Deep Learning Approach for Suspicious Activity Detection from Surveillance Video

Amrutha C.V, C. Jyotsna, Amudha J.
Dept. of Computer Science & Engineering, Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, India
amruthacvvtk@gmail.com, c_jyotsna@blr.amrita.edu, j_amudha@blr.amrita.edu

Abstract— **Video Surveillance plays a pivotal role in today's world. The technologies have been advanced too much when artificial intelligence, machine learning and deep learning pitched into the system. Using above combinations, different systems are in place which helps to differentiate various suspicious behaviors from the live tracking of footages. The most unpredictable one is human behaviour and it is very difficult to find whether it is suspicious or normal. Deep learning approach is used to detect suspicious or normal activity in an academic environment, and which sends an alert message to the corresponding authority, in case of predicting a suspicious activity. Monitoring is often performed through consecutive frames which are extracted from the video. The entire framework is divided into two parts. In the first part, the features are computed from video frames and in second part, based on the obtained features classifier predict the class as suspicious or normal.**

Keywords—suspicious activity, video surveillance, deep learning.

I. Introduction

Human behavior recognition in the real-world environment finds plenty of applications including intelligent video surveillance, shopping behavior analysis. Video surveillance has vast application areas especially for indoor outdoor and places. Surveillance is an integral part of security. Today security camera becomes part of life for the safety and security purposes. E-surveillance is one of the main agendas in Digital India, development programme of Indian government. Video surveillance remains as a part of it. Advantages of video surveillance are effective monitoring, less manpower required, cost effective auditing capability, adopting new security trends etc. Currently, the tracking has been performed by human. Since we are dealing with huge amount of video data, this is easy to make people feel tired and the manual intervention will also produce omissions. It greatly affects the efficiency of the system. This has been solved by the automation of video surveillance. Today, manual monitoring of all the events on the CCTV (Closed Circuit Television) camera is impossible. Even if the event had already happened, searching manually the same event in the recorded video wastes a lot of time. Analyzing abnormal events from video is an emerging topic in the domain of automated video surveillance systems.

Human behavior detection in video surveillance system is an automated way of intelligently detecting any suspicious activity. Number of efficient algorithms is available for the automatic detection of human behavior in public areas like airports, railway stations, banks, offices, examination halls etc. Video surveillance is the emerging area in the application of Artificial Intelligence, Machine Learning and Deep Learning. Artificial intelligence helps the computer to think like human. In machine learning, important components are learning from the training data and make prediction on future data. Nowadays GPU (Graphics Processing Unit) processors and huge datasets are available, so the concept of deep learning is used.

The combination of computer vision and video surveillance will ensure public safety and security. Computer vision methods involves the following stages: modelling of environments, detection of motion, classification of moving objects, tracking, behavior understanding and description, and fusion of information from multiple cameras. This method requires lot of pre-processing to extract features in different video sequences. The classification techniques are supervised and unsupervised classification. Supervised classification uses manually labelled training data whereas unsupervised classification is fully computer operated and do not require any human intervention.

Deep Neural Networks is one of the best architecture used to perform difficult learning tasks. Deep Learning models automatically extract features and builds high level representation of image data. This is more generic because the process of feature extraction is fully automated. From the image pixels, convolutional neural network (CNN) can learn visual patterns directly. In the case of video stream, long short term memory (LSTM) models are capable of learning long term dependencies. LSTM network has the ability to remember things.

The proposed system will use footages obtained from CCTV camera for monitoring the human behavior in a campus and gently warn when any suspicious event occurs. The major components in intelligent video monitoring are event detection and human behavior recognition. Automatic understanding of human behavior is a challenging task. In a campus, different areas are under video surveillance and various activities are to be monitored. The video footage obtained from campus has been used for testing.

