VIOLENCE ACTIVITY RECOGNITION USING COMPUTER VISION

BY

Shuvo Podder
ID: 171-15-888

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Refath Ara Hossain
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Fahad Faisal
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

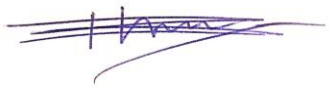
DHAKA, BANGLADESH

JUNE 2021

APPROVAL

This Project titled "**VIOLENCE ACTIVITY RECOGNITION USING COMPUTER VISION**", submitted by Shuvo Podder, ID: 171-15-888 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *01-06-2021*.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Jueal Mia — **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Ms. Refath Ara Hossain, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms. Refath Ara Hossain
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Shuvo Podder
ID: 171-15-888
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

I really grateful and wish my profound my indebtedness to **Ms. Refath Ara Hossain**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Computer Vision*" to carry out this project. Her endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Professor Dr. Touhid Bhuiyan** and Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

I proposed an intelligent system algorithm to address real-time violence activity using computer vision. Sometimes in our absent different violence activity occurs in our daily life. As a part of a smart surveillance system detecting real-time violent activity plays a key role. A video is several frames of the pixel so analyzing and classify them is a challenging research topic in the field of computer vision. Deep learning nevertheless CNN is the key part of computer vision. In previous research action recognition mostly focus on real-life activities but not enough for predicting violence. Considering all possible situation to recognize real-life violence more accurately in this research I follow Convolutional Long Short-Term Memory (CONVLSTM). The model finds spatial features from video and analysis the correlation. Datasets collected from various source and comparatively I get an adequate accuracy result. The research project finished with several experiments using different deep video analyzing algorithms. I compared and differentiated different deep learning model and finalize the best one which about 90% accuracy result. Finally, real-time video footage set to classify with my trained model. The model returns the relevant output whether the scenario is violent or not at the same time the result sent through the cloud to my developed mobile application for further action.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Addressing real-time violence action from the video is an important part of computer vision. Robotics, video surveillance systems are key field of action recognition. It is interesting and useful to recognition action because getting a visual relationship and makes the computer see without human interference [4,17,28,31,32]. As the world going towards automation this could be a key part of the surveillance automation system. Action recognition from a video is not the same as 2D image recognition. Action recognition is a process of finding pattern or correlation from several frames considering spatial feature [2,3]. A single period of the video contains an average of 30 frames of 2D images. In video sometimes violence occurring or not, finding correlation and detecting action from several frames of video is challenging. In previous detection system mostly based on the different object and activity detection like walking, running, cooking, gesture [1,4,6,7,13,14,29,33] but not exactly based on violence or crime detection. In this research, I will focus on crimes that exactly happening in our daily life. In this research, I will go through several deep learning models to recognize violent actions accurately. Because in real life there is the various situation like people with a gun, knife, something fighting, hijacking, snacking. Analyzing those videos, I find some violent activity could be recognized via different CNN algorithm [6,15,20,21,26] on images and give high accuracy and some violent activity need to determine the frame motions feature with period time. In a violent scene for example people with some violent element (gun, knife) could detect by 2D CNN easily where there are some chances to doing a violent activity. And secondly, sometimes people doing violence by fighting where no gun or knife available. In a solution, the model prediction needs to determine correlative motion with time. In this research project, I am going to use kind of deep learning algorithm which determines spatial feature with frames correlation from video data.

1.2 Motivation

In this era of technology, we are using several smart devices in our daily life. But we are a little bit away from monitoring real-time crime and its solution. Violence activity still occurring in front of surveillance cameras. I noticed several hijacking videos from social media collected from surveillance camera footage. I also find several violent activities occurring under the CCTV zone but have not any real-time solution to prevent the crime. That thing inspired me to research in the field of violent action recognition. Studying previously published research works in this area I find there were limitation and research gap in the field of violence action recognition. I decided to develop something which gives real-time effect and make a solution of violent crimes that occurred under the surveillance security zone.

1.3 Rationale of the Study

Computer vision is one of the most advanced frontiers and potentially revolutionary technologies. Bless of this computer vision we get the self-driving car [22], robot. Now the computer gets the ability to see but the most advanced machine computer still stuck in make sense of vision to the blind man. Computer vision indirectly able to helps the blind man [25]. There is a significant difference in human vs computer vision although the main process is common. Activity recognition is a field of computer vision which applied in different sectors in the modern world. Lots of research works have done previously based on action recognition. That research mostly based on daily life general event recognition [1,4,6,7,11,13,14,28,29,33]. There is not much research done to focus on violent action recognition. In a country like Bangladesh most violent activity occurring in a densely or crowded public place. So, things should be more specific in this area, where available datasets not focused on this type of common scenario happening in Bangladesh. So rarely research is done in Bangladesh with violent action. Studying this various scenario, it encourages me to go with this research area. Recognizing violence from a video is a little bit challenging where a single video is several pixels of frames data.

1.4 Research Questions

I am going to determine real-time violence activity from video data and make a solution to prevent crime. As analyzing video is a challenging and complex task in computer science I must go through several topic and methodology to complete this research project. During my research study, I will be trying to study and focus on the following questionaries.

1. What is computer vision?
2. How does computer vision work in action recognition?
3. Deep learning for computer vision?
4. Different violence action recognition algorithm?
5. How to recognize violent action with high accuracy?

1.5 Expected Output

The main target output of this work is to find a novel way and model to recognize violent action more accurately and implement them in a real-life surveillance system. I hope this research work will be a key part of the surveillance security system and also make a good impact on our daily social life. Most research works are just mostly limited to theory. I am expecting good research work with a successful implementation of this research in real life. I believe this work would have the ability to contribute to government security agencies to prevent and detect crime.

Finalize model will predict violence from real-time footage and the system able to alert the nearest authorized person.

1.6 Project Management and Finance

The research project based on real-time violence activity recognition, so I have to consider a huge dataset to make an efficient model. Training a deep learning model with huge video datasets as usual pc, Google Colab will take lots of times to train. Considering those things, I choose a virtual cloud platform to train the model. And collect dataset from different internet source as the research work done by fully self-funding. This project has relation with map and real-time data I have to use cloud functionality to complete the system. I choose to google for this system cloud platform which is easy to use and affordable compared to other available platforms in the market.

1.7 Report Layout

This research project report contains 7 chapters in total.

In chapter 1 I discussed the preliminary introduction of our research work with motivation, the rationale of the study, research question and expected output.

In chapter 2 I will discuss the background study. The basic introduction of computer vision and its background process discussed by comparing human biological vision. This chapter also contains the neuron and its working procedure, basic of CNN, problem discussion, related works.

In chapter 3 the discussion based on the methodology of this research work. In the beginning, I describe the data collection and process procedure. Describe the main system and main mechanism used for the model with system requirements.

In chapter 4 I provide the output and key finding of our experiments in several mathematical terms. This chapter will also contain the comparison table for performance result of the different method I used during our research. I will also discuss the key finding in this chapter.

In chapter 5 I will describe the social impact of our research work and how the research will sustainable.

In chapter 6 the final chapter of the research work where I will summarize and conclude the research work by providing the future scope in this area of research.

