Event Detection in Video Surveillance System

Submitted By

Pushti Pethani

16MCEC31

DEPARTMENT OF COMPUTER ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2018

Event Detection in Video Surveillance System

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Pushti Pethani

(16MCEC31)

Guided By

Dr. Priyanka Sharma

DEPARTMENT OF COMPUTER ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2018

Certificate

This is to certify that the major project entitled **"Event Detection in Video Surveillance System"** submitted by **Pushti Pethani (16MCEC31)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in t major project part-I, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Priyanka Sharma
Guide & Professor,
CE / IT Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Priyanka Sharma
Professor
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Alka Mahajan
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Pushti Pethani**, **16MCEC31**, give undertaking that the Major Project entitled "**Event Detection in Video Surveillance System**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Dr. Priyanka Sharma

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Priyanka Sharma**, Associate Professor, Computer Engineering Department, Institute of Technology, Nirma University, Ahmedabad for her valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. Her guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. Alka Mahajan**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Pushti Pethani

16MCEC31

Abstract

Surveillance videos are ubiquitous in the safety critical places. Detecting any type of abnormality in the place using intelligent surveillance system is an active area of research. Traditionally, handcrafted features like HoG (Histogram of Gradients) are used to detect objects from video and . Recently, deep learning techniques have outperformed the state of the art object detection techniques. Convolutional Neural Networks (CNN) are used to solve many of the computer vision problems. CNNs learn the features required for detecting object or event from video frames unlike the traditional systems where features are obtained by a user written program. In this report, we have done the comparison of the HoG features and learned features for object and behaviour detection by implementing two separate machine learning models, Support Vector Machines(SVM) and K-means classifier on custom data-set. Here, performance of the detection system with both features was analyzed. We have also done event detection using CNN and LSTM with the help of GPU and high speed processor. Event detection using CNN, uses only spatial data for classification of video with the help of convolutional neural network. Event detection using LSTM, learns features using pre-trained modal(Inception-V3) and classify video on both spatial and temporal data with the help of LSTM network.

Abbreviations

HOG	Histogram of Gradients
SVM	Support Vector Machine
GOP	Group of Picture
SURF	Speeded Up Robust Features
SIFT	Scale-invariant feature transform
ROI	Region of interest
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory

Contents

List of Figures

Chapter 1

Introduction

Currently there is a huge demand for video surveillance systems as the requirement for ensuring safety of people and property is growing day by day which necessitates the establishment of cameras in almost every corner of the world[1]. Video recording devices used in surveillance are ubiquitous and pervasive in nature. These surveillance systems produce large quantity of video data and thus arise the need of automatic analysis of the video. Video surveillance systems of present day can detect the objects and people automatically. They can identify an abnormal activity or a person which is significantly useful in analyzing the scene in intelligence monitoring system. Hence, object detection and recognition are the most important activities for a video surveillance system. Object detection confirms the presence of an object in the video and also the exact spatial position of the object. Object recognition identifies the object being detected based on the label associated with the object[2].

In real life applications, object detection can be used for various purposes. In this report, we have applied object detection technique in the following couple of scenarios. First, when we want to get the information of till what time a specific person is there in the video scene by extracting only those frames in which a person is actively shown. Second, when we have a specific area which is prohibited for free access where a very few persons are allowed to be entered and rest of the people are restricted to enter and we want to detect the trespasser who entered the prohibited area. Our target is to detect the person as soon as he/she enters the prohibited area and an alarm should be given to the security personnel. Difficulty lies in recognizing the target person and alerting in real time. In such scenarios, intelligent video surveillance system can successfully improve

security in the safety critical locations. In addition to that, it reduces the work load of the human observer in the traditional surveillance system.

The abnormality detection in surveillance video is a process of identifying if an unauthorized person is present in the video scene or not. The process involves a static video recording camera which gives a video recording in compressed format. The constituent frames of the video are first extracted and represented as digital images. These frames are then pre-processed before extracting features from them. Background subtraction and filtering of frames are mainly performed in the pre-processing module. Features are the important characteristics of the images which give information about the visual content of the image. Hence, they are also termed as visual features. Further, these features are fed into a machine learning model to detect the objects or to detect the behavioral pattern of the object. A post processing activity is required to decide if there is any abnormality or anomaly in the scene. So, whenever event or scenario based classification is done, there is another way to extract and learn features. And those extracted features are called as deep features and which are trained in the large and very complexed recurrent neural networks

The remainder of this repot is organized as follows. chapter II describes the Related work. chapter III presents the Object Detection, followed by Event Detection in chapter IV and conclusion in chapter V.

1.1 Motivation

The huge accumulation of digital data in this new century has become an interesting circumstance where storage and processing of such quantities of information are the key factors to satisfy user requirements and expectations. Multimedia data such as video sequences in visual surveillance systems is a very important topic and probably one of the most illustrative examples of this circumstance because the large demand for analysis and synthesis that is needed to understand the contents to determine specific actions based on registered events. Events are phenomena or circumstances that happen at a given place and time which can be identified without ambiguities, for example, a person entering in a forbidden place, a suspicious object abandoned in a public place or a car parking in a garage.

Digital video recording devices are now ubiquitous and pervasive in our daily lives. They are mounted indoors and outdoors everywhere: offices, rooms, halls, banks, hotels, hospitals, casinos, airports, parking lots, buildings, military sites, streets and intersections; some vehicles even have cameras recording passengers and the surroundings of the car.

1.2 Problem Statement

In video surveillance, high accuracy can not be achieved by human operator and this task is very time consuming. Even human operator can not provide accurate evidence when something abnormal is happened. In surveillance system camera should be on 24*7 and video recording and saving is done nonstop so Continuous monitoring using surveillance device causes data overhead.

1.3 Objective

Objective of this project is systems capable to detect events automatically in video surveillance applications and reducing or suppressing human interaction with the system. Also it reports alerts based on the events detected.

1.4 Scope

Scope or we can say applications of video surveillance system are listed below:

- Access control in special areas.
- Person-specific identification.
- Crowd flux statistics and congestion analysis.
- Anomaly detection and alarming.
- Interactive surveillance using multiple cameras.

Chapter 2

Literature survey

2.1 Overview

The main goal of this chapter is to present a review of the state-of-the-art research work in computer vision and machine learning that addresses surveillance applications in the area of event detection.

2.2 Related work

In this section, we briefly give an outline of the features used in detection of objects. Event detection comprises of Object Classification, Tracking, Behaviour Understanding and Description, Personal Identification, Fusion of Information from Different Cameras [3], where object classification can be divided into two categories: i) shape based event detection and ii) motion based event detection. Shape based event detection includes rigid or non-rigid objects like humans, vehicles etc and motion based event detection includes all types of motion. Shape based event detection is not difficult as compared to Motion based detection because shape based event detection procedure works on frame level and motion based detection works on video shots. Shots are a bunch of frames which belong to a scene in the video. All these tasks involve extraction of visual features from the video frames.

Various methods to identify human objects are found in the literature[4, 5, 6, 7]. The traditional approach of human object detection was based on the handcrafted features and using machine learning algorithm to identify the object. These object detection systems

can identify during what time of interval, specific object was present in the scene. in video processing, there is a task, which is mandatory to perform and it is a frame extraction. Frame extraction is done by FFmpeg software and using openCV-python's method. The features are extracted using different feature extraction techniques like that Histogram of Oriented Gradients (HOG)[4], Speed-ed Up Robust Features (SURF)[8], Scale-invariant feature transform (SIFT)[9], Features from accelerated segment test (FAST)[10], Haar wavelets[11] etc. Among all these handcrafted features, HoG features are the most popular ones as they accurately identify the human objects. These features are used to classify the frames of the video by training a machine learning model. Before classification we can do feature reduction also. Feature reduction is used for reducing the dimensionality. There are many techniques to do feature reduction which include Decision Trees[12], High Correlation[13], Backward Feature Elimination[14], Principal Component Analysis (PCA)[13], Linear Discriminant Analysis(LDA)[1], [15].

