Deep Surveillance

Project Report Submitted in Partial Fulfilment of the Requirements for the Degree of

Bachelor of Engineering
in
Computer Science and Engineering

Submitted by

MOHIT SONI(Roll No.18UCSE4021)

VIKRAMADITYA SINGH KACHHWAHA(Roll No.18UCSE4028)

VINEET KUMAR MEENA(Roll No.18UCSE4029)

SEM : VIII

Under the Supervision of

Mr. ABHISEK GOUR
Assistant Professor

Department of Computer Science and Engineering
M.B.M. Engineering College
Faculty of Engineering & Architecture
Jai Narain Vyas University, Jodhpur

JULY 2021

Department of Computer Science & Engineering

M.B.M. Engineering College, Jai Narain Vyas University
Ratanada, Jodhpur, Rajasthan, India –342011

CERTIFICATE

This is to certify that the work contained in this report entitled **"**Deep Surveillance " is submitted by Mr. Mohit Soni (Roll. No: 18UCSE4021),Mr. Vikramaditya Singh Kachhwaha (Roll. No: 18UCSE4028) and Mr. Vineet Kumar Meena (Roll. No: 18UCSE4029) to the Department of Computer Science & Engineering, M.B.M. Engineering College, Jodhpur, for the partial fullfilment of the requirements for the degree of **Bachelor of Engineering** in **Computer Science and Engineering**.

They have carried out their work under my supervision. This work has not been submitted else-where for the award of any other degree or diploma.

The project work in my opinion, has reached the standard fulfilling of the requirements for the degree of Bachelor of Engineering in Computer Science and Engineering in accordance with the regulations of the Institute.

Mr Abhisek Gour
Assistant Professor
(Supervisor)
Dept. of Computer Science & Engg.
M.B.M. Engineering College, Jodhpur

Prof. N.C. Barwar
(Head)
Dept. of Computer Science & Engg.
M.B.M. Engineering College, Jodhpur

DECLARATION

We, ***Mohit Soni, Vikramaditya Singh Kachhwaha*** **and** ***Vineet Kumar Meena*** hereby declare that this project titled **"**Deep Surveillance**"** is a record of original work done by me under the supervision and guidance of **Mr Abhisek Gour .**

We, further certify that this work has not formed the basis for the award of the Degree/Diploma/Associateship/Fellowship or similar recognition to any candidate of any university and no part of this report is reproduced as it is from any other source without appropriate reference and permission.

Mohit Soni (Roll No.18UCSE4021)
Vikramaditya Singh Kachhwaha (Roll No.18UCSE4028)
Vineet Kumar Meena (Roll No.18UCSE4029)
VIII Semester, CSE

ACKNOWLEDGEMENT

This is to thank all those who supported and helped us throughout the commencement of this seminar report. We would like to thank specially Mr. Abhisek Gour Sir for his continuous guidance. We would also like to thanks faculty of the department of CSE & IT and my friends and juniors for their continuous help and encouragement.

ABSTRACT

In this video surveillance project, we have introduced a spatio temporal autoencoder, which is based on a 3D convolution network. The encoder part extracts the spatial and temporal information, and then the decoder reconstructs the frames. The abnormal events are identified by computing the reconstruction loss using Euclidean distance between original and reconstructed batch.

In this project basically, we choose a video file and then the model will find the abnormal events or activities like fighting, stealing, etc.

Our project have two main components, one is a deep neural network model and another is its frontend.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 **MOTIVATION**

In the last decade, there have been advancements in learning algorithms for deep surveillance. These advancements have shown an essential trend in deep surveillance and promise a drastic efficiency gain. In the places which are prone to many crimes and attacks, security is assured by implanting CCTV cameras. Cameras and recording devices are relatively expensive and require human intervention to monitor the camera footage. Along with it, there is a large amount of data generated which requires vast storage. All this, demands more manpower and increased cost. In an era where we have smart gadgets, it is time to get our surveillance cameras smart as well. To do so we empower our surveillance cameras with deep learning algorithms. Our goal is to focus on automatic identification of unauthorized entries in an area by alerting the authorities. This implementation has a wide range of applications which includes personnel identification, bank security, company security, shop security etc. Also, it helps with instant crime detection which helps the community to identify the culprit easily.

1.2 **SCOPE OF PROJECT**

The scope of this project is to minimize the theft happening in future and maximize the protection of a data in highly confidential region. This will be more essential in industry areas.

- This will minimize the theft happening
- The administrator will be notified at the time of theft happening.
- The data will be protected with high security in the confidential region.
- The administrator will be notified regarding the abnormal activity in his/her place by a means of short message service or mail system using pop3 configuration.
- The process is fast and highly securable.

This will summarize the past methodology and help to serve better for future purpose.

Fig 1.2.1 Abnormal Event

Fig 1.2.2 Abnormal Event

1.3 REQUIREMENTS

1.3.1 PYTHON IDE

IDE stands for Integrated Development Environment. It's a coding tool which allows US to write, test, and debug our code in an easier way, as they typically offer code completion or code insight by highlighting, resource management, debugging tools.In our project specifically we are using jupyter notebook as our IDE.

1.3.2 LIBRARIES

1.3.2.1 OpenCV

OpenCV-Python is a library of Python bindings designed to solve computer vision problems. It can process images and videos to identify objects, faces, or even the handwriting of a human.

1.3.2.2 KERAS

Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result as fast as possible is key to doing good research.Keras is simple, flexible and powerful.

1.3.2.3 IMUTILS

Imutils is a series of convenience functions to make basic image processing functions such as translation, rotation, resizing, skeletonization, and displaying Matplotlib images easier with OpenCV and ***both*** Python 2.7 and Python 3.

1.3.2.4 FLASK

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries.[2] It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

1.3.3 WSGI

The Web Server Gateway Interface (WSGI) is a standard interface between web server software and web applications written in Python.

1.3.4 HARDWARE ARCHITECTURE

1.3.4.1 RAM

For this project we specifically need RAM around 32 GB.

1.3.4.2 PROCESSOR

Processor with configuration of i3 or higher.