CHAPTER 2

Background Studies

2.1 Preliminaries/Terminologies

Computer vision is one of the most advanced frontiers and potentially revolutionary technologies. We get the self-driving car [21,22], automated robots. Now computer gets the ability to see but most advanced machine computer still stuck in make sense of vision to the blind man. There is a significant difference in human vs computer vision although the main basic process is common [23,24].

Human Vision

The human body is the most complex part of the world where human vision is the key part. A human vision works in three-phase eye, brain and result. Lights bounce off the image and enter the eye. The cornea inside the eye directs it for the next processing. Brain receptors access the view and the visual cortex finalize the scenario. I learn various scenario from childhood. The brain stores them and later use them to detect them. In deep neurons works for processing the scenario. Human learns through the real-life experiences and examples. Following Figure 2.1 describe the scenario of human vision.

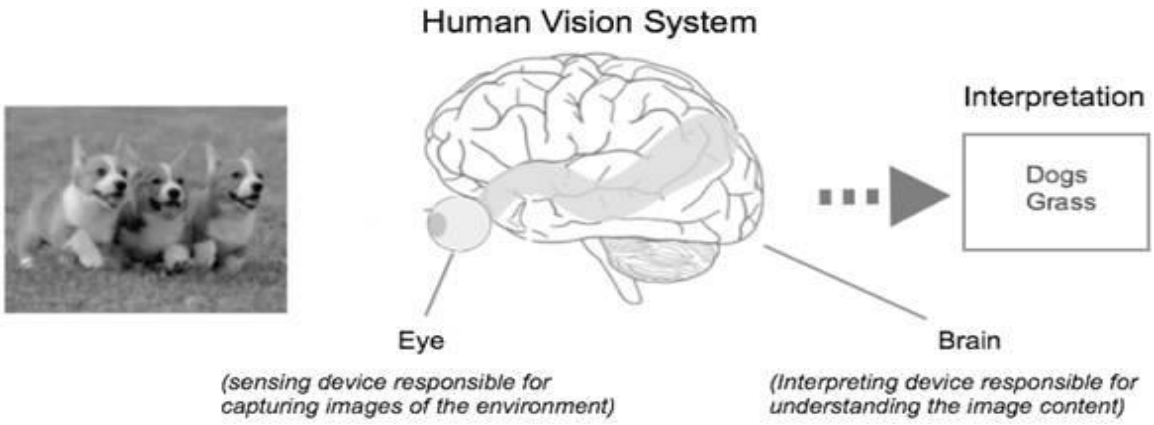


Figure 2.1: Human vision system

Computer Vision

Computer vision is the subfield of Artificial Intelligent. The main aim of computer vision is to gives the computer the ability to see. Computer Vision follows the biological process of human vision to see. Considering the biological visual process of human-computer have camera or image sensing device which relevant to the eyes, machine intelligent to learn and predict which follow the inner process of brain neurons. Computers represent and use a mathematical model of a visual image the generate model, algorithm and different perception which give visual ability. Following Figure 2.2 gives a view of computer vision basic procedure.

Figure 2.2: Scenario of basic computer vision process.

Everything in the universe could represent by math. Computer images are just numbers. Two-dimensional arrays of numbers known as pixels. An image made by a matrix of pixels. Each pixel has its own unique value 0~1 or 0~255. In a grayscale image each pixel has one value but in color image matrix of three-channel Red, Green and Blue. A computer makes and represents each color image pixel to 0 and 1, array representation of three-channel for RGB value where each channel represents by a two-dimensional array.

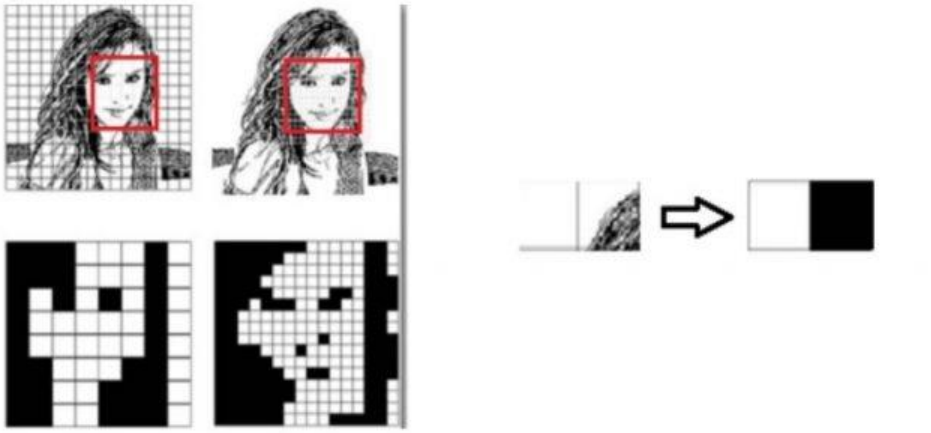


Figure 2.3: Binary representation of an image

Now the computer has got the images but needs to recognize an image. Get back to the human biological vision system in the second stage brain works as a processing system. Humans learn from experiences and real-life example to recognize images. In computer vision, I train a computer to learn which make the computer ability to recognize. I use algorithm and model to teach a computer about recognizing images. For this task, there are several ML models proposed. The most advanced model deep learning use in this process, CNN is one of them for computer vision [20].

Neurons/ Deep Learning

Deep Learning [21] is the subfield of AI where learning algorithm made considering a biological system of human brain neurons. The brain consists of billions of highly connected neurons. Neurons used to process and make sense of what we see. Neurons are the key players of the brain. There are three different kinds of neurons where sensory neurons responsible for seeing. The signal from other neurons arrives on the dendrites, get processed into the cell body to then move along the axon for other neurons. Biological Neurons represent in the following figure 2.4

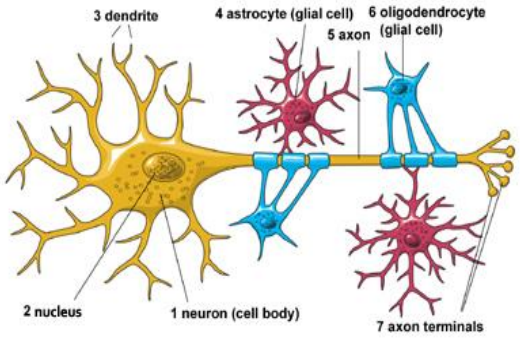


Figure 2.4: Biological Neurons

A deep learning algorithm neural network structured similar to the organization of biological neurons in the brain (Figure 2.4). Founding father of deep learning Geoffrey Hinton took this approach because till now the human brain is the most powerful and advanced computational engine. Neurons used in computer for deep learning also called an artificial neural network. Inspired by the biology of the human brain artificial neurons have three layers called input, hidden and output layer. The input layer is taken signal from others. The hidden layer performs some calculation. The output layer sends signals through the synapse to other neurons.