The entire process of training a surveillance system can be summarized in to three phases: data preparation, training the model and inference. The framework consists of two neural networks CNN and Recurrent Neural Network (RNN). CNN is used for the purpose of extracting high level features from the images so that the complexity of the input can be reduced. RNN is used for the classification purpose, which is well suited for processing of video stream. The proposed system is using a pre-trained model called VGG-16(Visual Geometry Group), which is trained on the ImageNet dataset. Currently, model is training in such a way to predict behavior from the footage. The model is able to predict suspicious or normal human behavior in the footage which is used to aid the monitoring process.

Most of the current system uses the footages obtained from CCTV cameras. If any crime or violence happens, this video will be used for investigation purpose. But if we consider a system which will automatically detects any unusual or abnormal situation in advance and a mechanism to alert the respective authority is more interesting and which can be applied to indoor and outdoor places. The proposed method is to design such a system in an academic area.

The paper is organized as follows: section II briefs the related works in the area of behavior analysis for detecting suspicious activities. Overall view of the proposed method is explained in section III. Implementation details are described in section IV followed by conclusion and future works in section V.

II. Literature Survey

The related works suggests different approaches for detecting human behaviors from video. The objective of the works was to detect any abnormal or suspicious events in a video surveillance.

Advance Motion Detection (AMD) algorithm was used to detect an unauthorized entry in a restricted area [1]. In the first phase, the object was detected using background subtraction and from frame sequences the object is extracted. The second phase was detection of suspicious activity. Advantage of the system was the algorithm works on real time video processing and its computational complexity was low. But the system was limited in terms of storage service and it can also be implemented with hi-tech mode of capturing of videos in the surveillance areas.

A semantic based approach was proposed in [2].The captured video data was processed and the foreground objects were identified using background subtraction. After subtraction, the objects are classified into living or non-living using Haar like algorithm. Objects tracking were done using Real-Time blob matching algorithm. Fire detection was also detected in this paper.

Based on the motion features between the object, suspicious activities were detected in [3]. Semantic approach was used to define suspicious events. The object detection and correlation technique was used to track objects [2]. The events are classified based on motion features and temporal information. The computational complexity of the given framework was less.

Abnormal events from a university were detected by divided into zones and estimated the optical flow in each zone using Lucas-Kanade method. Then they created the histogram of magnitude of optical flow vectors. Software algorithms are used for analyzing the content of a video to classify events as normal and abnormal [4].

A system was designed to distinguish the abnormal events from normal events based on the analysis of movement information from video sequences. HMM method was used to learn the histograms of optical flow orientations of the video frame. It compares the captured video frames with the existing normal frames and identified the similarity between these frames. The system was evaluated and validated on different datasets such as UMN dataset and PETS [5].

The unusual events in video footages could be detected by tracking of people. Human beings are detected from the video using background subtraction method. The features are extracted using CNN and which was fed to a DDBN (Discriminative Deep Belief Network). Labelled videos of some suspicious events are also fed to the DDBN and their features are also extracted. Then a comparison of features extracted using CNN and features extracted from the labelled sample video of classified suspicious actions was done using a DDBN and various suspicious activities are detected from the given video [6].

A real time violence detection system using deep learning was developed to prevent the violence behavior of crowd or players in sports. In a spark environment, frames were extracted from real-time videos. If the system detects any violence in football, then alert the security people. To prevent the violence in advance, the system detects the video actions in real time and alerts the security forces. VID dataset was used and achieved an accuracy of 94.5% for detecting violence in football stadium [7].

The abnormal event detection consists of different modules for the processing of video data. Deep architectures were used to detect human behavior. The proposed CNN and LSTM based models used UT Interaction dataset. One of the drawbacks of the system was similar human behaviors like pointing or punching is difficult to identify [8].

Understanding crowd behavior using a deep spatiotemporal approach classifies the videos into pedestrian future path prediction, destination estimation and holistic crowd behavior.es three different categories. Spatial information in the video frames was extracted using a convolutional layer. LSTM architecture was used learn or understand the sequence of temporal motion dynamics. Data sets used in the proposed system were PWPD, ETH, UCY and CUHK. The accuracy of the system can be improved by using deeper architectures [9].