Chapter 2

RELATED WORK

2.1 Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture

Malik Ali Gul , Muhammad Haroon Yousaf , Shah Nawaz , Zaka Ur Rehman and HyungWon Kim

In this paper, You Look only Once (YOLO) network is utilized as a backbone CNN model. For training the CNN model, a large dataset of patient videos by labeling each frame with a set of patient actions and the patient's positions are contructed.
They retrained the back-bone CNN model with 23,040 labeled images of patient's actions for 32 epochs. Across each frame, the proposed model allocated a unique confidence score and action label for video sequences by finding the recurrent action label.

The present study shows that the accuracy of abnormal action recognition is 96.8%.Their proposed approach differentiated abnormal actions with improved F1-Score of 89.2% which is higher than state-of-the-art techniques.

The results indicate that the proposed framework can be beneficial to hospitals and elder care homes for patient monitoring.

The foremost contribution of this research is the development of a high-speed recognizing algorithm and making a custom dataset of abnormal human activities for patient monitoring. For fast and easy computation, frames are processed independently at the desired rate by ignoring the temporal redundancy and we chose YOLO as the back-bone CNN model implementation of the YOLO framework for abnormal human activity recognition to meet the requirement of real-time computation. YOLO uses a single CNN simultaneously for classification as well as for localizationof the object instead of using different neural networks methods to first classify and then localize it.
The proposed system when implemented using the computer machines specified above can process the patient images at a speed of 40–90 frames per second. As a result, the improved performance has been achieved with minimum latency.
This work was demonstrated on video datasets with individual subjects present in the field view of the camera. This assumption limits the scope of real-world application of the work, where multiple human subjects may appear in a single scene.

2.2 Abnormal activity recognition based on deep learning in crowd

Yan Fu,Tao Liu,Ou Ye

In this paper, a new hybrid deep learning structure was proposed to fuse the extracted features, which integrates convolutional neural network (CNN) and long short-term memory network (LSTM). Firstly, the video was preprocessed and extracted visual features by CNN. Next, LSTM was used to learn the temporal features of visual features and added attention mechanism to select important features. Finally, the video feature vector obtained layer by layer to judge abnormal activity. An experiment is used to test the ability of the model on the standard dataset UMN to recognize abnormal activity.

In order to verify the effectiveness of the algorithm in this paper, this paper uses accuracy rate, precision rate and recall rate to evaluate the effectiveness of each algorithm. Where, accuracy represents the proportion of samples correctly classified in all classifications. The precision rate indicates how many of the predicted samples (such as positive samples) are actually samples of a certain type. Recall is how much of a sample of a sample is correctly predicted.

UMN dataset was selected for experiments, and three groups of experiments were conducted on the UMN dataset to verify the proposed method, containing SIFT+BoW, CNN+LSTM and BoG, and the comparison of proposed approach with existing method

The experimental result shows that the proposed method has been tested on the UMN dataset and outperforms the existing used methods, which proves the efficiency of the proposed method. The proposed approach cannot only fully extract the deep features of video frames, but also focus on the behavioral features that have a greater impact on results. Therefore, it has a greater potential compared with common deep learning and traditional manual feature extraction methods. However, due to the large amount of calculation, the real-time performance of this method is difficult to be applied to the multi-channel recognition system with high.

2.3 Sparse Representation and Weighted Clustering Based Abnormal Behavior Detection

Dongliang Jin, Songhao Zhu*, Songsong Wu, Xiaoyuan Jing

First, the hybrid histogram of optical flow feature is extracted; then, the double sparse representation is proposed to tackle the issue of global abnormal activity detection; finally, for the issue of local abnormal activity detection, the foreground of region of interest within the current frame is first detected, and then the method of online weighted clustering is utilized to detect local abnormal activity.
Experiments results conducted on UMN datasets and UCSD datasets validate the advantages of the proposed method.

To evaluate the efficiency of the proposed global abnormal activity detection method, other three methods are are selected as the comparative method, including the method based on Optical Flow in the method based on Social Force Model in and the method based on Energy Model and Threshold (EMT) .
Experimental results in terms of ROC curve, AUC value and EER value fully evaluate the advantages of the proposed method. The time consumption of the proposed method is 2000ms in ped1 dataset and 2100ms in ped2 dataset, which demonstrates the proposed method has lower computational complexity in local abnormal activities detection.

In this paper, a novel method of global abnormal activity detection and local abnormal activity detection is introduced. First, at the feature level, a new descriptor called hybrid histogram of optical flow is presented. Then, the proposed global abnormal activity detection consists of two sparse processes obtaining two probabilities respectively, and abnormal activity is judged by the fuzzy integral method. For the issue of local abnormal activity detection, they first extracted a small patches of region of interest, then the online weighted clustering algorithm is used to determinate whether the local abnormal activity exists or not, and finally we adopt multiple target tracking method to filter the noise. Experimental results demonstrate that the proposed method has higher accuracy and effectiveness in global and local abnormal activity demonstrate that the proposed method has higher accuracy and effectiveness in global and local abnormal activity detection.

2.4 Abnormal activity detection using shear transformed spatio-temporal regions at the surveillance network edge

Michael George , Babita Roslind Jose1 , Jimson Mathew

This paper presents a method of detecting abnormal activity in crowd videos while considering the direction of the dominant crowd motion. One main goal of their approach is to be able to run at the edge of the surveillance network close to the surveillance cameras so as to reduce network congestion and decision latency. To capture motion features while considering the direction of dominant crowd direction we propose a generalised shear transform based spatio-temporal region.

To detect abnormal activity, an autoencoder based method is adopted considering the requirement for running the method at the network edge. During training, the autoencoder learns motion features for each spatio-temporal region from video frames containing normal activity.

The AUC/EER (in %) metrics on UCSD Ped1, UCSD Ped2, Subway Entrance and Subway Exit are 0.8475/22.98, 0.954/10.7, 0.849/23.4, and 0.835/19.9 respectively. The frame processing rates obtained on an edge device were between 3.08 to 23.12 fps. It was also observed that frame-rates could be improved with a trade-off with accuracy by lowering frame size.