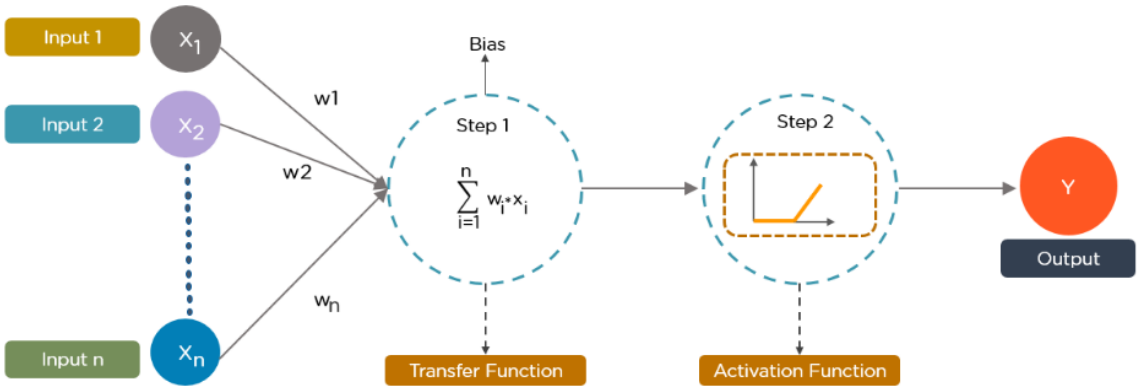


Figure 2.5: Artificial Neurons

The simplest neural networks is perceptron. Synapse is associate with weight which is important in the overall neural network. When neurons receive input neurons Xi

process the sum with its corresponding weights Wi and passed through an activation function. Considering figure 2.5 the neurons take input.

Mathematically,

$$y = \emptyset\left(\sum_{k=0}^{n} w_i x_i\right)$$

Activation functions Ø used to compute the weighted sum of input and biases, which decided whether a neuron can be activated or not. It is used to determine the activation of a node. Four kinds of activation layers work in deep learning. Each of them has its functionality.

CNN:

CNN is a kind of feed-forward neural network DL algorithm which mostly used in Computer vision to detect and classify objects [20][21]. A CNN model works in a three-part input layer. hidden layer and output layer. As our last discussion images could be represented in an array of pixel values. Input layer fed by the formation of images. Several layers of networks used here. Secondly, the hidden layer works for feature extraction contains the convolution layer, ReLU layer, pooling layer. The final fully connected layer provides the final output of the prediction. Convolution layers have several filters to slide over the image and produce a newly convoluted feature matrix. ReLU layer taking the output of the convolution layer and convert it to rectified feature map (negative = 0, positive = return the value). R(z) = max(0,z)

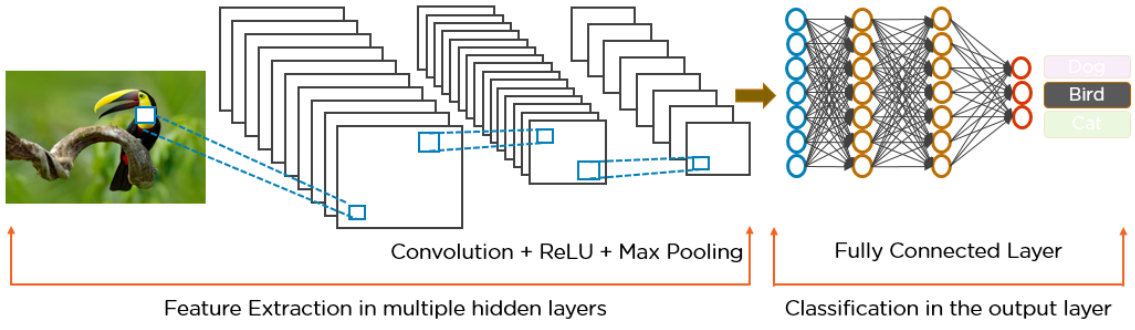


Figure 2.6: Basic architecture of CNN

The pooling layer reduces the dimensionality of the feature map and helps to identify different parts of images like edges, corners. Flattening means converting the input to a single linear output.

2.2 Related Works

"Two-Stream Convolutional Networks for Action Recognition in Videos" research done by the researchers (Karen Simonyan and Andrew Zisserman) of Visual Geometry Group, University of Oxford[1]. Video can describe in spatial and temporal components. The main contribution of this papers was to use two-stream convolutional networks first spatial feature extraction which learns from the still images, second one temporal feature extraction which works on the multiple-image frames optical flow. In this research, they generalize the best performing hand-crafted features within a data-driven learning framework using UCF-101 and HMDB-51 datasets [11].

A group of researchers from Istanbul Technical University Istanbul, Turkey recently worked on "Vision-based Fight Detection from Surveillance Cameras" [3]. They follow LSTM based approach for action recognition with enhancing the traditional CNN+LSTM model. Their Nobel recognition model split into three-part first one subsection to determine the length of frames second one feature extraction which contains VGG-16 and fight-CNN and thirdly classification parts where Bi-LSTM + attention layer used. Bi-LSTM learns from past and current information, attention layer defines and manages the importance from the past and current event. Newly dataset used for this research which contains 1500 video clips from different source like Hockey Dataset, Peliculas Dataset, Surveillance Camera Dataset.

Activitynet[4] is the first database for human activity recognition organized under a rich semantic taxonomy. Activitynet made for human activity understanding with large scale video benchmark. The research covers a large scale of daily living human activities. A large set of video sample used for each class. Activitynet used 137 videos per class in a total of 203 activity classes. The datasets are about 849 video hours in length. Datasets split

into 50:25:25 for train: test: validation. The detection algorithm works in three phases untrimmed video classification, trimmed activity classification and activity detection.

2.3 Comparative Analysis and Summary

After several years of research, the activity recognition model algorithm improved. The algorithm which considers the spatial feature with time series provides comparatively the highest accuracy. I shortly make a summary of previous research on activity recognition.

TABLE 2.1: PREVIOUS RESEARCH ON ACTIVITY RECOGNITION

Title	Datasets	Methodology	Result
Two-Stream Convolutional Networks for Action Recognition in Videos [1]	UCF-101 and HMDB-51	Two-stream convolutional networks	88%
Vision-based Fight Detection from Surveillance Cameras [3]	1500 video clip collected from various source	Fight-CNN + Bi-LSTM + attention	72%
ActivityNet: A large-scale video benchmark for human activity understanding [4]	27801 videos	MS, FS, DF	50%

2.4 Scope of the Problem

As I am going to analyze violent action from real-time video from the public place there have some chances of facing complexity in this research. There are several datasets for recognizing action but there is not much for violence. As our expected area in Bangladesh so there are several unusual occurrences found in a public place which is more complex to determine. As I am working with video datasets, I need a huge time to deal with experiments for an adequate result.

2.5 Challenges

Studying previous research, I have known that action recognition is a complex task in computer vision. Most previous research was focused on daily household life human activity. As this research focused on violent action and I have not much data I need to collect them and preprocess them for further analysis. Analysis and extract feature from still images is easier than recognize action from video. Where a video contains several frames, and it is also related to spatial feature over time my deep learning model must consider those things correctly. The deep learning model took time to train and when the topic based on video things gets more complex and time-consuming. Another side some action seems like violence in action but not and some action seems like as usual but there could be something violent.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Titled "VIOLENCE ACTION RECOGNITION USING COMPUTER VISION" is a kind of research-based project in the field of computer vision. In this research project, I am using a custom deep learning model for detecting violence from real-time video. As the deep learning model is complicated to train, I need a good configuration pc. But for a faster training process I use Google Colab with Cloud Platform. Here I am providing all the instruments list used for this research.