In the past few years, Convolutional Neural Networks (CNN) have become a strong machine learning model that is able to solve various computer vision problems[16]. Hence, CNN is used to address the problem of human object detection and recognition very efficiently[17, 18]. The key steps in object detection using CNN include CNN feature extraction followed by objects classification and clustering based on the features extracted from the last convolutional layer of the CNN architecture.

To detect the object and recognize it accurately, we can apply machine learning techniques of classification or clustering depending on the type of data available in hand. If the training images are in such a way that we already know the set of classes to which it belongs, then it would be easy to apply the classification technique that is ultimately supervised learning. On the other hand, if training images are totally unlabeled i.e., we don't have prior knowledge of classes into which they fall, in such cases we have to go for clustering techniques of unsupervised learning[19].

Deep learning is an advance version of the machine learning. In the deep learning all networks are gives more accurate results compare to machine learning algorithm because in the deep learning all the networks used back propagation technique when it classify the data. LSTM is used to learn video's spacial and temporal both data. There are many networks who learn spacial data features like RNN, 3D CNN, Opticalflow, using Feature vector neural networks etc. If Event detection is done on only spacial data than 2D CNN

also learn best feaures and gives best results. There are many pre-trained models used to extract features and getting weights, like Xception, VGG16, VGG19, ResNet50, InceptionV3, InceptionResNetV2, MobileNet, DenseNet, NASNet etc. Here all are trained on different dataset. Classification on deep features is better than the classification on handcrafted features , long term dependancy exist between the the data.

Chapter 3

Object Detection

In this chapter we review the principal components of an object detection in video surveillance system as an introduction to the general concepts and algorithms associated with this thesis. Flow of to detect object in video surveillance system given in the below figure.

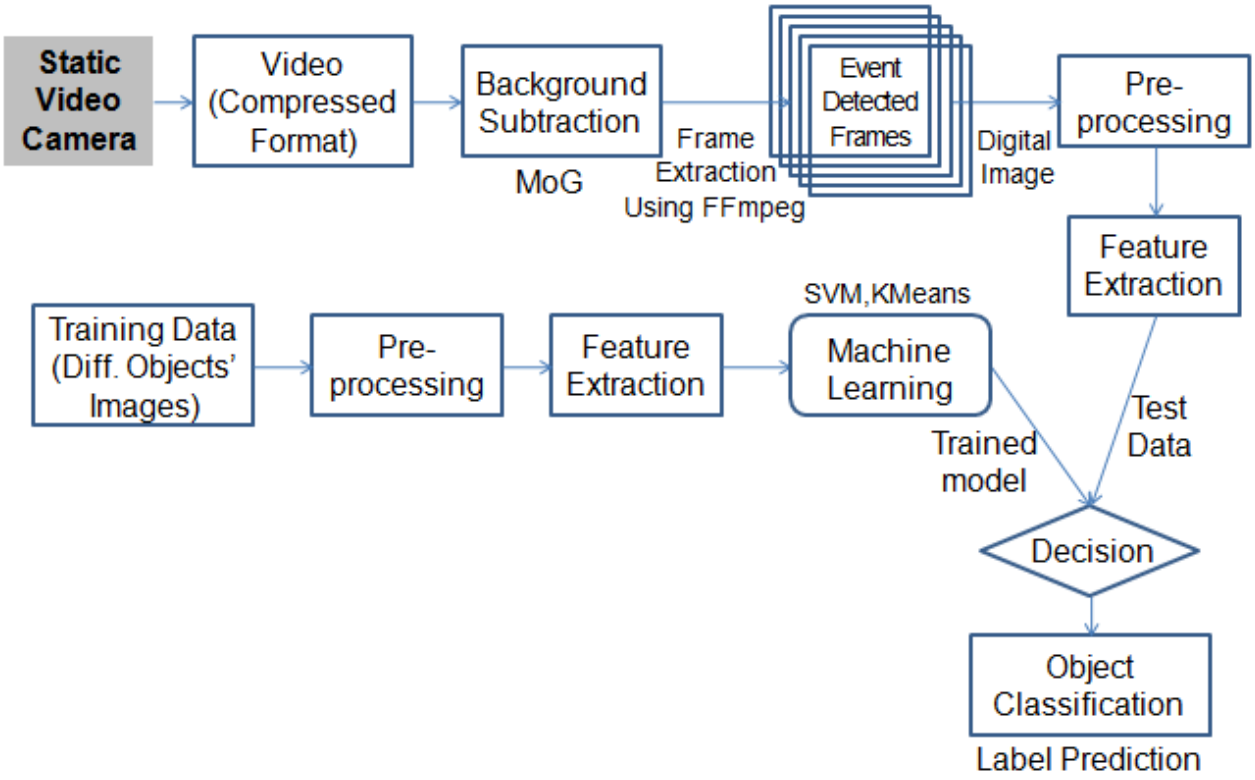

Figure 3.1: Flow of Object detection

In this research work, we implement a human object and behaviour detection system for a video surveillance system and analyze the importance of the features learned from the CNN over handcrafted features such as HoG. In this process, we have followed the

steps shown in Figure ?? to apply the HoG and CNN based features to detect human objects in the surveillance system. Following subsections describe the process in detail.

3.1 Pre-Processing

This is the flow of pre-processing. A brief explanation for each functional block is detailed below.

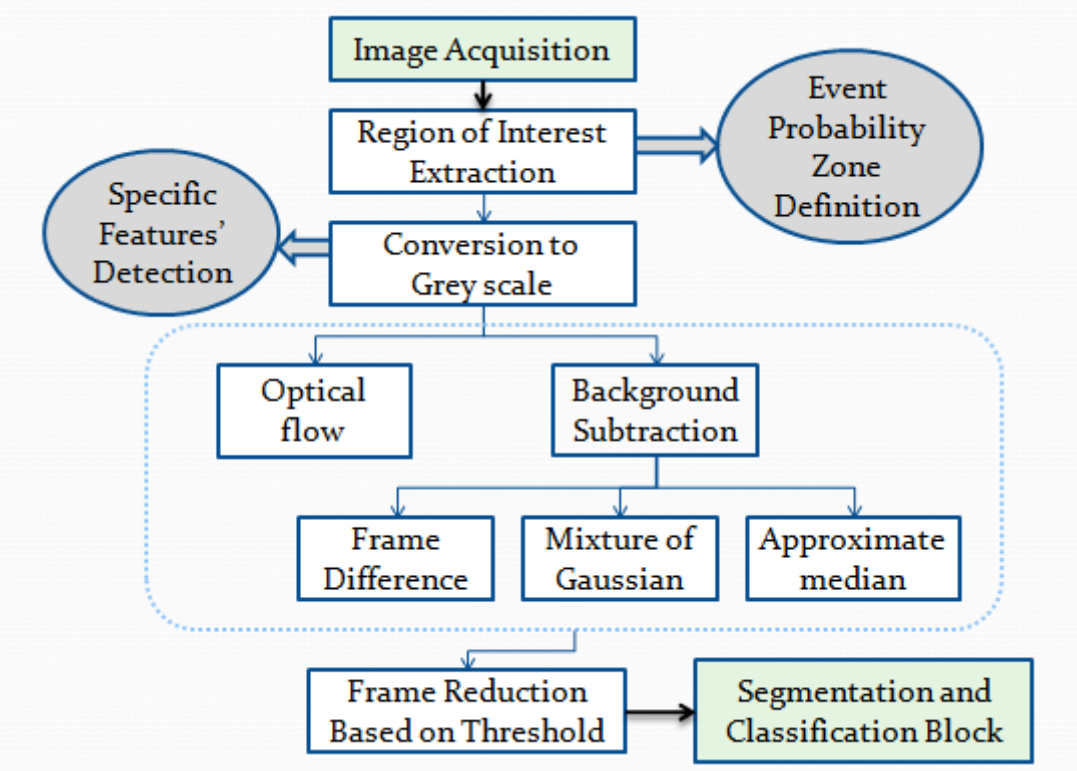

Figure 3.2: Flow of Pre-processing

3.1.1 Region of Interest (ROI) Extraction

The region of interest was extracted according to the area with the highest probability of the event to occur. which you want to filter or where you want to perform some other operation. A region of interest(ROI) is a portion of image that you want to filter or perform some other operation on.

