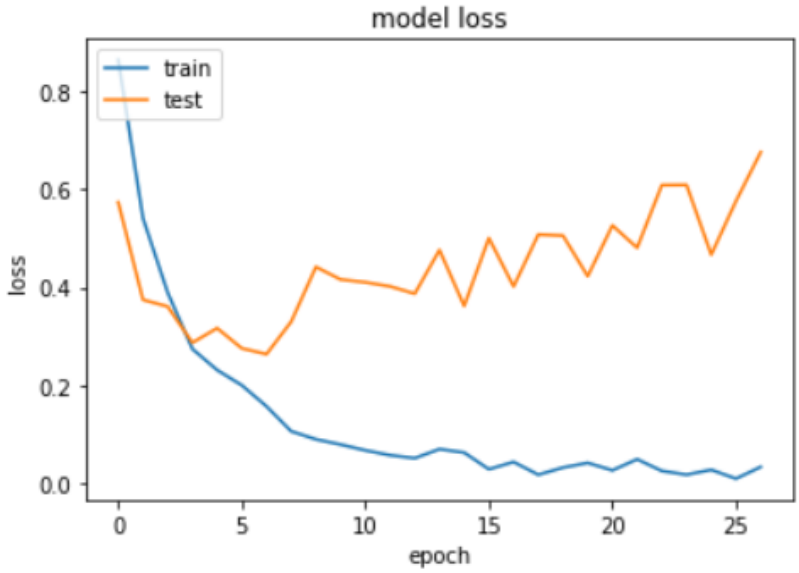

Figure 4.22 A Plot of Training Loss v/s Validation Loss on Real-Life Violence Dataset for Inception-V3

3. MOVIE DATASET

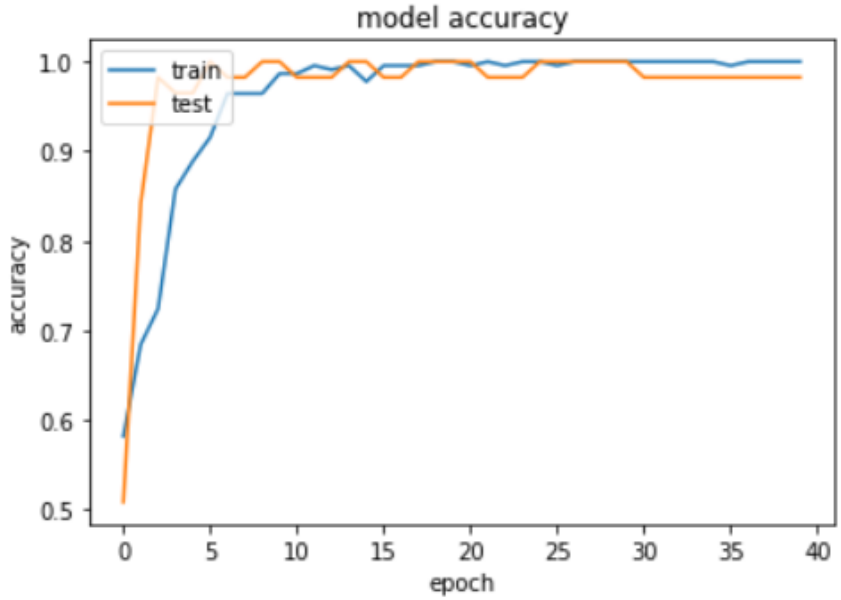

Figure 4.23 A Plot of Training Accuracy v/s Validation Accuracy on Movie Dataset for Inception-V3