Hardware & Tools:

- Intel core i5 6th generation
- 12 GB RAM
- 1TB HDD
- Google Colab
- Google Cloud Platform

Software & Libraries:

- Windows 10 Education
- OFFICE 365
- Python 3.7
- TensorFlow 2.3
- OpenCV
- Pandas
- NumPy
- Keras

3.2 Data Collection Procedure/Dataset Utilized

I collected in total 1000 video dataset from several sources containing surveillance camera footage, hijack mobile footage, fighting scene etc. I choose a small size video clip. All the data collected from various source like Kaggle, YouTube, Computer Vision Lab. After collecting data dataset preprocess into 40fps and sized to 90×90.

Data preparation:

- Load data
- Check for null and missing frames.
- Normalization
- Reshape
- Label encoding
- Split training and validation set.

3.3 Statistical Analysis

Final datasets for this research project contain 1000 datasets in total with two different categories 500 for violence and 500 for non-violence. I split the dataset 80:20 train, test ratio which contains 400 for train and 100 for testing for those two categories.

TABLE 3.1: DATASETS DESCRIPTION & ANALYSIS

Datasets	Train	Test	Total
Violence	400	100	500
Non-Violence	400	100	500

3.4 Proposed Methodology/Applied Mechanism

The main purpose of this research was to detect violent activity from real-time video. I consider the following method as a prediction model. Following Figure 3.1 contains the overview diagram of the system for violence action recognition.

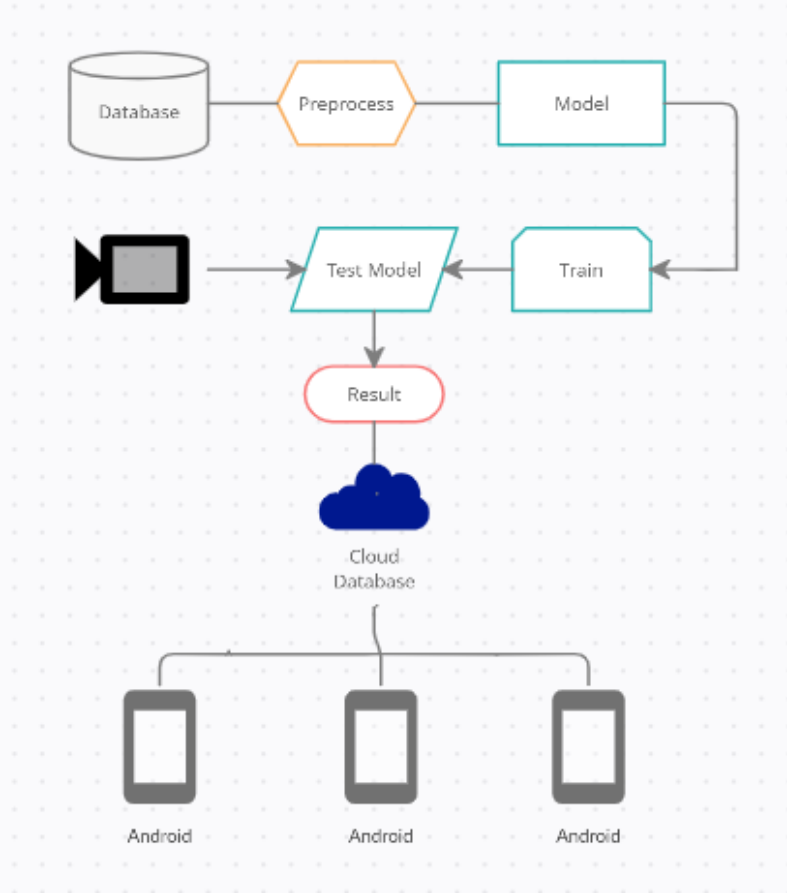


Figure 3.1: Overview diagram of the proposed system

I proposed two Deep Learning algorithm to analyze the video for recognizing action violence. The model detects violence activity from real-time video footage when the model detects violence and results sent to mobile application through cloud.

CONVLSTM: CONVLSTM works on spatial and temporal dimensions of data. I am working on violence action detection from videos. So, videos are a large num of images frame. To recognize violence accurately I must analyze several frames from a video and need to get the feature of how the motion happening with time. Focusing on the spatial and temporal dimension for analyzing a video CONVLSTM [1,2,3,5,12,13,30,33,35] encode the convolution spatial feature with time sequence. Following the process whole datasets train and encode with a localized spatial-temporal feature. Where normal LSTM failed to localize the spatial-temporal changes between frames because it extracts feature from a fully connected layer. That is why CONVLSTM suitable for our method. I proposed Convolutional LSTM, which is a structure designed specifically for spatiotemporal sequences. CONVLSTM uses convolution in the internal calculation.

Generally, LSTM works [13] following four steps forget, store, update output. First (f_t) forgetting irrelevant history from the previous stage and passing it through a sigmoid gate. Second (i_t) storing most relevant new information, what part of the new information and what part of the old information relevant are store into the cell state. Thirdly updating their internal cell state (C_t). final state (H_t) is further controlled by the output gate (o_t). CONVLSTM used the Convolution operation between those states, where LSTM find frames relationship with time, CONVLSTM extracts spatial-temporal feature like convolution and address their changing relationship with time. Convolution calculation in CONVLSTM denotes by "*" Reference Figure :3.2.

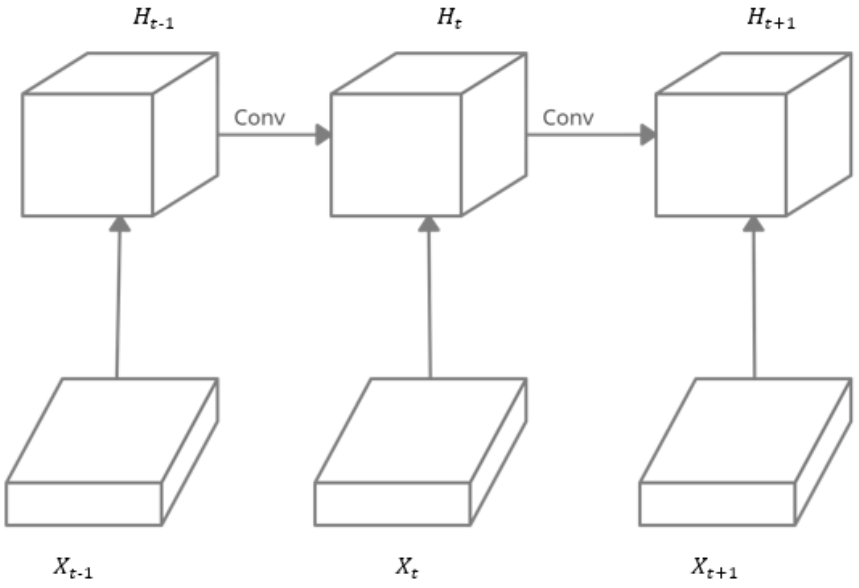


Figure 3.2: Architecture of CONVLSTM

The equation for the CONVLSTM are bellowed where "*" denotes the convolution operator, and "o" denotes the Hadamard product,

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o)$$
$$H_t = o_t \circ tanh(C_t)$$

And parameter size count by,

$$\text{ParamConvLSTM} = \text{K} \times \text{K} \times (\text{C}_{\text{in}} + \text{C}_{\text{out}}) \times \text{C}_{\text{out}} \times 4$$

$\text{X}_1, \ldots, \text{X}_\text{t}$ refers to the current input, cell states denote by $\text{C}_1, \ldots, \text{C}_t$, outputs as $\text{H}_1, \ldots . H_t$, and gates are i_t, f_t, o_t .