Daily human activities were captured from videos and classification of those videos in to household, work related, caring and helping. Sports related are done through deep learning. CNN was used for retrieving input features and RNN for classification purpose. They used Inception v3 model and UCF101, Activitynet as datasets. The accuracy achieved was 85.9% on UCF101 and 45.9% on Activitynet [10].

A system was developed to monitor students' behavior in examination using neural network and Gaussian distribution. It consists of three different stages: face detection, suspicious state detection and anomalous detection. The trained model decides whether the student was in suspicious state or not and Gaussian distribution decides whether the student performs any anomalous behavior [11]. The accuracy achieved was 97%.

Intelligent video surveillance for crowd analysis was discussed in [12]. This was a review paper which covers relevance of video surveillance analysis in today's world, various deep learning models, algorithms and datasets used for video surveillance analysis.

The majority of papers mentioned above were done with the help of computer vision using various algorithms or by neural networks for detecting human behavior analysis from videos. Computer vision methods require lot of pre-processing to extract the trajectories or motion pattern to understand the evolution of features in a video sequence [13]. Also, background subtraction is based on static background hypothesis which is often not applicable in real time scenarios. In the real world, most of the issues occur in the crowd. Above discussed methods lacks efficiency while handling crowd. Based on the literature survey a deep architecture can be modelled for suspicious activity prediction using 2D CNN and LSTM, so the accuracy of the system can be improved. In deep learning approach, most of the papers detect only the suspicious activity. So an efficient mechanism is needed to alert the security in the case of any suspicious behavior.

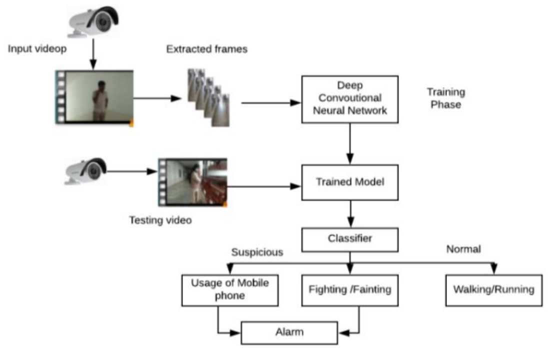
III. System Overview

The proposed system will use footages obtained from CCTV camera for monitoring students' activities in a campus and send message to the corresponding authority when any suspicious event occurs.

A. System Architecture

The architecture has different phases like video capture, video pre-processing, feature extraction, classification and prediction. The general layout of the system architecture is shown in Fig.1.The system classifies the videos into three classes.

1)Students using Mobile phone inside the campus- Suspicious class
2)Students fighting or fainting in campus-Suspicious class
3)Walking, running- Normal class

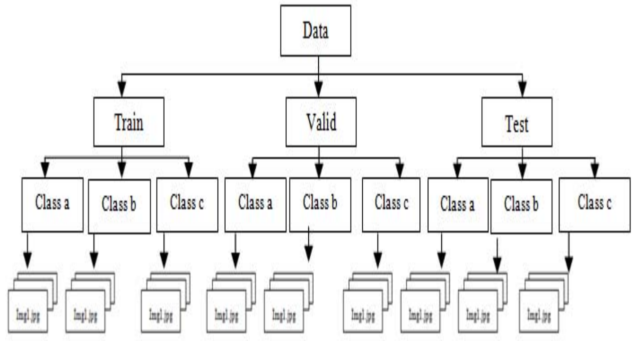


B. Video capture

Installation of CCTV camera and monitoring the footage is the initial step in video surveillance system. Various kinds of videos are captured from different cameras, covering the whole area of surveillance. The processing in our implementation is carried out using frames, so the videos are converted to frames.