The paper motivated the need to process frames at the edge in a large video surveillance network from the viewpoint of latency and bandwidth usage. A generalised shear transform based approach that is suitable for an edge-based abnormal activity detection was proposed.
This approach was tested on an edge device using standard abnormal activity detection.
Datasets

2.5 Detecting abnormal events in traffic video surveillance using superorientation optical flow feature

Joshan Athanesious , Vasuhi Srinivasan, Vaidehi Vijayakumar, Shiny Christobel, Sibi Chakkaravarthy Sethuraman

The paper proposed a novel scheme called super orientation optical flow (SOOF)-based clustering for identifying the abnormal activities. The key idea behind the proposed SOOF features is to efficiently reproduce the motion information of a moving vehicle with respect to superorientation motion descriptor within the sequence of the frame. Here, the authors adopt the mean absolute temporal difference to identify the anomalies by motion block (MB) selection and localisation. SOOF features obtained from MB are used as motion descriptor for both normal and abnormal events.

The proposed scheme detects anomalies with 94.6% accuracy and 98.55% recall for highway followed by 92% accuracy with 96.17% recall for city datasets.

In this paper, motion vectors(orientation) are used to capture the vehicle flow and the MB selection is obtained using the mean absolute temporal difference.
Additionally, finding the optimal cluster centre for a traffic scene by means of cluster labelling technique outperforms the state-of the-art model in terms of accuracy. It is also realised that the nearest-neighbour search is efficient in detecting anomalies with negligible false positives. Experimental results on three datasets i.e. real-time situation, iLIDS AVSS 2007 and UCF-Crime (road accidents) datasets confirm that the proposed system is efficient in identifying anomalies

2.6 Spatio-TemporalEncoder-Decoder Fully Convolutional Network for Video-based Dimensional Emotion Recognition

Zhengyin Du, Suowei Wu, Di Huang, Member, IEEE, Weixin Li, Member, IEEE, and Yunhong Wang, Senior Member, IEEE

In this paper, a novel encoder-decoder framework is used. It adopts a fully convolutional design with the cascaded 2D convolution based spatial encoder and 1D convolution based temporal encoder-decoder for joint spatio-temporal modeling. In particular, to address the key issue of capturing discriminative long-term dynamic dependency, our temporal model, referred to as Temporal Hourglass Convolutional Neural Network (TH-CNN), extracts contextual relationship through integrating both low-level encoded and high-level decoded clues. Temporal Intermediate Supervision (TIS) is then introduced to enhance affective representations generated by TH-CNN under a multi-resolution strategy, which guides TH-CNN to learn macroscopic long-term trend and refined short-term fluctuations progressively.Extensive experiments are conducted on three benchmark databases (RECOLA, SEWA and OMG) and superior results are shown compared to state-of-the-art methods, which indicates the effectiveness of the proposed approach.

The model is compared with several state-of-the-art methods including both handcrafted feature based and CNN feature based ones on Recola. This model is also compared with several recent deep learning based methods on SEWA. This TH-CNN model is more convenient for parallel computing due to the characteristic of convolution operation.

In this research, a novel and effective framework for video-based dimensional emotion recognition model is proposed. Specifically, present an encoder-decoder fully convolutional network integrating a spatio-temporal encoder and a temporal decoder to support spatial and long-range contextual dependency.

2.7 Dynamic Basic Activity Sequence Matching Method in Abnormal Driving Pattern Detection Using Smartphone Sensors

Thi-Hau Nguyen , Dang-Nhac Lu , Duc-Nhan Nguyen , and Ha-Nam Nguyen

In this work,a novel method,name lydynamicbasic activity sequence matching (DAS) is presented, a combination of machine learning methods and flexible threshold based methods for distinguishing normal and abnormal driving patterns. Indeed, DAS relies on the activity detection module (ADM) to analyze each driving pattern as a sequence of basic activities—stopping (S), going straight (G), turning left (L), and turning right (R). Moreover, an efficient framework composing of two phases: in the first phase, the normal and abnormal driving patterns are distinguished by relyingon DAS is proposed here. In the second phase,the detected abnormal patterns are further classified into various specific abnormal driving patterns—weaving, sudden braking, etc.

Prediction accuracy for identifying various specific abnormal driving patterns.

Methods		Weaving	Sudden Braking	Average
Dynamic Time Wrapping (DTW)		97.65%	96.45%	97.05%
Machine Learning	Random Forest (RF)	100%	93%	96.50%
	Convolution Neural Network (CNN)	96.20%	97.70%	96.95%
Proposed Method		98.04%	97.83%	97.94%

Table 2.7.1

The experiments illustrate that for distinguishing normal and abnormal driving patterns problem, the longer window size leads to the higher prediction accuracy, as the duration of each basic activity is often long in the normal driving scenarios. In contrast, for specific abnormal driving pattern detection problem, the shorter window size leads to the higher prediction accuracy, as the duration of each basic activity is quite short in the abnormal driving scenarios.

2.8 Non-Intrusive human activity recognition and abnormal behavior detection on elderly people: a review

Athanasios Lentzas and Dimitris Vrakas

With the world population aging at a fast rate, ambient assisted living systems focused on elderly people gather more attention. Human activity recognition (HAR) is a component connected to those systems, as it allows identification of the actions performed and their utilization on behavioral analysis. This paper aims to provide a review on recent studies focusing on HAR and abnormal behavior detection specifically for seniors. The frameworks proposed in the literature are presented. The results are also discussed and summarized, along with the datasets and metrics used. The absence of a universal evaluation framework makes direct comparison not feasible, thus an analysis is made trying to divide the literature using a taxonomy. Solutions on the challenges identified are proposed, while discussing future work.

This review, presented recent advantages in those fields extensively. In total seventeen approaches were discussed, while performing an analysis on their results and categorizing them based on design choices and purpose. Reviewed material, allowed us to detect open issues and challenges on the area.

The main challenge at the moment when performing activity recognition, is the use of the same system without retraining on different subjects. Sztyler et al. (2017) tried to achieve cross subject recognition using subject grouping with physical characteristics and the results were promising. Using subject grouping, though requires a large and diverse amount of training data, in order to create sets with enough examples to train each classifier. In future, cross subject recognition has to be addressed, as no need for retraining, could make HAR systems accessible to a larger number of households.