3.5 Implementation Requirements

For a computer vision system algorithm development, we need at least minimum system requirements. The desired model based on video datasets and the system will work with real-time video, I need to set up pc with 8GB ram,40GB HDD and 4GB GPU with a Core i5 processor higher configuration system works much better. To work with video or photo need to be installed OpenCV. Python with Jupyter Notebook used to design the system algorithm. As my desire classification model took much time in as usual pc I go for Cloud Platform and used Google Cloud. Datasets processed in Google Drive. Model training purpose a highly configured virtual pc need. Mobile application finished with android studio and firebase database used as a storage drive. Google cloud functionality used for communication data media. For implements and experiment minimum software and hardware requirements list given below.

- Set up virtual pc with Google Cloud.
- Set Google Colab for model train
- Setting up TensorFlow
- Setting up OpenCV
- Processor core i5
- GPU 4 GB ,8GB ram
- Memory size 40 GB
- Jupyter Notebook
- Google drive
- Firebase Database

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

The datasets contain two files with violence and no violence video data. I load and preprocess the datasets in height, width values to 90 & 90 and the length of the sequence for a single video is 40fps. The datasets encoded to four-dimensional NumPy array, where the dimensions of the array. The datasets split into 80:20 ratio for train, test.

CONVLSTM:

The model trained from scratch using the TensorFlow library with Keras interface. CONVLSTM has taken 5D tensor data (samples, frames, row, column, channel). I used 'channels_last' as a data format this is why the channel was taken at last. For each sample data frame size was 40 with 90×90 pixel per color frame which mean RBG three-channel. I define CONVLSTM2D with 72 filters in the first layer. Return sequences set to False so the CONVLSTM return a 4D tensor (category, filter, row, column). Rectified linear unit (ReLU) non-linear activation is applied after each of the convolutional and fully connected layers. The final output was a Fully Connected layer with two output categories using the SoftMax activation function.

I define and set my classification model to the following steps:

- Model taken 40 frame per data with spatial size (90,90), input size (40,90,90,3)
- Convlstm2d layer: filter size 72, kernel size (3, 3), return sequences set as 'False', data format as 'channels_last',
- Model dropout .2 for preventing overfitting.
- Flatten the layer.
- Fully Connected Dense layer with value 254 and Rectified linear unit (ReLU) non-linear activation is applied after each of the convolutional and fully connected layers.
- Again dropout 0.3

- Fully Connected layer with two output categories using SoftMax activation function.

Total params: 141,817,724
Trainable params: 141,817,724

For training process loss function "categorical_crossentropy" used with 'Adam' optimizer, learning rate .0001. My model trainable data contains a large video dataset I used small batch size which was 5 with 25 epochs for refinement.

4.2 Experimental Results & Analysis

During my experiment, I have gone through algorithms and model to get the highest accurate model for the datasets. To test the accuracy of my train model I used Classification Metrics, Charts and Classification report. For violence action analysis the model most possible result is true or negative. In more description follow table no 2. True Positive (TP) when the predicted value matches the actual value means the actual result of a data is true and the model gives the predicted value true. True Negative (TN) denoted when the actual value is negative, and the trained model gives the predicted value negative. There are two types of error cases one is False Positive (FP) in case of falsely predicted like when the actual value is negative, but the model predicted value as positive, a second error occurs when the actual value is positive but the prediction gives negative output denoted as False Negative (FN).

Evaluation metrics

TABLE 4.2: EVALUATION METRICS SUMMARIZATION

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

I define the accuracy as mean average precision considering three concepts of measurement precision, recall and f1 score. The mathematical process of the evaluation given below,

Precision provided the total amounts of correctly predicted cases turned out to be positive. We could measure the model reliability with this precision value,

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall gives the correctness of a model, total positive cases the model able to predict correctly,

$$\text{Recall} = \frac{TP}{TP+FN}$$

To get a combined idea about those two values I used F1-score. It is finding the harmonic mean from precision and recall value.,

$$\text{F1} = \frac{2*precision*recall}{precision+recall}$$

This is how I will calculate the accuracy of a model,

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

Following Figure 4.1 for the final accuracy report. For the CONVLSTM model train I get about 90% accuracy,

	precision	recall	f1-score	support
0	0.88	0.91	0.89	96
1	0.91	0.88	0.90	104
accuracy			0.90	200
macro avg	0.89	0.90	0.89	200
weighted avg	0.90	0.90	0.90	200

Figure 4.1: Accuracy Report

Model accuracy during the training period I use a graphical chart below in following figure 4.2.

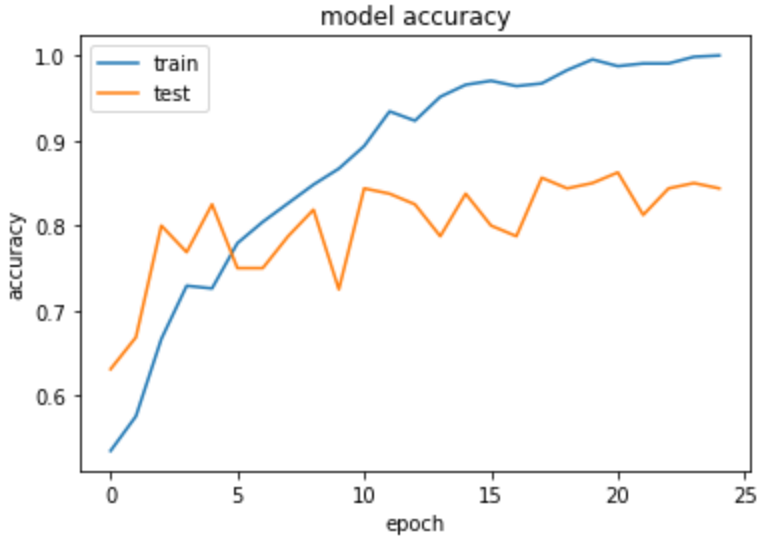


Figure 4.2: Train Model Accuracy

Following figure 4.3 contains the graphical review of total model loss during training.

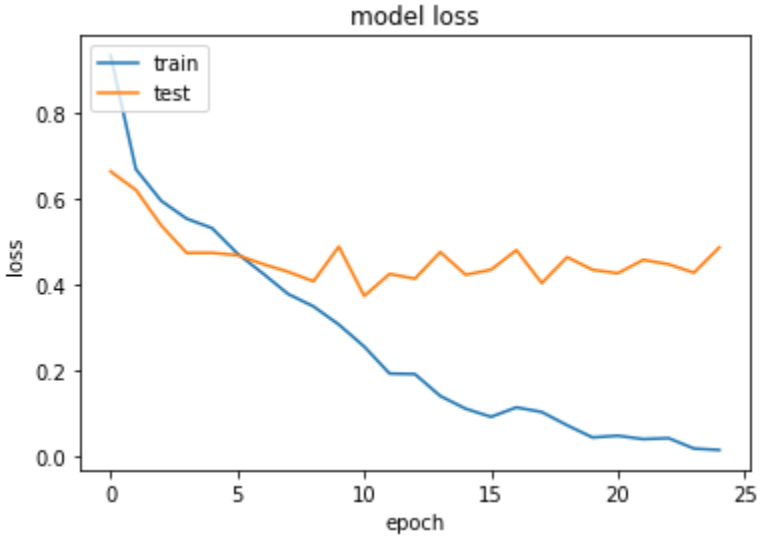


Figure 4.3: Train Model Loss

Confusion metrics:

Summarize the result of a machine learning model by showing the correlation between the label and the model's classification with N×N table. Considering the following confusion matrix in following figure 4.4 that describe my train model that the model correctly classified 87 as violence (TP) and 9 are incorrectly classified. Similarly, 92 (TN) no violence classified correctly where 12 incorrectly classified (FP).