3.1.2 Conversion to Grayscale

Color components in images and video sequences can be useful during process specific features detection, such as skin color or face recognition of human being. Grayscale images are related to the luminance component in the YCbCr color space, Where Y is

the luminance component that gives the average brightness of the image and Cr and Cb are chrominance for red and blue color components. The conversion of color image into gray scale image is converting the RGB values(24 bit) into gray scale value(8 bit). So, processing time of gray image is 3 times lesser than the RGB image.

3.1.3 Foreground/Background Estimation

As a prepossessing step, mainly two techniques are used. First one is Background subtraction and second is optical flow generation. If camera is static then background subtraction is used and if camera is moving while recording the video, then optical flow technique is used to prepossess the data. In this research work, Closed-Circuit Television (CCTV) camera is used. Hence, background subtraction is applied as shown in Figure 3.3. By using background subtraction technique we got the difference between frames. There are techniques like Frame difference, Mixture of Gaussian (MoG) and Approximate median for background subtraction. Mixture of Gaussian (MoG)[20] is good enough for subtracting background from different frames.

Let the Video Frames be represented as:

$$V = \{F_1, F_2, F_3, ..., F_n\} \tag{3.1}$$

where V is a video and F_i is the constituent frame with i ranging from 1 to n. The number of frmaes in a video is represented by n. Background subtraction is carried out by subtracting the frame from its previous frame.

$$S_n = F_n - \{F_{n-1}, F_{n-2}, F_{n-3}...F_m\} \tag{3.2}$$

where S_n is the difference of frames and n is grater than m. After subtraction, objects are either moved from one position to other or disappear in the difference frame and the pixels of background remain same. Change in background is considered as noise. The changes in pixels at the boundaries of object is more important[21].

Here Mixture of Gaussian (MoG) is used for background subtraction[22]. Every pixels in those frames are modeled by Mixture of K Gaussian distribution. The probability of

certain pixel that has a value of X_N at time N can be written as:

$$p(f_N) = \sum_{i=1}^{K} \omega_i \eta(f_N; \theta_i) \tag{3.3}$$

where ω_k is the k^{th} Gaussian component weight parameter and the k^{th} component normal distribution represented by:

$$\eta(f; \theta_j) = \eta(f; \mu_j, \Sigma_j) \tag{3.4}$$

$$\eta(f; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_j)^T \sum_j^{-1} (x-\mu_j)} \tag{3.5}$$

where μ_k is represented as mean and $\sum_j = \sigma_j^2 I$ is the j^{th} component co-variance. The first D distributions are used as mixtures of the background model of the scene, where D is evaluated as:

$$D = arg_d min \left(\sum_{i=1}^{d} \omega_i > T \right) \tag{3.6}$$

Threshold T should be minimum fraction of background model; it is lowest prior probability that background in particular frame. Subtraction of background is done by marking the foreground pixels and pixels which is greater than 2.5 standard deviations away from any D's distribution. Here equations are being updated by updating first Gaussian component that matches the test value.

Figure 3.3: Process of background subtraction

Figure 3.3 shows the subtracted images after Mixture of Gaussian (MoG) background subtraction. In these images moving objects are separated from background of particular frame/image. Unwanted information is removed from original frame/image.

3.1.4 Optical flow

The optical flow methods used to evaluate the optical flow vectors of pixels in every different frame were successfully tested. The Horn-Schunck method showed an accurate evaluation of vectors, but the processing time is very long as it is shown in the time analysis section.On the other hand, the Lucas-Kanade method performs much faster in processing time and the optical flow density can be handled changing the Window-Size parameter.

3.1.5 Frame Reduction

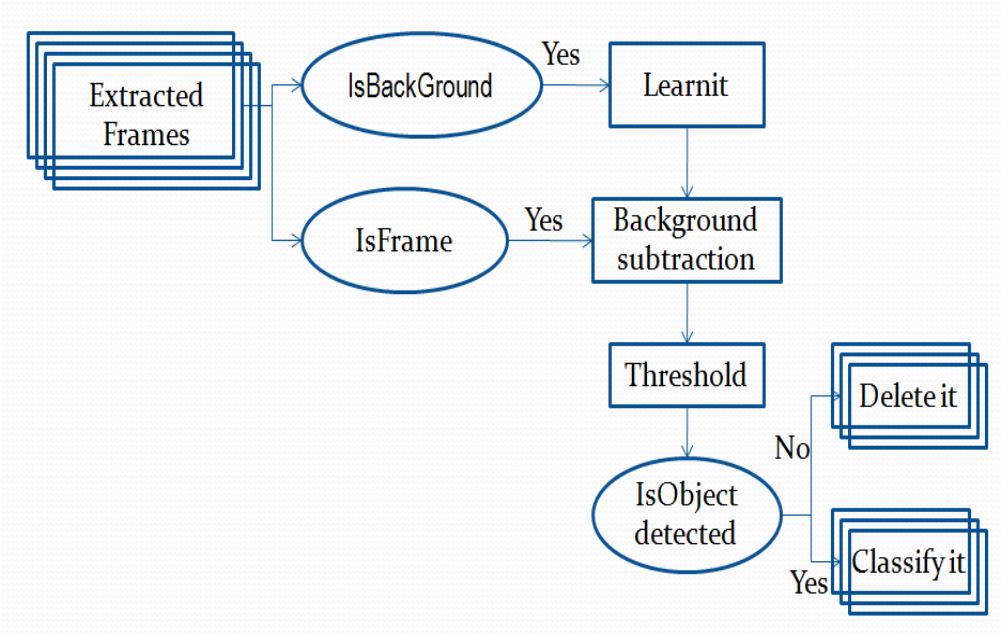

Figure 3.4: Flow of Frame Reduction

In this step, all extracted and background subtracted frames are processed on their pixel values. Here, a threshold value is set for reducing the frames. If number of black pixels in any frame is higher than the threshold value then that frame is dropped because in those frames no event would have taken place. Background subtraction output is used for taking the decision if a human object is present or not. Generally, in this procedure, large number of frames are extracted in a very small amount of time span. So data overhead will occur to process those large number of frames. There is an alternative to

reduce frame generation in given amount of time, which is done by setting the parameter FPS(Frames Per Second) to a comparatively large value. But the major disadvantage of setting the large value of frame rate is, we may lose some important frames in very small amount of time span. And we can't capture object in CCTV footage too. The frame reduction procedure is depicted in Figure 3.4.

3.2 Features Extraction

A feature is a significant piece of information extracted from an image which provides more detailed understanding of the image.Good feature extraction is important for the performance of classification.

3.2.1 Types of feature

- Low-level Visual Features
  - Color
  - Texture
  - Shape
  - Motion
  - Shot Boundaries
- Mid-level Semantic Content
  - People/Objects
  - Location
  - Actions
  - Time
- High-level Semantic Content
  - Story
  - Concept
  - Event

3.2.2 Feature extraction

In this step, we extract HoG and CNN based features from the video frames which are preprocesed in the previous step. HOG is an edge orientation histograms based on the orientation of the gradients in localized region called cells. HOG can represent a rough shape of the object, that is why it has been used for general object detection and recognition, such as people or cars. On the other hand, rotation and scale changes are not supported using HoG.

HOG is used to extract global features whereas other techniques are used for extracting local features. So, for object detection, global feature extraction techniques give better results in identifying different objects from images. The algorithm of HoG feature extraction is given in Algorithm 1.

Algorithm 1 : HoG

Input: Constituent Frames of Surveillance Video

Output: Histogram of Gradients

Steps:

1. The input image is transformed to gray-scale scale image
2. The luminance gradient is computed at every pixel in the image
3. To generate a HoG orientation for each cell
    - Feature quantity becomes robust to changes of form
4. Normalization and Descriptor **Blocks**
    - Feature quantity becomes robust to changes in illumination

---

Deep learning is gaining popularity in solving the major computer vision problems like image classification, object detection, human action recognition etc. CNNs have shown promising results in all these areas by replacing the complicated procedure of extracting handcrafted features[23]. The architecture of a CNN model is shown in Figure 3.5.