2.9 Abnormal Crowd Behavior Detection Using Heuristic Search and Motion Awareness

Imran Usman and Abdulaziz A. Albesher

In this paper, a novel approach to detect anomalous human activity using a hybrid approach of statistical model and Genetic Programming is proposed. The feature-set of local motion patterns is generated by a statistical model from the video data in an unsupervised way. This features set is inserted to an enhanced Genetic Programming based classifier to classify normal and abnormal patterns. The experiments are performed using publicly available benchmark datasets under different real-life scenarios.

The below figure shows the performance comparison of the proposed technique with some recent state of the art anomaly detection techniques using the UCSD dataset using the Area under the Curve (AUC) metric. It can be seen from the figure that the proposed technique outperforms most of the previous techniques.

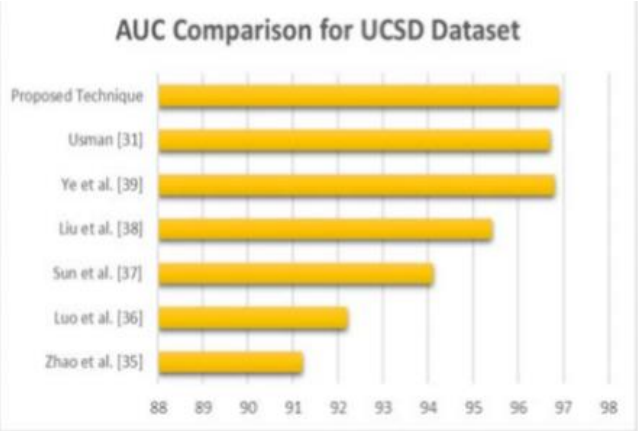

Fig 2.9.1

In this paper, we introduce an automatic crowd video sequence anomaly detection system. For crowd motion, the proposed methodology uses a gradient-based approach with activation

function and a motion aware feature. GP-based training simulation is used to evolve the best classifier in a stepwise enhancement process.

2.10 A Robust Framework for Abnormal Human Action Recognition Using R-Transform and Zernike Moments in Depth Videos

Chhavi Dhiman and Dinesh Kumar Vishwakarma , Senior Member, IEEE

The aim of the algorithm is to detect the abnormal actions that are more prone to elderly people in order to make them more independent and improve their quality of life. The framework is structured to construct a robust feature vector by computing R-transform and Zernike moments on average energy silhouette images (AESIs). The AESIs are generated by the integral sum of the segmented silhouettes obtained from the Microsoft's Kinect sensor v1. The proposed feature descriptor possesses scale-, translation-, and rotation-invariant properties that are also less sensitive to noise and minimizes data redundancy. It enhances the proposed algorithm's robustness and makes the classification process more efficient. The proposed work is validated on a novel abnormal human action (AbHA) dataset and three publically available 3D datasets—UR fall detection dataset, Kinect Activity Recognition dataset, and multiview NUCLA dataset.

The proposed framework exhibits superior results from other state-of-the-art methods in terms of average recognition accuracy (ARA). The experimental results demonstrate 96.5%, 96.64%, 95.9%, and 86.4% ARA on the UR fall detection dataset, the KARD dataset, the AbHA dataset, and the multi-view NUCLA dataset, respectively.

In this paper, a vision based Abnormal Human Action Recognition approach is presented to monitor the daily life actions of elderly people to recognize the abnormal action in order to providethem better services. The proposed framework encrypts an action in terms of AESI image which is rich in the spatio-temporal details and introduce less computational cost.

Chapter 3

PROJECT DETAILS

3.1 TECHNOLOGY USED

3.1.1 DEEP LEARNING

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data.

In our projects we have used CNN and LSTM.

CNN

A convolutional neural network (CNN) is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data.CNNs are powerful image processing, artificial intelligence (AI) that use deep learning to perform both generative and descriptive tasks.

LSTM

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

3.1.2 HTML

HTML(Hyper text Markup Language) is used for creating web pages and web applications.

3.1.3 CSS

CSS stands for Cascading Style Sheets.It describes how HTML elements are to be displayed on screen,paper or any in other media.

3.1.4 JAVASCRIPT

JavaScript is a dynamic computer programming language. It is lightweight and most commonly used as a part of web pages, whose implementations allow client-side script to interact with the user and make dynamic pages. It is an interpreted programming language with object-oriented capabilities.

3.1.5 FLASK

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

3.2 PROJECT MILESTONE

3.2.1 PROJECT FORMULATION

In this video surveillance project, we have introduced a spatio temporal autoencoder, which is based on a 3D convolution network. The encoder part extracts the spatial and temporal information, and then the decoder reconstructs the frames. The abnormal events are identified by computing the reconstruction loss using Euclidean distance between original and reconstructed batch.

In this project basically, we choose a video file and then the model will find the abnormal events or activities like fighting, stealing, etc.

We tested our project on two datasets for accuracy purposes, to find out how much is the accuracy of our model on the data collected by us compared to some other available dataset on internet.

Our project have two main components, one is a deep neural network model and another is its frontend. So, first we will discussing about its model which is made in a sequential manner. Different phases of this model are as follows:

3.2.2 DATA COLLECTION

In this project we had two datasets one which is available on internet i.e. CUHK campus avenue dataset.

The videos are captured in CUHK campus avenue with 15328 training frames.This dataset accompanies paper "Abnormal Event Detection at 150 FPS in Matlab, Cewu Lu, Jianping Shi, JiayaJia, International Conference on Computer Vision, (ICCV), 2013".

The training videos capture normal situations. Testing videos include both normal and abnormal events. Two samples are shown as follows:

AvenueDataset : http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

Fig 3.2.2.1

Fig 3.2.2.2

For this project we have collected a variety of video footages from our college cctv cameras. For this we have taken the data or videos from 3 cameras which are:

- Camera near the computer center.
- Camera near the physics department.
- Camera in mechanical department.

Fig 3.2.2.3 Camera near the computer center

Fig 3.2.2.4 Camera near the physics department

Fig 3.2.2.5 Camera in mechanical department

For testing we needed some abnormal activities too since in campus it is not normal to find out those especially during covid times. So, we shooted some abnormal activities in front of those cameras.