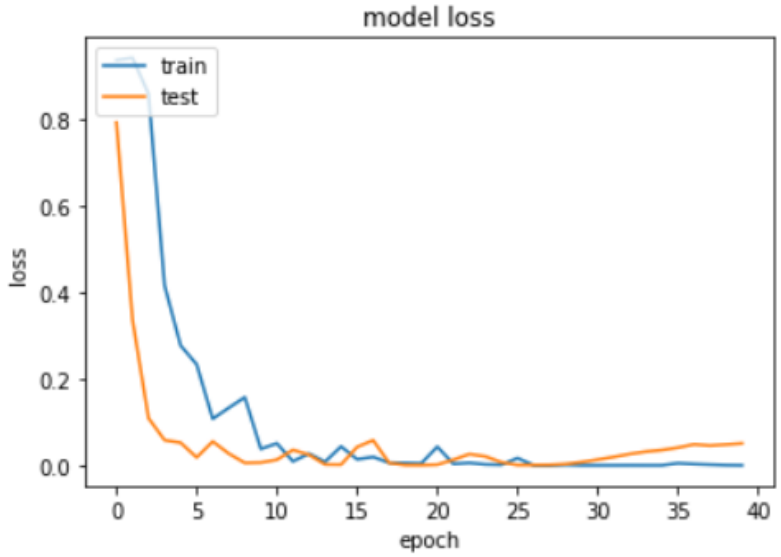

Figure 4.24 A Plot of Training Loss v/s Validation Loss on Movie Dataset for Inception-V3

It can be analyzed from data presented in the above images and tables that out of the 3 pre-trained models used in our proposed architecture, the best results are obtained in Inception-Resnet-V2 based architecture. And hence, our proposed novel Deep ConvNet architecture is based on Inception-Resnet-V2. It can be noted that even though training accuracy reaches a 100% mark, training loss is yet a non-zero value. The reason behind this behavior is that accuracy and loss often appear to be inversely proportional but there is no concrete mathematical relationship between these two metrics[64]. Accuracy can be seen as a count of correct predictions whereas loss can be seen as a difference between true value and predicted probability.

4.4 COMPARITIVE ANALYSIS

To evaluate the proposed novel Inception-Resnet-V2 based Deep ConvNet for the task for Violence Detection, a comparative analysis is presented in table 4.4 and table 4.5 which showcases that the proposed model either outperforms or works at par with the existing technologies and methodologies proposed so far by other researchers. The Real-Life Violence Dataset being quite new to the domain isn't used extensively yet, the best testing accuracy achieved by [63] was 88.2% which is higher than achieved by our proposed model.

Table 4.4 Violence Detection Accuracy (%) Comparison on Hockey Fight Dataset

Method	Accuracy
ViF [34]	81.6%
OViF [36]	84.2%
ViF + OViF [36]	86.03%
HOMO	89.3%
Our Proposed Approach	93.33%

Table 4.5 Violence Detection Accuracy (%) Comparison on Movie Dataset

Method/Classifier		Feature Extracted	Accuracy
[52]		Spatiotemporal	100%
[54]		Spatiotemporal	99.9%
ViF [4]	SVM	Motion	96.7%
	Adaboost		92.8%
	Random Forest		88.9%
Our Proposed Approach		Motion	100%

CHAPTER 5

FUTURE WORK AND CONCLUSION

5.1 CONCLUSION

In this thesis, a novel deep architecture focused on single-stream RGB-DMI is introduced that capitalizes on motion features obtained from violent or non-violent videos. DMIs are simple yet extremely powerful representations of videos summarizing videos into a single image. DMIs can encrypt the summary of the video in a very compact environment allowing for excellent performance by using motion features. Any existing or future CNN architectures can also utilize the DMI as input to achieve better results. Furthermore, incorporating DMIs in the motion stream used in multi-stream state-of-the-arts can enable in learning meaningful results as well. Applying DMIs with very recent very deep CNN – pre-trained InceptionResnetV2, Resnet50, and Inception-V3 with fine-tuning enabled in achieving more speed and better accuracy as compared to existing violence detection methods based on motion feature extraction. Experiments on publically available violence detection dataset benchmarks are conducted to validate the performance and experiments demonstrate the effectiveness of DMIs in achieving impressive performance despite their simplicity. With the rising population and increasing need for surveillance has posed growing demands of systems that are capable to detect violent acts automatically.

Violence detection is already an interesting research field and CNN's have made an exceptional breakthrough in the field of detecting violence. Every upgrade or breakthrough can and will help in creating more and more systems that are robust and utmost accurate in identifying violent content in video and in helping to create highly efficient automated intelligent video analysis systems for detecting violence. The behavioral analysis of human crowds will be the focus of attention due to the possible potential applications, for eg - multi-camera crowd counting, Real-time Processing, and Generalization, Multi-Sensor Information Fusion. This problem can involve researchers from diverse backgrounds.