A CNN model consists of convolutional layers followed by sub-sampling or max-pooling layers and the fully-connected layers neural layers at the end. The stack of convolution and sub-sampling layers learns the generic features from the input images and these features are fed into the fully-connected layers for classification. The output of

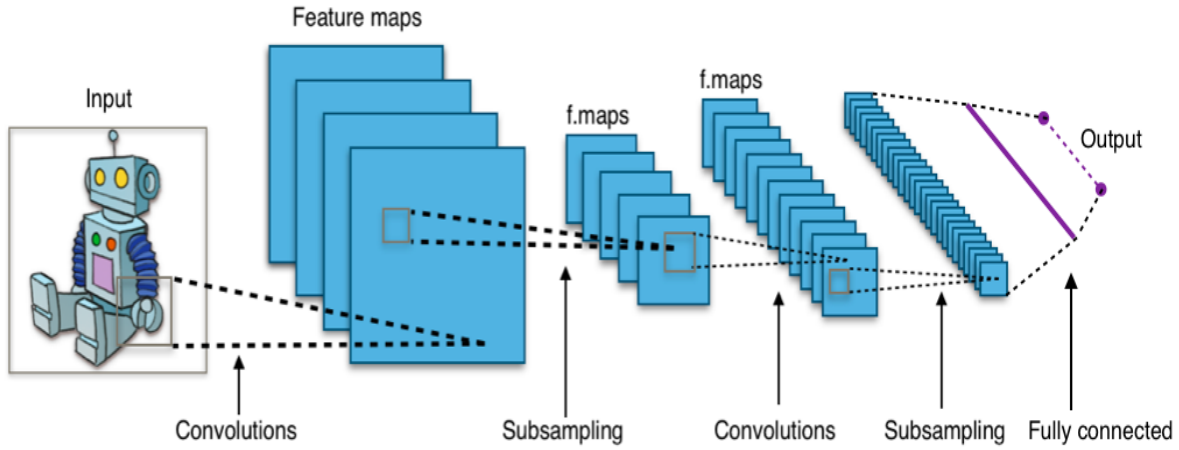

Figure 3.5: Architecture of CNN model

last sub-sampling layer work as the visual feature descriptions. These are called learned features as they are learned by a CNN model automatically. These models work very well when trained on a large dataset. But, training takes a long time depending on the size of the network and computational capabilities available. Hence, we used a pre-trained CNN model called VGG-19 [24] which was trained on 1000 classes of Imagenet[25] data. This is a medium sized CNN model with 19 layers. The output of the last layer of max-pooling is taken as the learned feature.

3.3 Frame Classification

Using classification of frames in the testing phase, we can detect anomaly or specific object from the frame according to the anomaly defined during training phase. At training module, all the features of images consisting of the different types of objects are given as an input. Support Vector Machines(SVM) are used to classify the objects in the current implementation. Then the features from the pre-processed frames which are actually testing data, are passed into the created SVM model to predict the object is anomaly or not. Here we used multi-class SVM to do classification. It divides the training data into multiple class using their feature space and finding the pattern from those images which are having the same class. The testing samples are predicted by probability distribution to related class using multi-class SVM classification model. The training and testing processes are illustrated in Figure 3.6 and Figure 3.7 respectively. The features extracted from the frames are denoted as X_n where n is the number of frames in the video as shown

in Equation 3.7.

$$X_n = \{X_0, X_1, X_2, ..., X_n\} \tag{3.7}$$

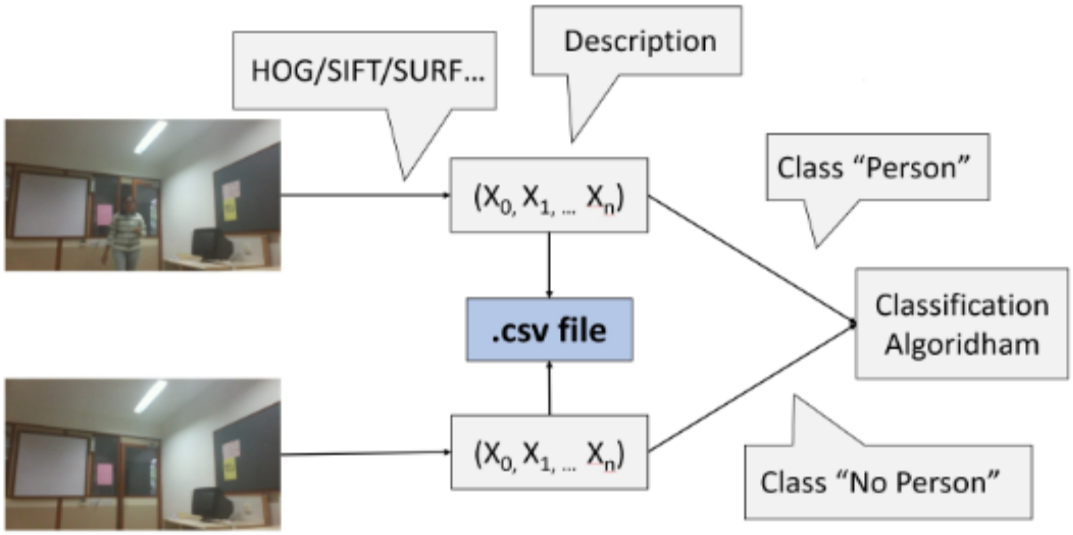

Figure 3.6: Training process

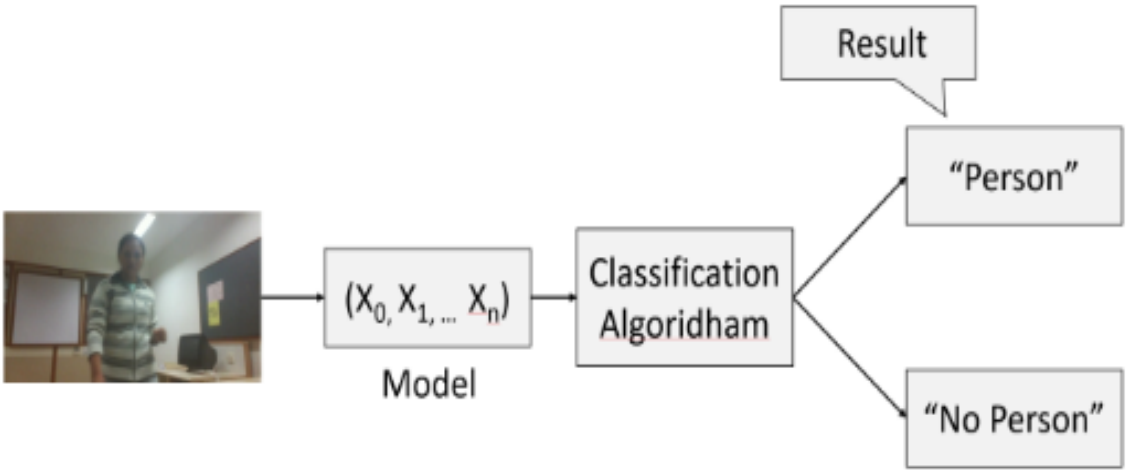

Figure 3.7: Testing process

Here we train model using our own data set which consist of images of multiple objects. and testing is on the pre-processed extracted frames of CCTV footage.

After classification, post processing of classified data is carried out to find out that actually which frames have abnormality using the temporal relation. Temporal relation is defined as the correlation between the consecutive frames of the video. The steps used for post processing are given in Algorithm 2.