3.2.3 DATA CLEANING

In CUHK avenue dataset we had around 10 minutes of content with 15328 training frames out of which we trained on 1500 frames by skipping every 10frames such that there was no loss of features as there were many repeated frames. And for testing also there were around 10 minutes of content with normal and abnormal activites.

Now in MBM campus dataset we have over more than 15 GB (12-14 hours) of data so first we had to select the data or videos for our training data set which had normal activities like walking, moving cars, etc. For this we selected and trimmed the videos manually. After that we selected some data for testing too, outof which we labeled some frames for test accuracy too.

The videos were first converted into frames and then fed into the model for training. Since each and every frame of the video was not necessary as continuous frames are repetitive therefore every 10^{th} frame of the video is selected as frame.

There were total 14990 frames for training. These frames were converted into array as gray scale images as model will be used mainly for classification between normal and abnormal events.

These images which are stored as array are now standardized because it makes sure that data is internally consistent; that is, each data type has the same content and format. Standardized values are useful for tracking data that isn't easy to compare otherwise.Then we have finally stored our data in a numpy file.

A value is standardized as follows:

y = (x - mean) / standard_deviation

3.2.4 MODEL SELECTION

Here, we used a spatial autoencoder model. Autoencoders (AE) are neural networks that aims to copy their inputs to their outputs. They work by compressing the input into a latent-space representation, and then reconstructing the output from this representation. This kind of network is composed of two parts :

1. **Encoder:** This is the part of the network that compresses the input into a latent-space representation. It can be represented by an encoding function h=f(x).
2. **Decoder:** This part aims to reconstruct the input from the latent space representation. It can be represented by a decoding function r=g(h).

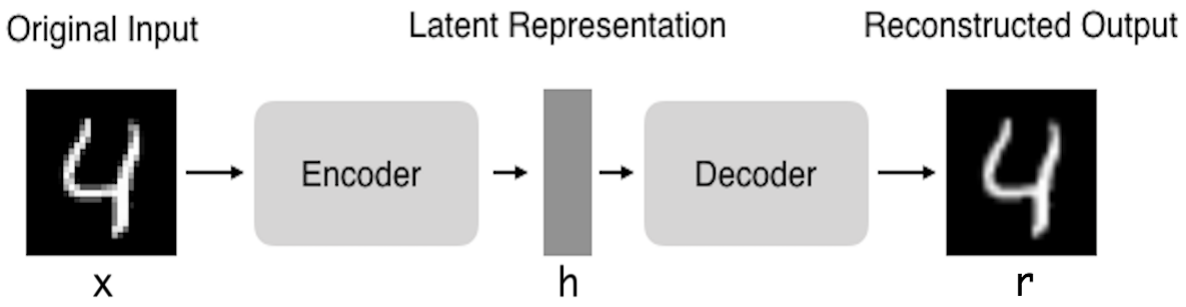

Fig 3.2.4.1 Image through Encoder-Decoder

Autoencoders are trained to preserve as much information as possible when an input is run through the encoder and then the decoder, but are also trained to make the new representation have various nice properties.

Here, we have used **Spatio-Temporal AutoEncoder** (ST AutoEncoder or STAE), which utilizes deep neural networks to learn video representation automatically and extracts features from both spatial and temporal dimensions by performing 3-dimensional convolutions.

3.2.5 MODEL TRAINING

First the model was trained on CUHK avenue dataset then on MBM campus dataset.

In CUHK avenue dataset,

- The model was trained on video content of around 10 minutes.
- There were around 1500 frames in total for training.

The model was trained for 10 epochs with batch size of 10 which took 40-45 minutes.

In MBM campus dataset,

- The model was trained on video content of around 2 hours.
- There were around 14990 frames in total for training.

The model is trained is for 10 epochs with batch size of 10 which took 3-4 hours.

```
model=Sequential()

model.add(Conv3D(filters=128,kernel_size=(11,11,1),strides=(4,4,1),padding='valid',input_shape=(227,227,10,1),activation='relu'))
model.add(Conv3D(filters=64,kernel_size=(5,5,1),strides=(2,2,1),padding='valid',activation='relu'))
model.add(ConvLSTM2D(filters=64,kernel_size=(3,3),strides=1,padding='same',dropout=0.4,recurrent_dropout=0.3,return_sequences=Tru
model.add(ConvLSTM2D(filters=32,kernel_size=(3,3),strides=1,padding='same',dropout=0.3,return_sequences=True))
model.add(ConvLSTM2D(filters=64,kernel_size=(3,3),strides=1,return_sequences=True, padding='same',dropout=0.5))
model.add(Conv3DTranspose(filters=128,kernel_size=(5,5,1),strides=(2,2,1),padding='valid',activation='relu'))
model.add(Conv3DTranspose(filters=1,kernel_size=(11,11,1),strides=(4,4,1),padding='valid',activation='relu'))

model.compile(optimizer='adam',loss='mean_squared_error',metrics=['accuracy'])
```

Fig 3.2.5.1 Spatio Temporal AutoEncoder Model

.

3.2.6 Web User Interface

This is the step were we have converted our python script into a **Python web application** using the **Flask web framework.**

A web application often requires a static file such as a javascript file or a CSS file supporting the display of a web page. Usually, the web server is configured to serve them for us, but during the development, these files are served from static folder in our package and it will be available at /static on the application. In our project there are 3 sub folders in static folder containing the css, js, imgs(containing screenshots showing abnormal and normal activity)

The whole work of creating web application gets divided into two parts, Front end and Back end.

Front End

- We have used HTML for creating webpage and CSS for styling of the content.
- It contains name of our Group “AI Buddies” and provides information related to Deep Surveillance to users.