5.2 FUTURE WORK

In the future, the plan is to design a system that uses the proposed novel deep ConvNet model based on extracting motion features from DMI along with a separate stream of RGB still images as a stack extracted from RGB video which extracts spatial features followed by LSTM for extracting temporal features and combining both the streams to generate predictions which will be much more accurate. Designing a front-end where videos can be uploaded can also be incorporated so that detecting suspicious and violent acts can be done in real-time. If such a feat is achieved, the prototype can be integrated with a surveillance device like a camera to detect violent suspicious activity or criminal activity at highly crowded places like markets, shopping malls, and railway stations, etc. The moment this system detects any such activity, nearby security personnel could also be alerted by activating an alarm as done by[54]. The current model was trained using only the hockey fight, movie, and real-life violence dataset but for training purposed an amalgamation of datasets can be done, for example- Violent-Flows datasets, Caviar, Behave datasets and UFC101 datasets can be clubbed together to provide better training examples. Although different training examples can also be provided to detect different kinds of activities in real-time like training a system to detect bullying by installing such a system in school or college.

In particular, more kinds of violent acts can be detected with their characteristics, so the NN must learn independently given that the dataset is good enough. More sections like fire, explosion, gunshot, etc. can be added for categorizing violent activity. A multi-stream architecture to make violence detection more robust enhancing the performance of the system which can predict in real-time and also alert the nearby law enforcement agencies can be a fruitful solution to also identify a perpetrator is an object and person identification is also integrated with the system. In all, the current times and advancements in computer vision, and image processing related domain did open the gates for creating better software that can achieve the best result and help in simplifying the human tasks.

REFERENCE

[1] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6479–6488, 2018.

[2] H. U. Keval, "Effective design, configuration, and use of digital CCTV," *Crit. Rev.*, no. April, pp. 1–289, 2009.

[3] https://vidimensions.com/, "Vi Dimensions | We Discover. Going Beyond Detection." [Online]. Available: https://vidimensions.com/. [Accessed: 10-Dec-2019].

[4] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T. K. Kim, "Fast fight detection," *PLoS One*, vol. 10, no. 4, pp. 1–19, 2015.

[5] O. Deniz, I. Serrano, G. Bueno, and T. K. Kim, "Fast violence detection in video," *VISAPP 2014 - Proc. 9th Int. Conf. Comput. Vis. Theory Appl.*, vol. 2, pp. 478–485, 2014.

[6] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Rob. Auton. Syst.*, vol. 77, pp. 25–38, 2016.

[7] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "Unified framework for human activity recognition: An approach using spatial edge distribution and ℜ-transform," *AEU - Int. J. Electron. Commun.*, vol. 70, no. 3, pp. 341–353, 2016.

[8] D. K. Vishwakarma and K. Singh, "Human activity recognition based on spatial distribution of gradients at sublevels of average energy silhouette images," *IEEE Trans. Cogn. Dev. Syst.*, vol. 9, no. 4, pp. 316–327, 2017.

[9] L. Tian, H. Wang, Y. Zhou, and C. Peng, "Video big data in smart city: Background construction and optimization for surveillance video processing," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 1371–1382, 2018.

[10] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent Interaction Detection in Video Based on Deep Learning," *J. Phys. Conf. Ser.*, vol. 844, no. 1, 2017.

[11] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, no. August 2018, pp. 21–45, 2019.

[12] M. Ramzan *et al.*, "A Review on State-of-the-Art Violence Detection Techniques," *IEEE Access*, vol. 7, pp. 107560–107575, 2019.

[13] P. D. Garje, M. S. Nagmode, and K. C. Davakhar, "Optical Flow Based Violence Detection in Video Surveillance," *2018 Int. Conf. Adv. Commun. Comput. Technol. ICACCT 2018*, pp. 208–212, 2018.

[14] Z. Guo, F. Wu, H. Chen, J. Yuan, and C. Cai, “Pedestrian violence detection based on optical flow energy characteristics,” *2017 4th Int. Conf. Syst. Informatics, ICSAI 2017*, vol. 2018-Janua, no. Icsai, pp. 1261–1265, 2017.