Algorithm 2 : Classification

Input: Output of SVM classifier

Output: Frame-level object detection and recognition

Steps:

Level-1 Post-processing

1. **for** all filtered frames
2. **if** All blocks in a L1 frame are normal/abnormal **then**
3. Frame <– Normal/Abnormal
4. **else**

Level-2 Post-processing

5. **if** All blocks in L2 frame normal/abnormal **then**
6. Frame <– Normal/Abnormal
7. **else**
8. **Vote** (all neighbouring blocks)
9. Blocks <– Normal/Abnormal
10. go to step-6
11. **end if**
12. **end if**
13. **end for**

Temporal post-processing

14. **Perform** temporal post-processing

In level-1 post processing, simply the frame is identified as normal or abnormal based on the class labels obtained by SVM classifier. In level-2 post processing, identification is on a group of frames if that block contains abnormal objects or not.

3.4 Experiments and Results

For the experiments, we have used the video of 9 seconds as a testing data-set. In that video, there is a girl walking inside a room and through CCTV camera we captured that scene. In that video, girl actually comes in video at 5^{th} second and she is actively available till 9 seconds. Total extracted frames are 339 and out of 339 frames, girl is visible in only 144 frames and so they are selected as useful frame. Remaining frames have no event of any object so we can discard them. We have provided few images of 3 different persons as an input training set as shown in Figure 3.8. Out of those images, first type of images is the set of images of the girl who is captured in the camera during testing. Few images are of another girl and rest of the images are of a boy's images. We have worked upon two algorithms, namely SVM and K-means clustering and built the model using the training data. SVM gives higher accuracy of 95% compared to K-means because K-means algorithm is behavioral algorithm which considers overall gestures and postures of the person. And in the testing video, girl's gesture and postures are somewhat similar to boys.

Figure 3.8: Samples from training images

Table 3.1: Results in terms of accuracy

	CNN Features		HoG Features	
	SVM	**KMeans**	**SVM**	**KMeans**
Object Detection	**94%**	**-**	**87%**	**-**
Behavioural detection	**-**	**93%**	**-**	**75%**

Chapter 4

Event Detection

4.1 Types of Event Detection

- Object Tracking
- Behaviour Understanding and Description
- Personal Identification and Observation
- Anomaly detection and alarming.
- Fusion of Information from Different Cameras.

4.2 Pre-processing

In this module, pre-processing consist of two steps, one is re-size of images and second one is normalization of image. After image re-sizing of image, Image of any sized will be converted into image pixel size of (299,299). It is easy to process images of constant pixel dimensions. In image normalization, pixels value of re-sized image will be normalized between 0 to 1.

Image normalization will reduce input data size for next processing module, which will help to process input data faster. At the same time, image normalization will help to remove noise also. In proposed method pre-processing, which includes image re-sizing and image normalization, is used in both Event detection By classification of images/frames using CNN and Event detection By classification of shots/clips using LSTM.

4.3 By classification of images/frames using CNN

Methodology of Event detection By classification of images/frames using CNN is explained in the below figure. In proposed method, First step is to extract frames from different types of event's video clips using FFmpeg software. Here, in ordered to extract frame, video clips read one by one. Those extracted frames will be written in the folder, from which clips are saved. Next, all the frames are pre-processed as discussed earlier.

Here, temporal data or we can say it as temporal features of the different types of events clips are ignored for the classification of video. Here, classification is done only spacial data using the 2D Convolutional Neural Network. Now, Pre-processed data/frames are used as input for convolutional neural network. In the Convolutional Neural Network, only top layers will be trained using weights of pre-trained model. In proposed method, InceptionV3 is used as pre-trained model. Now we have trained model which we can use further for testing purpose. Now, For testing, We passed any random image into new trained model for prediction of event, but image for testing must having picturization of any event which event is in the training data-set at-list.

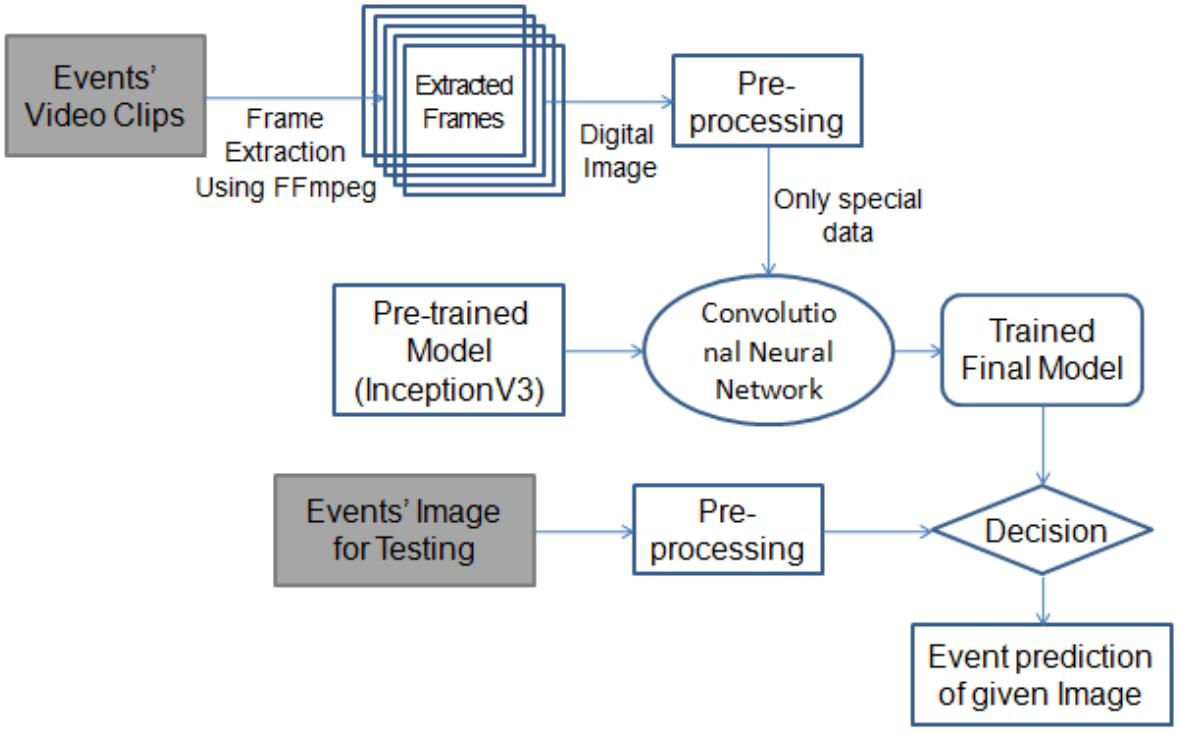

Figure 4.1: Flow of Event detection By classification of images/frames using CNN

4.3.1 Description of the Architecture

The below figure describes architecture of Convolutional Neural Network. In our proposed method Event detection is done by classification of images/frames using CNN is done only on spacial data. CNN having multiple layers like Convolutional layer, pooling layer, fully connected layer and input-output layer. In this network starting layers such as convolutional and pooling layers are used to extract features, where fully connected and top dense layers are used to classify the given input. In CNN Convolutional Here temporal data is ignored. There are various layers in the CNN, In our methodology we train our data only on the top layer, other layers are freezed out because we've used pre-trained modal. For top layer input is pre-processed image and fine-tuned weights of pre-trained model (Inception-V3). While fine tuning the weights, only top dense layer will get included. Top layers of CNN have two layers which both are dense layers. First dense layer's input is GlobalaveragePooling2D layer's output, which having 1024 neurons and RELU activation. Second dense layer has five number of neurons which are equal to number of classes. The activation of second dense layer is Softmax. After training of top layers we get new trained modal which is used for further processing. Test data will be passed and trained into new trained modal and it will give classification of input image/frame as a output.