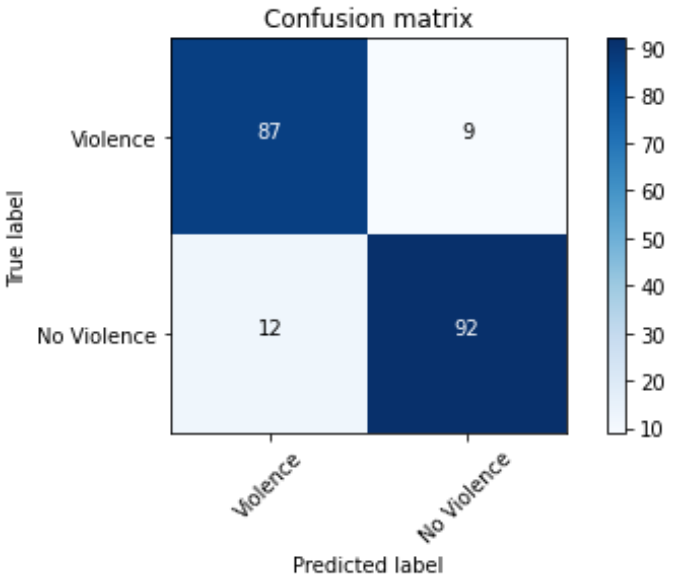


Figure 4.4: Confusion Matrix

Statistical difference of my experiment with different training model:

TABLE 4.2: STATISTICAL DIFFERENCE OF MY EXPERIMENT

Model	Accuracy
3D CNN	81%
CONVLSTM	90%
3D CNN+ CONVLSTM	81%

4.3 Discussion

During my experiments, I tried several methods, including the algorithm with different parameter. During that experiment, I focused on the accuracy rate. Deep Learning is a lengthy process to train. At first, I split the datasets into a small part and train them for the experimenting purpose then define the top accurate model and finally train with the final datasets. I collected a few datasets for this thesis. But the output I get more than expected. As a result of my work, I get an effective model for recognizing violent action. The model works accurately which is better than my expectation.

To get the result and test my trained model I split 5 frames group per prediction. The model analyzes those group of images with the spatial and temporal feature that the scenario is violent or not violent.

Violence

No Violence

Figure 4.5: Violence and No Violence Scenario

Experimenting with several model and algorithm on my collected datasets I concluded the following findings:

- 3D CNN and CONVLSTM give more accurate result in recognition. But CONVLSTM provides the best result.
- 3D CNN is faster than CONVLSTM.
- To improve the training time, combine a layer of 3D CNN with CONVLSTM.

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

We live in the age of modern technology. All the modern society in the world are advanced in the uses of technology. A modern society gives more modern service and has a better lifestyle. Security is a major part of society. When a society or place is safe from violence and crime, we say the safest place to live in. Everybody wants to be in a safe and modern society where they get the best services and environment. Day by day modern society gets updated with digital services. At the same time, the medium of crime changing. So, things getting complex to control, monitor and protect. This research would add great value in modern society with a surveillance system to improve our lifestyle, control, monitor and digitally protect our place.

5.2 Impact on Environment

This research would be a key part of a digital country. Nowadays all modern and smart city trying to increase the uses of Artificial intelligence and IoT. A fast-developing country like Bangladesh also improving the lifestyle in a city with those intelligent technologies. For a most populated country maintain the crime rate is challenging. AI and Robotics give the best result where human power is not sufficient. Successful implementation and uses of this research will change city life and make it easier to detect crime and improve the crime rate. An extraordinary value will be added with a surveillance system in modern city life. Where in a dangerous situation citizen in Bangladesh have to call 999 and for other countries they use their specific national security number, then the connected officer makes a connection to the authorized and appropriate person for help. This system takes a huge time to support a victim. For example, the victim should have described the situation, exact location, and other relevant information. To solve this, matter this research will have to detect the violent situation and directly notify the nearest authorized person. The system automatically detects violence like hijacking, gender violence, snacking, and all other violence and sent notification to the nearest security authority track by GPS with a picture,

place and other relevant information. Overall, this will be an intelligent automation system that will control, monitor and secure city life with less human power. Using this system, the national security service also improves where they faced several fake call report. Nowadays we must check the CCTV footage all day so monitoring all the city with CCTV is a little bit harder. This automatic system changes the life of the city's people and help the security forces to prevent crime. Successful implementation of this research makes a city top listed among all digital cities that using artificial intelligence directly to automate the security system with security forces.

5.3 Ethical Aspects

Thinking about public welfare this research idea was proposed. People are helpless in a dangerous situation when they did not find anyone to help. It is not easy to monitor and provide security everywhere. Provide an automated security service, secure people from different dangerous situation. Security is the first key thing of basic needs in citizen life. Automated security service with less human effort in different zone helps a lot of people to improve and secure life.

- This research can add and improve the value of a country security service.
- Help all residence to leads a better lifestyle.
- Secure life and economy.

5.4 Sustainability Plan

As this research related to security. The proposed model will be applied in a real-life event. And try to connect with government security agencies to apply this research for public safety. For private use, implementing software following this research methodology with existing CCTV improve the security system and economically effective. This model will be sustainable because in Bangladesh till now there is no other company or organization that are not giving this service. Not only Bangladesh we can have a look in other modern countries they are still a little bit away from this kind of service. Some country doing research and invest in this area of work but have not any real-life implementation till now. And this deep learning model will be extremely new in our country and the same pipeline could be used in a different situation. Not only in government service the implemented system can add extra value in some private service like in hospital, office, factory. And from another side in this era of technology using ai in real life for detecting violence or crime will be a groundbreaking innovation, I think. People are helpless when they are in danger while using public places. Ordinary CCTV does not work with these situations because there is no real-time data communication system or alert system to let someone know to help them and it is harder to monitor the entire city all time using CCTV. My proposed system makes an automation system without any human interference and helps people during danger by providing an alert to nearest authorized person for further action. My research idea will add extra value to security in the city with the surveillance system. Overall people will get a digital environment and I believe the implementation of this research in real life will reduce the crime rate and improve our citizen lifestyle.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

During the study of this research project, I went through several difficulties and learn several topics in the fields of computer vision and deep learning. In the first stage, I learn how the computer sees and common things with the biological vision of human. During this basic vision study, I study about neuron and neural network as well as deep learning. Then I discussed several deep learning models which are related to action analysis. I have been studied previously published research works and their result of accuracy. Choose a related methodology for my classification model. Dataset from different sources in two categories. As the datasets were not prepared for classification in this stage I prepare and preprocess for classification model train. My previously chosen model trained with the datasets with several experiments. Google Cloud platform was used for training purpose. During the experiments, I also collect information and their behavior. Finally, desire and set a model best fit for my datasets and topic. Using different evaluation process like evaluation metrics, the chart I show the model performance. At the end of finalizing the model, the system was developed for addressing and alert via mobile application. The research works done with good accuracy. A short review of the steps I follow to finish this research work bellowed.