[15] “New Delhi Streets Turn Into Battleground, Hindus vs. Muslims - The New York Times.” [Online]. Available: https://www.nytimes.com/2020/02/25/world/asia/new-delhi-hindu-muslim-violence.html. [Accessed: 21-Jun-2020].

[16] “Fiery Clashes Erupt Between Police and Protesters Over George Floyd Death - The New York Times.” [Online]. Available: https://www.nytimes.com/2020/05/30/us/minneapolis-floyd-protests.html. [Accessed: 21-Jun-2020].

[17] B. Jiang, F. Xu, W. Tu, and C. Yang, “Channel-wise Attention in 3D Convolutional Networks for Violence Detection,” *Proc. - 2019 Int. Conf. Intell. Comput. Its Emerg. Appl. ICEA 2019*, pp. 59–64, 2019.

[18] E. Ditsanthia, L. Pipanmaekaporn, and S. Kamonsantiroj, “Video Representation Learning for CCTV-Based Violence Detection,” *TIMES-iCON 2018 - 3rd Technol. Innov. Manag. Eng. Sci. Int. Conf.*, pp. 1–5, 2019.

[19] M. Baba, V. Gui, C. Cernazanu, and D. Pescaru, “A sensor network approach for violence detection in smart cities using deep learning,” *Sensors (Switzerland)*, vol. 19, no. 7, pp. 1–17, 2019.

[20] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, “Bidirectional convolutional LSTM for the detection of violence in videos,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11130 LNCS, pp. 280–295, 2019.

[21] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] J. Schmidhuber, “Deep Learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.

[23] J. Gu *et al.*, “Recent advances in convolutional neural networks,” *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

[24] J. Mahmoodi and A. Salajeghe, “A classification method based on optical flow for violence detection,” *Expert Syst. Appl.*, vol. 127, pp. 121–127, 2019.

[25] 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,” *Mach. Vis. Appl.*, vol. 28, no. 3–4, pp. 361–371, 2017.

[26] A. A. Einstein, “DETECTION OF REAL-WORLD FIGHTS IN SURVEILLANCE VIDEOS Mauricio Perez , Alex C . Kot School of Electrical and Electronic Engineering University of Campinas Institute of Computing,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 2662–2666, 2019.

[27] A. Singh, D. Patil, and S. N. Omkar, “Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 1710–1718, 2018.

[28] A. M. R. Abdali and R. F. Al-Tuma, “Robust Real-Time Violence Detection in Video Using CNN and LSTM,” *SCCS 2019 - 2019 2nd Sci. Conf. Comput. Sci.*, pp. 104–108, 2019.

[29] X. Yang, X. Yang, M. Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, “Step: Spatio-temporal progressive learning for video action detection,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 264–272, 2019.

[30] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional Two-Stream Network Fusion for Video Action Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, no. i, pp. 1933–1941, Apr. 2016.

[31] T. Z. Ehsan and M. Nahvi, “Violence detection in indoor surveillance cameras using motion trajectory and differential histogram of optical flow,” *2018 8th Int. Conf. Comput. Knowl. Eng. ICCKE 2018*, no. Iccke, pp. 153–158, 2018.

[32] L. Lazaridis, A. Dimou, and P. Daras, “Abnormal behavior detection in crowded scenes using density heatmaps and optical flow,” *Eur. Signal Process. Conf.*, vol. 2018-Septe, pp. 2060–2064, 2018.

[33] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, “Action recognition by dense trajectories,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3169–3176, 2011.

[34] T. Hassner, Y. Itcher, and O. Kliper-Gross, “Violent flows: Real-time detection of violent crowd behavior,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 1–6, 2012.

[35] P. Vashistha, C. Bhatnagar, and M. A. Khan, “An architecture to identify violence in video surveillance system using ViF and LBP,” *Proc. 4th IEEE Int. Conf. Recent Adv. Inf. Technol. RAIT 2018*, pp. 1–6, 2018.

[36] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, “Violence detection using Oriented VIolent Flows,” *Image Vis. Comput.*, vol. 48–49, pp. 37–41, 2016.