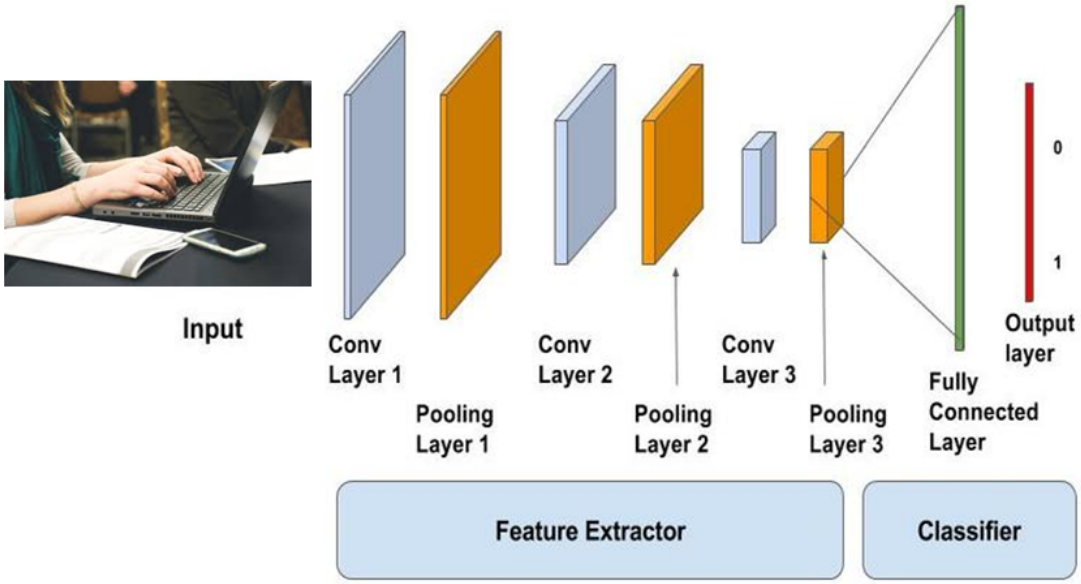

Figure 4.2: Architecture of CNN

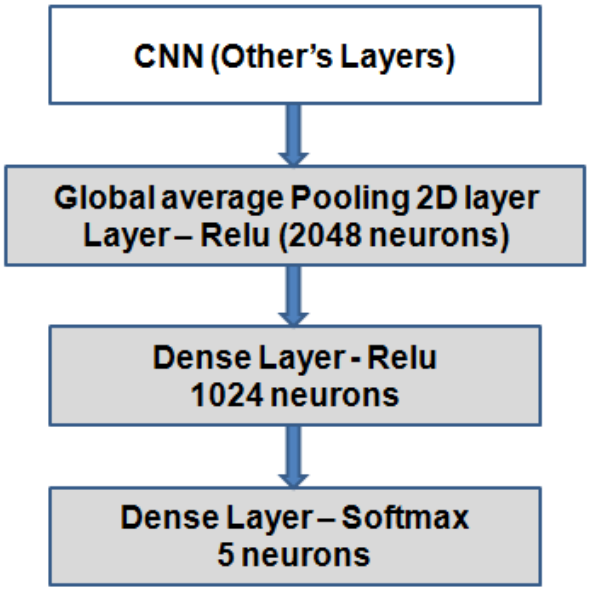

Figure 4.3: Architecture of CNN's Top layers

4.4 By classification of shots/clips using LSTM

Methodology of Event detection By classification of shots/clips using LSTM is explained in the below figure. First, We have extracted set of frames from the different types of events' shots. Here, In this proposed method, We have taken booth spacial as well as Temporal features as input for video classification. For consideration of temporal feature we will take new parameter as value of number of frames of a shot. So, For this reason we have considered only 40 frames from every shots in the pre-processing module. And it will be passed out for next module as deep feature extraction. To filter out the only object's features from the pre-processed data frames rather than any background's features, we have used inception v3 model to extract the features. So InceptionV3 model will extract deep feature of those object by whom event is occurred actually.

Here we have used LSTM model for the training n testing purpose. Whatever features we have extracted using InceptionV3 of 40 frames from each shot, which will be considered as a one pattern. We will pass pattern of 70% shots of data-set as a training set in the LSTM model and whenever the new trained model is generated, it will be further used for the testing. Now rest of the 30 data will be passed into new generated model which are considered as testing data, and it will give the output for the testing data and classify the testing video clips/shot's event into related class. If testing data will have the normal

event than output of the LSTM will be shown as the normal event's class and consider it as normal event else output of the LSTM will be shown as an abnormal event's class and consider it as abnormal event.

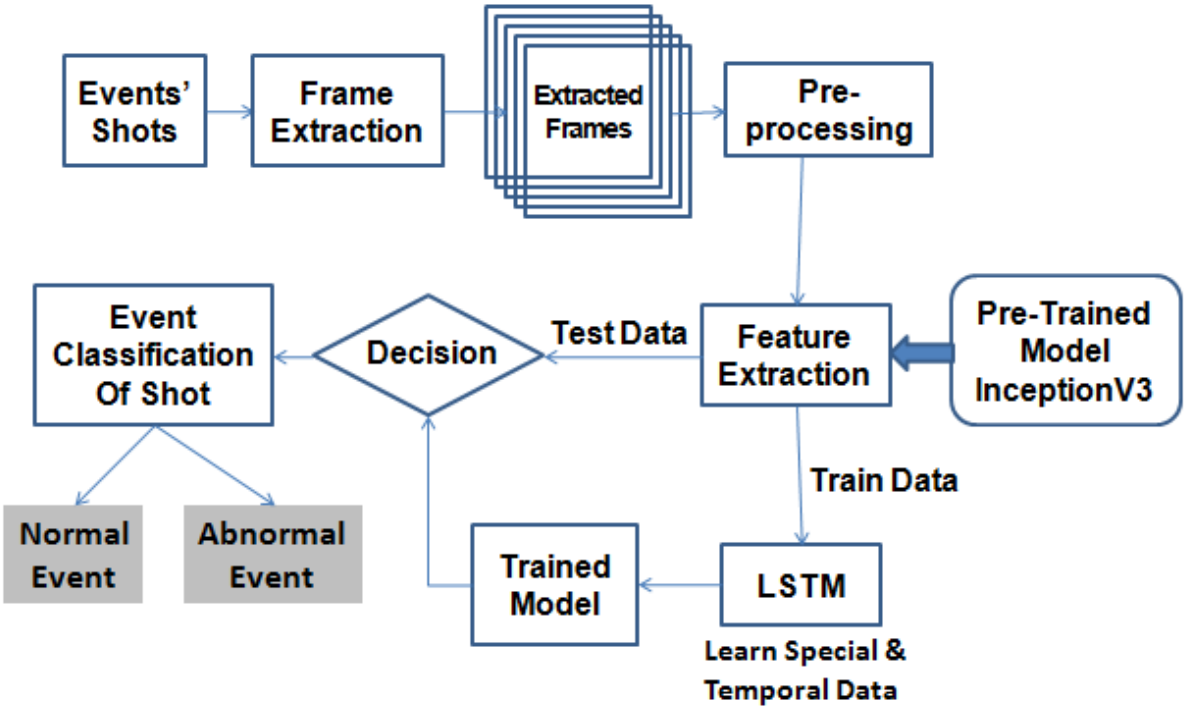

Figure 4.4: Flow of Event detection By classification of shots/clips using LSTM

4.4.1 Deep feature extraction

For decreasing the complexity of the input and obtaining high-level features of the images the first neural network which known as "Convolution Neural Network" is used. Here, "Inception" which is pre-trained model developed by Google is used by us. The third version of Inception is trained by using ImageNet Large Visual Recognition Challenge data-set. ImageNet data set have around thousands of classes for different types of objects. Classification of images and identifying different classes like "Person", "Animal" and "bird" by using different models is an ideal task in computer vision.

For applying the transfer learning method we are using this model. For perfect object recognition or detection we have to pass millions of parameters into neural network or we can say it as a model, and train those parameters but it will take much time. So by using fully-trained model those training process will get faster and retrain old weights for a new classes which are exist in our system, using some data-set like ImageNet.

In proposed method we are detecting event associated to some object. Pre-trained model(Inception-v3) is used to extract and filter out particular object features from the extracted frames. So, while training it will look for particular object and this will result in better training. Therefore, we can get better results at testing phase.

The below figure represents architecture of Inception-v3, Which consists of multiple layers. First layer is input, second layer is convolution and third layer is output layer. Convolution layer have multiple layers like average pool, Max. pool, Dropout, Concate, Fully-connected and Soft-max layer. In input layer pre-processed images/frames will taken as input. when frames get passed through convolutional layer it will filter out object and in top layer features will get extracted of those object.