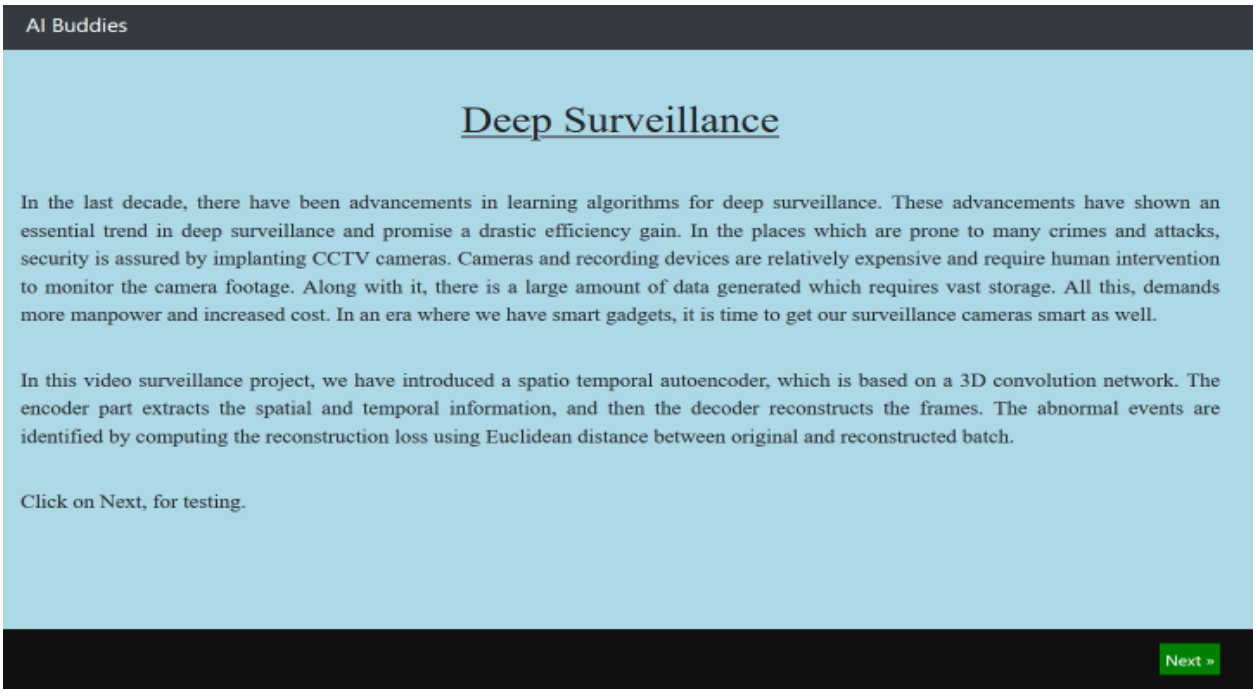

Fig 3.2.6.1

- In order to test a video for any abnormality detection, steps are shown on web page .

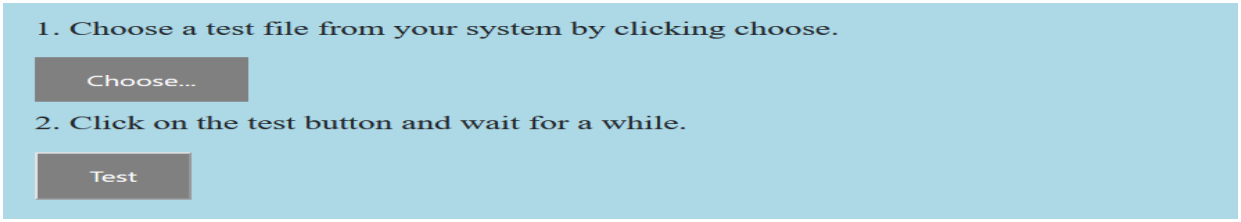

Fig 3.2.6.2

- After choosing the test file and clicking the Test button, after some time video starts playing.

Back End

- Here comes the role of jQuery(a JavaScript framework) and Flask.

jQuery

As soon as the user clicks on the Test button, the uploaded file gets store in our .js file, which then by ajax function call (used for exchanging data with a server, and updating parts of a web page - without reloading the whole page.) delivered to python script running on backend.

```
$(document).ready(function () {

    // Test
    $('#btn-test').click(function () {
        var form_data = new FormData($('#upload-file')[0]);

        // Make prediction by calling api /test
        $.ajax({
            type: 'POST',
            url: '/test',
            data: form_data,
            contentType: false,
            cache: false,
            processData: false,
            async: true,
        });
    });
});
```

Fig 3.6.2.3

Flask

- In the python script, Flask facilitates us to upload the files easily.
- Flask uses **jinja2** template engine. A web template contains HTML syntax interspersed placeholders for variables and expressions (in these case Python expressions) which are replaced values when the template is rendered.
- An object of Flask class is our **WSGI** application.
- The Web Server Gateway Interface(**WSGI**) is a simple calling convention for web servers to forward requests to web applications or frameworks written in the Python programming language.
- Modern web frameworks use the routing technique to help a user remember application URLs. It is useful to access the desired page directly without having to navigate from the home page. The following code has **'/'** URL rule that displays **'index.html'** from the templates folder. Here URL **'/upload'** rule is bound to the **upload()** function.
- The server-side flask script fetches the file from the request object using request.files[] Object. On successfully uploading the file, it is saved to the desired location on the server which is in our case is uploads folder. it is recommended to obtain a secure version of it using the **secure_filename()** function.

```
app = Flask(__name__)


@app.route('/', methods=['GET'])
def index():
    # Main page
    return render_template('index.html')


@app.route('/test', methods=['GET', 'POST'])
def upload():
    if request.method == 'POST':
        # Get the file from post request
        f = request.files['file']

        # Save the file to ./uploads
        basepath = os.path.dirname(__file__)
        file_path = os.path.join(
            basepath, 'uploads', secure_filename(f.filename))
        f.save(file_path)
```

Fig 3.6.2.4

3.3 Team Member's Contribution

Mohit Soni : Worked on Deep Learning Model and Project Report

Vikramaditya Singh Kachhwaha : Worked on Deep Learning Model and Project Report

Vineet Kumar Meena : Worked on Web user Interface and Project Report

Chapter 4

PROJECT OUTCOMES

4.1 Training Accuracy

The model is trained for two datasets, so a comparison is done on both datasets.

In case of MBM campus dataset,

We got a training accuracy of **~71%** trained with 10 epochs with a batch size of 10.