[37] A. A. Mishra and G. Srinivasa, “Automated detection of fighting styles using localized action features,” *Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018*, no. Icisc, pp. 1385–1389, 2018.

[38] R. Ghosh, “Deep Learning for Videos: A 2018 Guide to Action Recognition,” 2018. [Online]. Available: http://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review. [Accessed: 14-Dec-2019].

[39] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, “Violence detection in video by using 3D convolutional neural networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8888, pp. 551–558, 2014.

[40] I. Laptev, “On space-time interest points,” *Int. J. Comput. Vis.*, vol. 64, no. 2–3, pp. 107–123, 2005.

[41] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, “Action Recognition with Dynamic Image Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2799–2813, 2018.

[42] C. Dhiman, M. Ieee, D. K. Vishwakarma, and S. Member, “View-invariant Deep Architecture for Human Action Recognition using late fusion.”

[43] C. Dhiman and D. K. Vishwakarma, “View-Invariant Deep Architecture for Human Action Recognition Using Two-Stream Motion and Shape Temporal Dynamics,” *IEEE Trans. Image Process.*, vol. 29, no. DI, pp. 3835–3844, 2020.

[44] J. Schmidhuber, “Deep Learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.

[45] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.

[46] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” *J. Physiol.*, vol. 195, no. 1, pp. 215–243, 1968.

[47] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[48] Y. LeCun *et al.*, “Handwritten digit recognition with a back-propagation network,” 1990.

[49] S. R. Dinesh Jackson *et al.*, “Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM,” *Comput. Networks*, vol. 151, pp. 191–200, Mar. 2019.

[50] M. Marsden, K. McGuinness, S. Little, and N. E. O’Connor, “ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification,” *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2017*, 2017.

[51] G. Mu, H. Cao, and Q. Jin, “Violent scene detection using convolutional neural networks and deep audio features,” in *Communications in Computer and Information Science*, 2016, vol. 663, pp. 451–461.

[52] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2017*, 2017.

[53] Z. Meng, J. Yuan, and Z. Li, “Trajectory-pooled deep convolutional networks for violence detection in videos,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10528 LNCS, pp. 437–447.

[54] 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 (Switzerland)*, vol. 19, no. 11, pp. 1–15, 2019.

[55] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4787–4797, 2018.

[56] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1725–1732, 2014.

[57] Q. Dai *et al.*, "Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning," *CEUR Workshop Proc.*, vol. 1436, pp. 5–7, 2015.

[58] G. Sakthivinayagam, R. Easawarakumar, A. Arunachalam, and M. Pandi, "Violence Detection System using Convolution Neural Network," *SSRG Int. J. Electron. Commun. Eng.*, vol. 6, pp. 6–9, 2019.

[59] B. Peixoto, B. Lavi, J. P. Pereira Martin, S. Avila, Z. Dias, and A. Rocha, "Toward Subjective Violence Detection in Videos," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 8276–8280, 2019.

[60] S. Mukherjee, R. Saini, P. Kumar, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Fight Detection in Hockey Videos using Deep Network," *J. Multimed. Inf. Syst.*, vol. 4, no. 4, pp. 225–232, 2017.

[61] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic Image Networks for Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 3034–3042, 2016.

[62] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6855 LNCS, no. PART 2, pp. 332–339, 2011.

[63] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," *Proc. - 2019 IEEE 9th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2019*, pp. 80–85, 2019.

[64] "Accuracy and Loss - AI Wiki." [Online]. Available: https://docs.paperspace.com/machine-learning/wiki/accuracy-and-loss. [Accessed: 03-Jul-2020].

LIST OF PUBLICATION BY CANDIDATE

[1] Aayush Jain and Dinesh Kumar Vishwakarma, "State-of-the-arts Violence Detection Using ConvNets" in *IEEE International Conference on Communication and Signal Processing*, 2020 [Presented].

[2] Aayush Jain and Dinesh Kumar Vishwakarma, "Deep NeuralNet For Violence Detection Using Motion Features From Dynamic Images" in *IEEE International Conference on Smart Systems and Inventive Technology*, 2020 [Accepted].