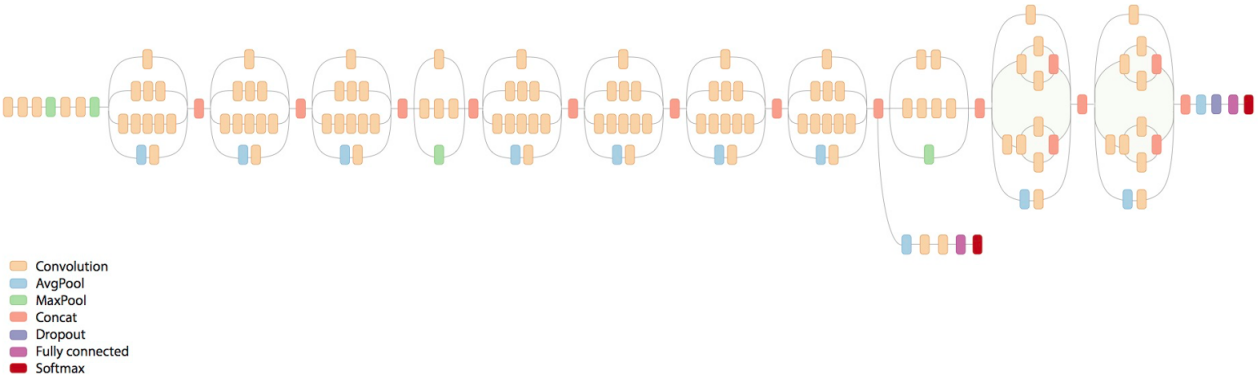

Figure 4.5: Architecture of InceptionV3

4.4.2 Long Short term Memory

LSTM stands for Long Short Term Memory. LSTM is type of Recurrent neural network, but is has one memory cell which will help in saving long-term memory. This Long Short Term Memory (LSTM) is used to learn temporal data. There is one memory unit in LSTM cell by which some long term dependency is processed and important things are saved. LSTM is used to reduced vanishing gradient problem.

Sigmoid functions have values 0 or 1. So this function is used to forget or remember the memory information. So If it hold '0' value than it will forget the memory information and if it holds the '1' value than it will remember the memory information.

The above figure describes the structure of Long Shot Term Memory cell with forget

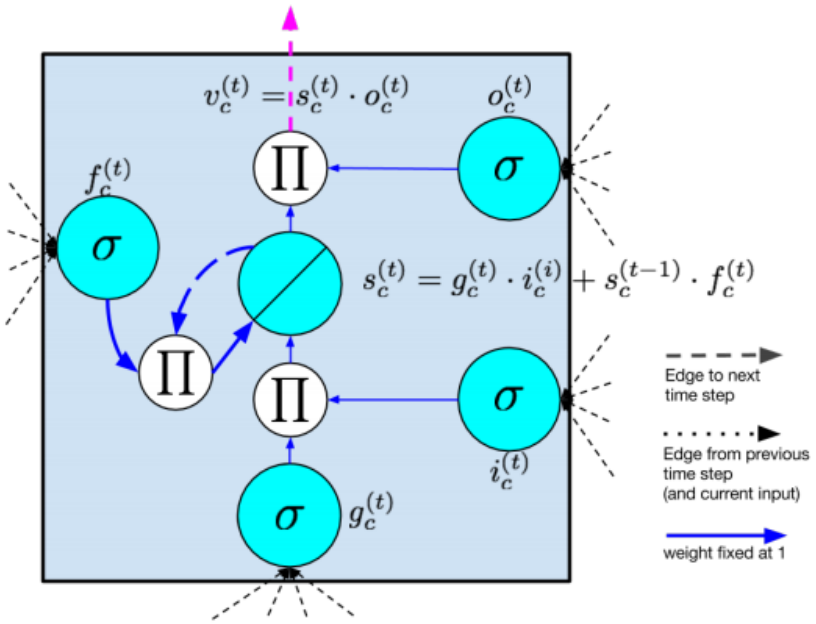

Figure 4.6: LSTM memory cell with a forget gate

gate. The Blue circles in the figure represents sigmoid functions. The white circles in the figure represents multiplicative node. In the image i_c represents input gate, f_c represents forget gate and o_c represents output gate. The value for input gate, forget gate and output gate is in binary format. So, they will hold value either '0' or '1'. Dashed line in figure represents edge to the next time stamp. Dotted line represents edge from previous time stamp (and current input).

Input i_c to the memory cell from current memory data is concatenated with input gate and the results will tell that how much important this current input is for this cell. s_c is very important unit for this memory cell data. It has self loop for recursion. Sc is calculated on the basis of previous memory cell and current memory. Importance of previous memory will be decided by f_c. If f_c is 0 than previous memory will get discarded and if f_c is 1 than previous memory will get passed to next step in memory cell. Now, s_c is calculated by adding current memory data and previous memory data. Now, how much s_c important for a next memory cell is denoted by O_c gate.

RNN(Recurrent Neural Network) is not more capable to save and processed long term dependency between data. For detecting an event in video we have to processed long term memory to identify what kind of event is happening. For event detection there are various time stamps in the video and there is some particular sub-event associated with each time

stamp. According to those time-stamp and associated sub-event we will be able to know the actual whole event. For this procedure we have to know some pattern which exist between time-stamp and sub event. so LSTM is used for learning this type of pattern easily.

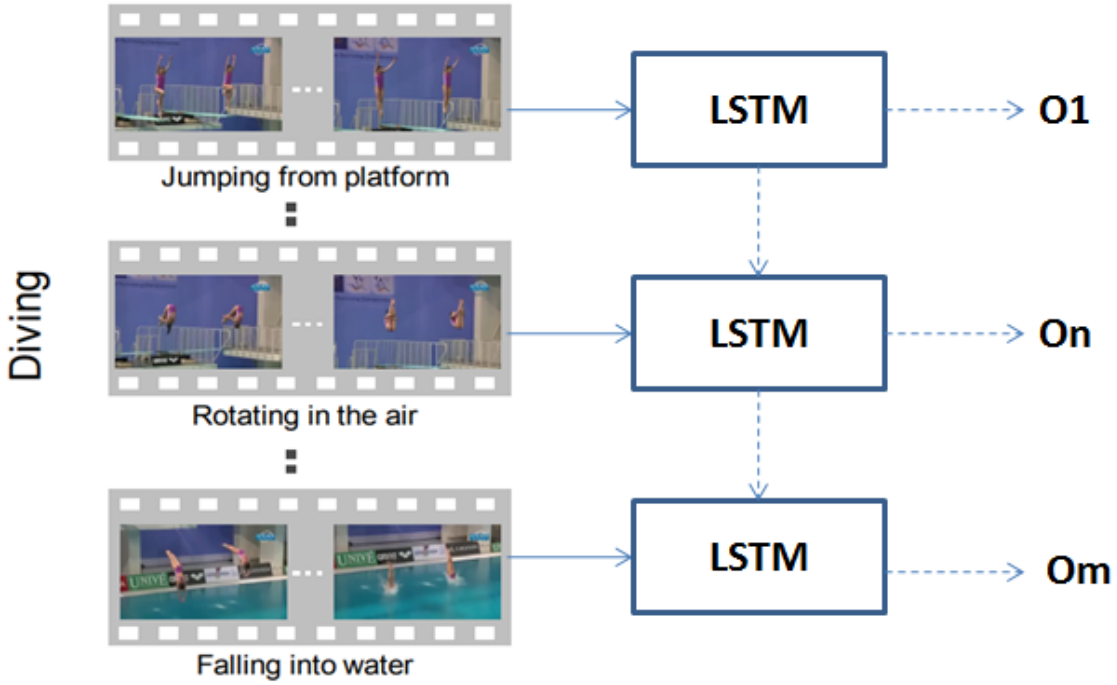

Figure 4.7: Event Detection using LSTM

This above figure shows one example of how event is detected by LSTM. In this example the whole procedure for diving event is get detected by three different sub-events happening in the sequence. For this first LSTM will detect the event of Jumping from platform at first time-stamp. In the second time-stamp LSTM will detect the event of Rotating in the air and in the third time-stamp LSTM will detect the event of falling into water. When this three sub-events is learned by LSTM than it will conclude that event of diving is happening.