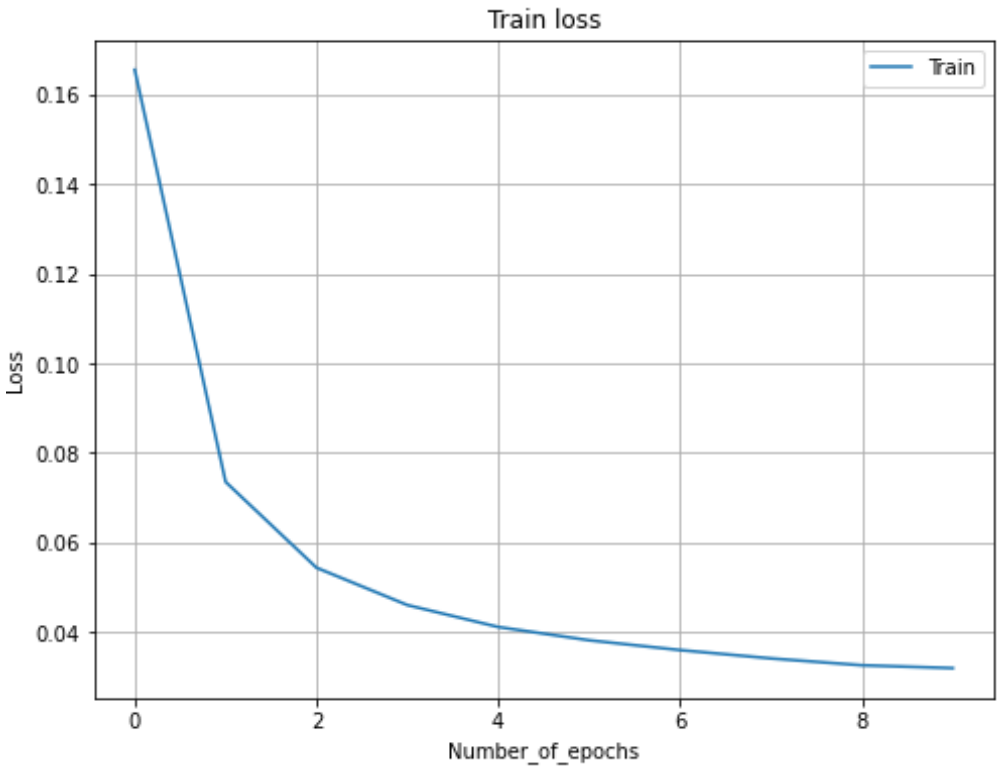

Fig 4.1.1 Training loss graph

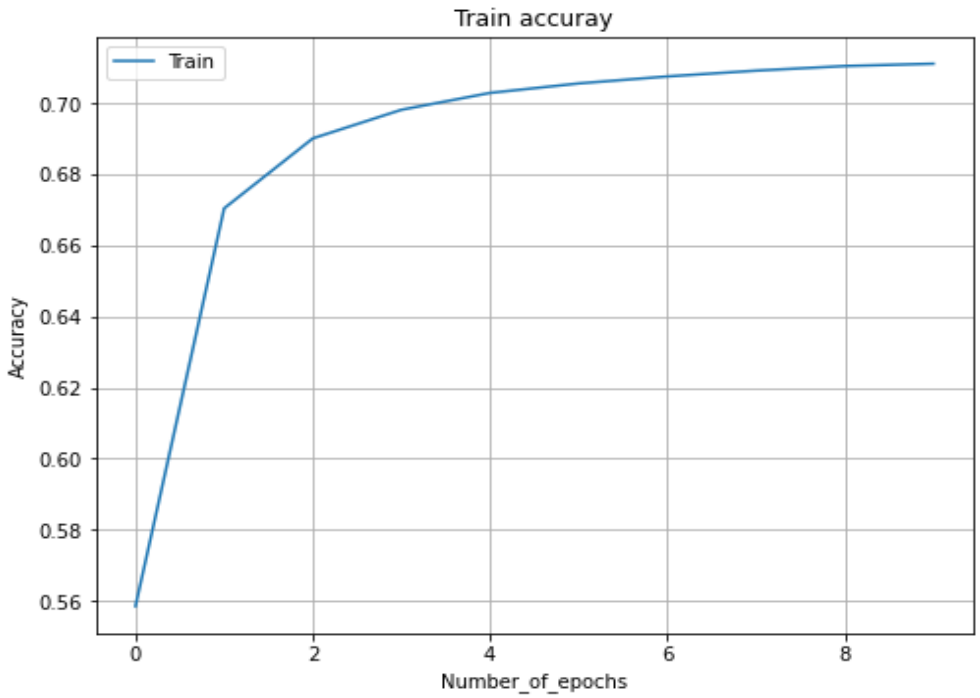

Fig 4.1.2 Training accuracy graph

4.2 Test Accuracy

Test accuracy is calculated on 244 labelled frames. We have measured the performance of our model using confusion matrix.

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy.

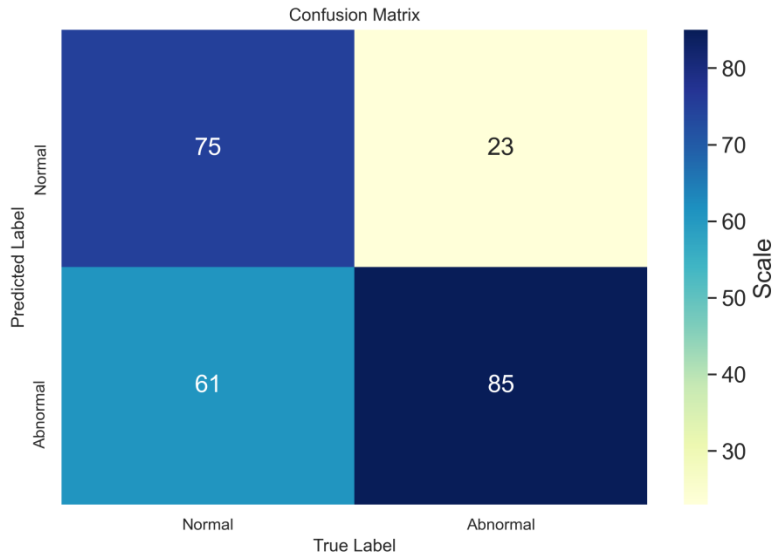

Fig 4.2.1 Confusion Matrix

The different values of the Confusion matrix are:

- **True Positive(TP)**= 85; cases which were actually abnormal and predicted as abnormal.
- **True Negative (TN)** = 75; cases which were actually normal and predicted as normal.
- **False Positive (FP)** =61; cases which were actually normal but predicted as abnormal.
- **False Negative(FN)** = 23; cases which were actually abnormal but predicted as normal.