- Literature review
- Data collection
- Preprocess data
- Define and train the model
- Showing the performance of the trained model
- Improve performance
- Setup for real time prediction

6.2 Conclusions

My main aim was to address violent activity from real-time video data. I completed this work with an effective result. My trained model able to determine violence, no violence activity from real-time video data and make an alert to the nearest authorized person for further step. During this research, I learn several human action recognition model and different field of computer vision analysis. I find two major and highly accurate method for my datasets CONVLSTM and 3D CNN. CONVLSTM achieve more accuracy compared to the 3D CNN model. During analysis, that method I also noticed and find that the 3D CNN is the faster method which is easy to train and predict but CONVLSTM gives us more accuracy because LSTM is a kind of RNN that works on frame correlation and helps to predict much higher accuracy. I tried to find the best method to analyze violent action during this research and explored several ways of recognition violent action from real-time video data. Research is a continuous process thing may be more improved in future with newly developed method.

6.3 Implication for Further Study

This research was done by using the CONVLSTM method to recognize action depending on the spatial correlation between videos. I also did some experiments and compared the 3D CNN model with other related modified deep learning algorithm. As research work is a continuous process this is not all, the recognition could also be more accurate with other algorithms and method. After the study, activity recognizes possible by human pose estimation [29] and give effective output. In this research, there were limited datasets collected from various internet sources. Future research could be more efficient with new specific category datasets. I used two categories one for violence another for no violence, in violence category contain surveillance camera violence video, fight, hijack, gun violence etc future work could be updated with newly collected datasets for a more effective and specific result. I also suggest observing and increase the uses of the cloud platform like a model API for better performance.

References:

[1] K. Simonyan, A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in Neural Information Processing Systems, 2014, pp. 568–576

[2] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[3] Ş. Aktı, G. A. Tataroğlu and H. K. Ekenel, “Vision-based Fight Detection from Surveillance Cameras,” 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2019, pp. 1-6, doi: 10.1109/IPTA.2019.8936070.

[4] F. C. Heilbron, V. Escorcia, B. Ghanem and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 961-970, doi: 10.1109/CVPR.2015.7298698.

[5] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, “Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4580-4584, doi: 10.1109/ICASSP.2015.7178838.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

[7] Chakraborty B., Bagdanov A.D., Gonzàlez J. , “Towards Real-Time Human Action Recognition” In: Araujo H., Mendonça A.M., Pinho A.J., Torres M.I. (eds) Pattern Recognition and Image Analysis. IbPRIA 2009. Lecture Notes in Computer Science, vol 5524. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02172-5_55

[8] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, “Beyond short snippets: Deep networks for video classification,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4694-4702, doi: 10.1109/CVPR.2015.7299101.

[9] Md. Zabirul Islam, Md. Milon Islam, Amanullah Asraf, “A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images,” Informatics in Medicine Unlocked, Volume 20, 2020, 100412, ISSN 2352-9148

[10] E. Ditsanthia, L. Pipanmaekaporn and S. Kamonsantiroj, “Video Representation Learning for CCTV-Based Violence Detection,” 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), 2018, pp. 1-5, doi: 10.1109/TIMES-iCON.2018.8621751.

[11] Ke R, Li W, Cui Z, Wang Y. “Two-Stream Multi-Channel Convolutional Neural Network for Multi-Lane Traffic Speed Prediction Considering Traffic Volume Impact”, Transportation Research Record. 2020;2674(4):459-470. doi:10.1177/0361198120911052

[12] D. Wang, Y. Yang and S. Ning, “DeepSTCL: A Deep Spatio-temporal ConvLSTM for Travel Demand Prediction,” 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489530.

[13] Zufan Zhang, Zongming Lv, Chenquan Gan, Qingyi Zhu, “Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions”, Neurocomputing, Volume 410,2020,Pages 304-316,ISSN 0925-2312

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, “HMDB: A large video database for human motion recognition,” 2011 International Conference on Computer Vision, 2011, pp. 2556-2563, doi: 10.1109/ICCV.2011.6126543.

[15] Nasaruddin, N., Muchtar, K., Afdhal, A. et al, “Deep anomaly detection through visual attention in surveillance videos,” J Big Data 7, 87 (2020), https://doi.org/10.1186/s40537-020-00365-y

[16] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” IEEE Trans. Pattern Anal. Mach. Intell., 30(3):555–560, 2008

[17] J. Johnson et al., “Image retrieval using scene graphs,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3668-3678, doi: 10.1109/CVPR.2015.7298990.

[18] S. Gupta,”"Facial emotion recognition in real-time and static images,” 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, pp. 553-560, doi: 10.1109/ICISC.2018.8398861.

[19] A. Jain and D. K. Vishwakarma, “State-of-the-arts Violence Detection using ConvNets,” 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 0813-0817, doi: 10.1109/ICCSP48568.2020.9182433.

[20] S. Albawi, T. A. Mohammed and S. Al-Zawi, “Understanding of a convolutional neural network,” 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[21] Abhishek Gupta, Alagan Anpalagan, Ling Guan, Ahmed Shaharyar Khwaja, “Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues,”Array,Volume 10, 2021,100057, ISSN 2590-0056, https://doi.org/10.1016/j.array.2021.100057.

[22] B. Barua, C. Gomes, S. Baghe and J. Sisodia, “A Self-Driving Car Implementation using Computer Vision for Detection and Navigation,” 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 271-274, doi: 10.1109/ICCS45141.2019.9065627.

[23] B. Zhang, “Computer vision vs. human vision,” 9th IEEE International Conference on Cognitive Informatics (ICCI'10), 2010, pp. 3-3, doi: 10.1109/COGINF.2010.5599750.

[24] Chong, E., Clark-Whitney, E., Southerland, A. et al. “Detection of eye contact with deep neural networks is as accurate as human experts.,” Nat Commun 11, 6386 (2020). https://doi.org/10.1038/s41467-020-19712-x

[25] Rodrigo A. Brant Fernandes, Bruno Diniz, Ramiro Ribeiro, Mark Humayun,“Artificial vision through neuronal stimulation,”Neuroscience Letters, Volume 519, Issue 2,2012,Pages 122-128,ISSN 0304-3940, https://doi.org/10.1016/j.neulet.2012.01.063.

[26] Fernandez-Carrobles M.M., Deniz O., Maroto F. , “ Gun and Knife Detection Based on Faster R-CNN for Video Surveillance,” In: Morales A., Fierrez J., Sánchez J., Ribeiro B. (eds) Pattern Recognition and Image Analysis. IbPRIA 2019. Lecture Notes in Computer Science, vol 11868. Springer, Cham. https://doi.org/10.1007/978-3-030-31321-0_38

[27] T. N. Nguyen and J. Meunier, “Anomaly Detection in Video Sequence With Appearance-Motion Correspondence,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1273-1283, doi: 10.1109/ICCV.2019.00136.