4.4.3 Description of the Architecture

In the proposed method for event detection, The Recurrent neural network is used which architecture is given below. this is used to find a pattern between different types of actions by giving the sequences of shots/clips. In this network Long Short Term Memory cell is exist as first layer. After than there are another three sequential hidden layers

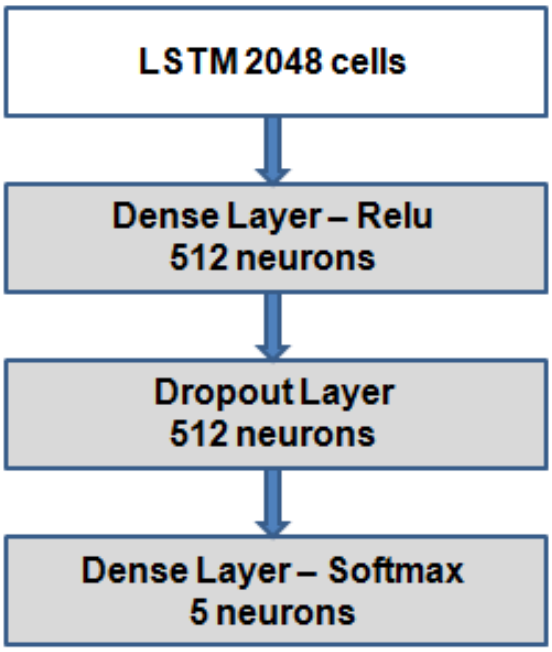

Figure 4.8: Recurrent neural network

exist for classification process. First is dense layer, Which having 521 neurons and RELU activation and it takes input from LSTM layer. Then second is Dropout layer which having also 512 neuron and drop out rate is 0.5. and last and third layer is Dense layer which having 5 neurons(same as no of class) and Softmax activation. This final

The second neural network used was a recurrent neural network, the purpose of this net is to make sense of the sequence of the actions portrayed. This network has an LSTM cell in the first layer, followed by two hidden layers (one with 1,024 neurons and RELU activation and the other with 50 neurons with a sigmoid activation), and the output layer is a three-neuron layer with Softmax activation, which gives us the final classification.

4.5 Experiments and Results

In proposed method, The data-set is of different five types of video clips/shots. There are 679 clips divided into five classes named as Typing, Writing on the board, Mopping on the floor, Jumping and Fighting. Here, First two are normal in the classroom for surveillance system where another three are abnormal event in the classroom. For frame extraction, Frame rate(Number of frames extracted in a second) is 3. In proposed method, Used data set is part of UCF-101 data-set. In every class, There are on an average 100 video

clips/shots of [1-2] min available. When, frame extraction is done, On an average there are 180 frames are extracted from a shot/clip.

In the Event detection By classification of images/frames using CNN methodology, There are around 1 lakh images used for training and around 40,000 images used for testing of Data-set. Where, In the Event detection By classification of shot/clip using LSTM methodology, There are 476 clips used for training and 203 clips used for testing of Data set. But in this direct video can not be passed into network directly so it's size of feature set of a shot/clip is (40,2048). Where, 2048 is feature length and 40 is number of frames which is considered as a pattern. Both classification is done with the help of Keras Library and many powerful hardwere devices(RAM : 7.7 GiB, Processor : Intel Core i7-4790 CPU @ 3.60GHz * 8, Graphics : Tesla K40c/PCIe/SSE2, OS : Ubantu 16.04 64-bit). Actually, Number of classes is very less for our video surveillance system. There are so many normal and abnormal events performed under surveillance system and we can take it and also classify it but it will take more time to train on above mentioned system also.

In this research work, In the CNN architecture, Data set is trained up-to 45 epochs on the Nvidia tesla k40 GPU and it took around 6hr for completing whole training process. Where In the LSTM architecture, Data set is trained up-to 38 epochs on the Nvidia tesla k40 GPU and it took around 2hr for completing whole training process. Accuracy and loss of both methods are given into below table.

Table 4.1: Results in terms of accuracy and loss

	Loss	**Accuracy**
Event Detection using CNN	**0.25**	**89.12%**
Event detection using LSTM	**0.097**	**94.87%**

Below figure is of results of event detection using CNN. In this figur, list of class name and predicted results according to given testing image are written behind the testing image. Here, Typing, Writing on the board and Mopping on the floor are considered as normal event, where Jumping and Fighting are the abnormal event under the surveillance system of classroom.

data/test/image1.jpeg
WritingOnBoard: 0.91
MoppingFloor: 0.06
Typing: 0.02
Fighting: 0.01
Jumping: 0.00

data/test/image2.jpeg
Fighting : 0.85
Jumping : 0.14
WritingOnBoard : 0.01
MoppingFloor: 0.00
Typing : 0.00

data/test/image3.jpeg
Typing: 0.96
WritingOnBoard: 0.02
MoppingFloor: 0.02
Fighting: 0.00
Jumping: 0.00

data/test/image4.jpeg
MoppingFloor: 0.79
Fighting : 0.13
WritingOnBoard : 0.8
Jumping : 0.00
Typing : 0.00

Figure 4.9: Results of Event Detection using CNN

Chapter 5

Conclusion

Video surveillance systems are expected to be intelligent enough to detect object automatically to reduce the human operators involvement in the process. Features extracted during the object detection play a major role in the process. In this paper, We have implemented a object detection system for video surveillance using both handcrafted HoG features and CNN based features and study on human object detection and behavioral detection using SVM and K-means algorithm. CNN features are learned using the VGG-19 architecture which was pre-trained on 1000 classes of Imagenet data. It is evident from the results that learned features are the best for the object detection.

In the event detection in the video surveillance system, Classification on spacial and temporal data using convolutional neural network of the video gives around 89% accuracy result to detect any particular event on the around 1 lakh images of the different types of events' images, where the classification on spacial data of video only using long short term memory network gives the around 95% accuracy on the almost 450 video clips on the different types of events' clips.So, conclusion is CNN based approach is batter than the LSTM based approach.

Chapter 6

Future work

- Will use appropriate live database of surveillance system and show comparative results.
- Classification will do on live surveillance dataset.
- Efficient learning and inference procedures will be presented for this project using comparative study.

Bibliography

[1] X. Li and Z.-m. Cai, "Anomaly detection techniques in surveillance videos," in *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on*, pp. 54–59, IEEE, 2016.

[2] Y. Amit, *2D object detection and recognition: Models, algorithms, and networks*. MIT Press, 2002.

[3] R. A. C. Jimenez, *Event detection in surveillance video*. Florida Atlantic University, 2010.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[6] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3626–3633, IEEE, 2013.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[9] T. Lindeberg, "Scale invariant feature transform," *Scholarpedia*, vol. 7, no. 5, p. 10491, 2012.

[10] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *European conference on Computer vision*, pp. 183–196, Springer, 2010.

[11] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[12] X. Geng, D.-C. Zhan, and Z.-H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 6, pp. 1098–1107, 2005.

[13] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[14] K. Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 629–634, 2004.

[15] K. Vignesh, G. Yadav, and A. Sethi, "Abnormal event detection on bmtt-pets 2017 surveillance challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 2161–2168, IEEE, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[19] C. Donalek, "Supervised and unsupervised learning," in *Astronomy Colloquia. USA*, 2011.

[20] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, pp. 28–31, IEEE, 2004.

[21] U. K. Kotikalapudi, "Abnormal event detection in video," *Tech Thesis, Indian Institute of Science, Banglore*, 2007.

[22] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for realtime tracking with shadow detection," 2001.

[23] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, "Action recognition with image based cnn features," *arXiv preprint arXiv:1512.03980*, 2015.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.