4.2.1 Accuracy: From all the classes (positive and negative), how many of them we have predicted correctly.

Accuracy=((tp+tn)/total)*100 => 65.57 %

4.2.2 Misclassification: Overall, how often it is wrong.

error=((fp+fn)/total)*100 => 34.42 %

4.2.3 Recall: From all the positive classes, how many we predicted correctly.Recall should be high as possible.

recall=((tp)/(tp+fn)) =>0.78

4.2.4 False positive rate: When it's actually no, how often does it predict yes.

fpr=((fp)/(tn+fp)) =>0.44

4.2.5 Precision: From all the classes we have predicted as positive, how many are actually positive.Precision should be high as possible.

precision=(tp/(tp+fp)) =>0.58

4.3 F-Score

An f-score is a way to measure a model's accuracy based on recall and precision. There's a general case F-score, called the F1-score.The higher an F-score, the more accurate a model is. The lower an F-score, the less accurate a model is.

The F1-score is the most commonly used F-score. It is a combination of precision and recall, namely their harmonic mean. We can calculate F1-score via the following formula:

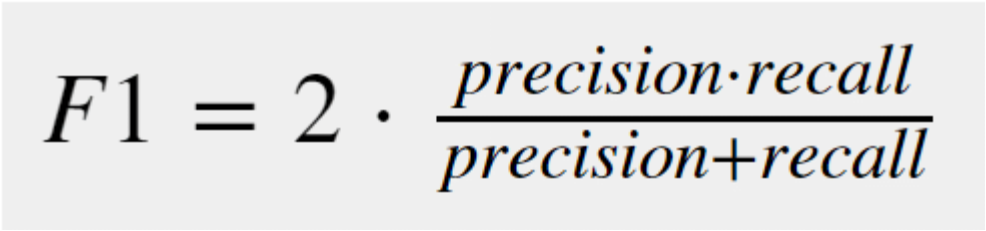

Fig 4.3.1

fscore=2*(precision*recall)/(precision+recall) =>0.66

In case CUHK avenue dataset we compared with our campus dataset:

	Training Accuracy	Testing Accuracy	Misclassification	Recall	False positive rate	Precision	F-score
MBM Campus	71%	65.57%	34.42%	0.78	0.44	0.58	0.66
CUHK avenue	74%	70.58%	29.41%	0.81	0.47	0.74	0.77

Table 4.3.1

CONCLUSION

In this project, we have introduced a spatio temporal autoencoder, which is based on a 3D convolution network. The encoder part extracts the spatial and temporal information, and then the decoder reconstructs the frames. The abnormal events are identified by computing the reconstruction loss using Euclidean distance between original and reconstructed batch. We compared the accuracy by using our dataset to the other available dataset on the internet. We identify the abnormal events based on the euclidean distance of the custom video feed and the frames predicted by the autoencoder. We set a threshold or Loss value for abnormal events.. In future, the work can be extended by the concepts of deep learning approaches. Further, block matching motion estimation techniques such as MPEG-2 and H-264 along with the image perspective can also be experimented. Also, by introducing some fully connected layers with loss function by labeling the frames prediction can be done automatically instead of find a threshold value manually.

REFERENCES

- Malik Ali Gul , Muhammad Haroon Yousaf , Shah Nawaz , Zaka Ur Rehman HyungWon Kim: 'Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture'.
- Yan Fu,Tao Liu,Ou Ye : 'Abnormal activity recognition based on deep learning in crowd'.
- Dongliang Jin, Songhao Zhu, Songsong Wu, Xiaoyuan Jing : 'Sparse Representation and Weighted Clustering Based Abnormal Behavior Detection'.
- Michael George , Babita Roslind Jose1 , Jimson Mathew : 'Abnormal activity detection using shear transformed spatio-temporal regions at the surveillance network edge'.
- Joshan Athanesious , Vasuhi Srinivasan, Vaidehi Vijayakumar, Shiny Christobel, Sibi Chakkaravarthy Sethuraman : 'Detecting abnormal events in traffic video surveillance using superorientation optical flow feature'
- Zhengyin Du, Suowei Wu, Di Huang, Member, IEEE, Weixin Li, Member, IEEE, and Yunhong Wang, Senior Member, IEEE : 'Spatio-TemporalEncoder-Decoder Fully Convolutional Network for Video-based Dimensional Emotion Recognition'
- Thi-Hau Nguyen , Dang-Nhac Lu , Duc-Nhan Nguyen , and Ha-Nam Nguyen : 'Dynamic Basic Activity Sequence Matching Method in Abnormal Driving Pattern Detection Using Smartphone Sensors'
- Athanasios Lentzas and Dimitris Vrakas : 'Non-Intrusive human activity recognition and abnormal behavior detection on elderly people: a review'
- Imran Usman and Abdulaziz A. Albesher : 'Abnormal Crowd Behavior Detection Using Heuristic Search and Motion Awareness'
- Chhavi Dhiman and Dinesh Kumar Vishwakarma , Senior Member, IEEE : 'A Robust Framework for Abnormal Human Action Recognition Using R-Transform and Zernike Moments in Depth Videos'.
- Avenue Dataset : http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
- Confusion Matrix : https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/#:~:text=A%20confusion%20matrix%20is%20a,related%20terminology%20can%20be%20confusing
- Flask : https://www.tutorialspoint.com/flask/flask_file_uploading.htm
- JQuery API : https://api.jquery.com/jquery.ajax/
- HTML,CSS,JAVASCRIPT : https://www.w3schools.com/