[28] Amit Singh, Albert Haque, Alexandre Alahi, Serena Yeung, Michelle Guo, Jill R Glassman, William Beninati, Terry Platchek, Li Fei-Fei, Arnold Milstein, “Automatic detection of hand hygiene using computer vision technology”, Journal of the American Medical Informatics Association, Volume 27, Issue 8, August 2020, Pages 1316–1320, https://doi.org/10.1093/jamia/ocaa115

[29] Haque A., Peng B., Luo Z., Alahi A., Yeung S., Fei-Fei L. ,"Towards Viewpoint Invariant 3D Human Pose Estimation," In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_10

[30] Rahman, S.A., Adjeroh, D.A. "Deep Learning using Convolutional LSTM estimates Biological Age from Physical Activity," Sci Rep 9, 11425 (2019). https://doi.org/10.1038/s41598-019-46850-0

[31] Krishna, R., Zhu, Y., Groth, O. et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," Int J Comput Vis 123, 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7

[32] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[33] G. Zhu et al., "Redundancy and Attention in Convolutional LSTM for Gesture Recognition," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 4, pp. 1323-1335, April 2020, doi: 10.1109/TNNLS.2019.2919764.

[34] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[35] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. "Convolutional LSTM Network: a machine learning approach for precipitation nowcasting," In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 802–810.

APPENDICES

Appendix A: Android Studio

In this research project, I used the android studio to develop the mobile application used for the system. The mobile app connected to the firebase database to get real-time data from the system through google cloud. When the system changes the value of the current situation in firebase through cloud my internal algorithm in the mobile app determines the action that should be taken. User also gets notification via mobile device when any dangerous scenario occurs.

Appendix B: Google Cloud and Firebase Database

Analyzing a video database is a complex and time-consuming task. In a solution, I used Google Cloud-based virtual environment to train my model. Secondly, the full system has an alert system via mobile application. Google Cloud used to pass data through the cloud and address the specific location map API that needs to be integrated. The mobile app used the firebase database as storage and real-time communication media with help of Google Cloud Services. Firebase database also manages and control the user information.

Appendix C: Real Time Violence Detection

I set real-time video data with my model, the model address violence and alert the nearest authorized person via the mobile app. The app alertly notifies and shows the DateTime with other information related to the incident.

Figure 7.1: Addressing violence activity and alert via mobile application.

Appendix D: Other Experimented Model

During this research I experiment with several models from them CONVLSTM, 3DCNN, CONVLSTM+3DCNN gives higher performance. I already declare and finalize CONVLSTM for this violence action recognition. But still, the other two models have a good result. Following figure 7.2 and figure 7.3 describe the performance of the other two trained model during my research work.

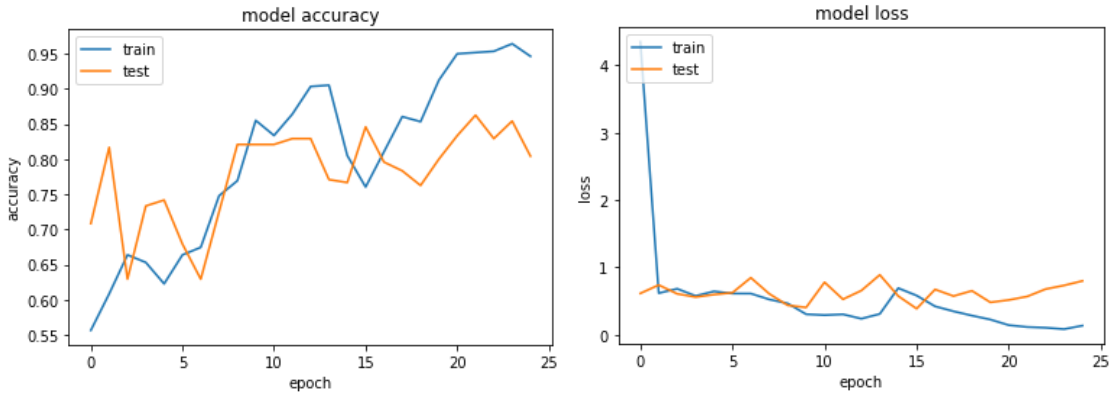


Figure 7.2: Performance of 3DCNN+CONVLSTM model

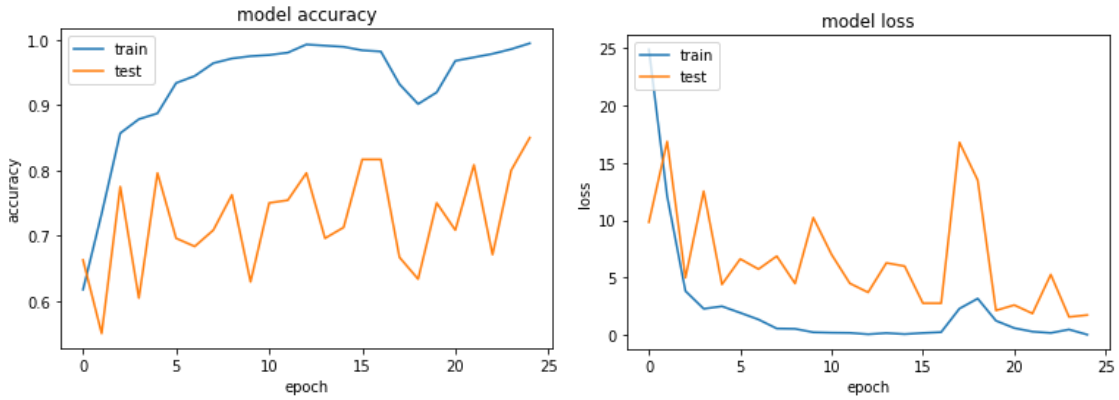


Figure 7.3: Performance of 3DCNN model

Appendix E: Open-Source Repository

Research is a continuous process. I would like to make some part of this research open source for further improvement. Interested researcher encourages to fork this repository and contribute.

Link: <<https://github.com/shuvopodder/Violence-Activity-Recognition-Using-Computer-Vision >>

Violence Activity Recognition using Computer Vision

ORIGINALITY REPORT

9%	7%	3%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	arxiv.org Internet Source	1%
4	www.analyticsvidhya.com Internet Source	1%
5	Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, Juan Carlos Niebles. "ActivityNet: A large-scale video benchmark for human activity understanding", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 Publication	1%
6	Ming-Huwi Horng, Cheng-Wei Yang, Yung-Nien Sun, Tai-Hua Yang. "DeepNerve: A New Convolutional Neural Network for the Localization and Segmentation of the Median	<1%

	Nerve in Ultrasound Image Sequences", Ultrasound in Medicine & Biology, 2020 Publication	
7	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2021 Publication	<1 %
8	Submitted to Rochester Institute of Technology Student Paper	<1 %
9	lup.lub.lu.se Internet Source	<1 %
10	李庆辉 Li Qinghui, 李艾华 Li Aihua, 王涛 Wang Tao, 崔智高 Cui Zhigao. "Double-Stream Convolutional Networks with Sequential Optical Flow Image for Action Recognition", Acta Optica Sinica, 2018 Publication	<1 %
11	"Proceedings of International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2021 Publication	<1 %
12	cs229.stanford.edu Internet Source	<1 %
13	doi.org Internet Source	<1 %

14	Submitted to University of Surrey Student Paper	<1 %
15	eprints.lancs.ac.uk Internet Source	<1 %
16	Xiang Chen, Yuanchang Liu, Kamalasudhan Achuthan, Xinyu Zhang. "A ship movement classification based on Automatic Identification System (AIS) data using Convolutional Neural Network", Ocean Engineering, 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		