C. Dataset Description

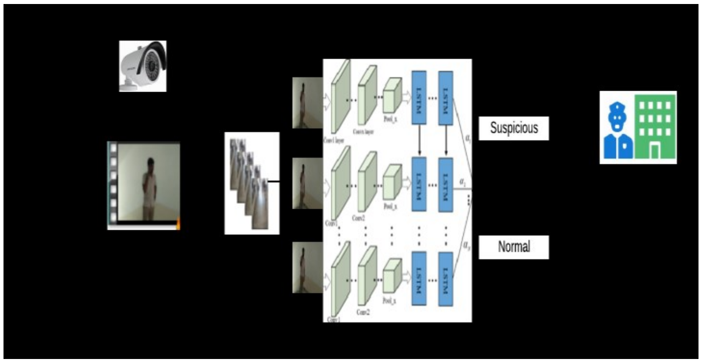
The KTH dataset is a standard dataset which has collection of sequences representing 6 actions and each action class has got 100 sequences. Each sequence has got almost 600 frames and the video is shot at 25 fps [14]. The model is trained on this dataset for normal behavior (running and walking). CAVIAR dataset, videos taken from campus and YouTube videos are used for training suspicious behavior (mobile phone using inside the campus, fighting and fainting). Around 7035 frames are extracted from different videos. The whole dataset is manually labelled and separated into 80% for training set and 20% for validation set. The directory structure of dataset is as shown in Fig.2. A combination of KTH, CAVIAR, YouTube videos and videos captured from campus are used in our system.



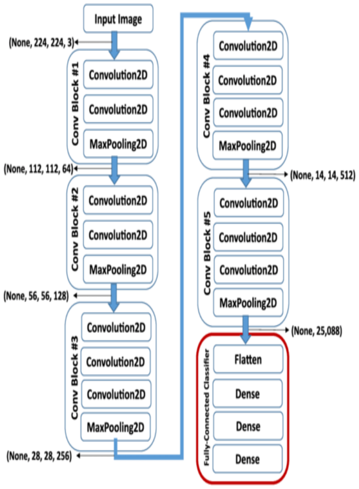
D. Video pre-processing

A deep learning network is using in our proposed system for suspicious activity detection from video surveillance. By deep learning architectures, the accuracy obtained can be higher and it also works better with large datasets. A detailed design overview is represented in Fig.3.

The input videos are taken from existing and created datasets. As part of pre-processing, frames are extracted from the captured videos. Based on the videos, three labelled folders are created and stored the frames in it. The entire video is converted to 7035 frames and the frames are stored as images in jpg format. Each frame is then resized to 224 ×224 to suite 2D CNN architecture and stored them. The testing video is also converted to frames and resized to 224 ×224 and stored in folder. OpenCV library in python is used for video pre- processing.



In image feature extraction, a pre-trained CNN model known as VGG-16 is trained on ImageNet dataset. VGG-16 architecture is shown in Fig.4. VGG-16 neural network [15] has convolution layers of size 3×3, max pooling layers of size 2×2 and fully connected layers at end, which makes a total of 16 layers was the deep learning architecture used here. The input image should be in the size 224×224×3 RGB form. Representations of the various layers which include convolution layers, ReLU (Rectified Linear Unit) layer i.e. activation function, max pooling layers, fully connected dense layers and normalization layers. The model can fine tune as per our requirement and the last layer of this model is removed. Then the model is trained on LSTM architecture. LSTM networks are a kind of RNN capable of learning order dependence in sequence prediction problems. We have ReLU activation function, dropout layer and fully connected dense layers. The count of neurons in the last layer is equal to the count of classes that we have and hence the number of neurons here is three.



The system classifies the videos as suspicious (students using mobile phone, fighting, fainting) or normal (walking, running). In the case of suspicious behavior, an SMS (Short Message Service) will be send to the respective authority.

IV. RESULT ANALYSIS

The aim of the project is monitoring the suspicious activities in a campus using CCTV footages and alerts the security when any suspicious event occurs. This was done by extracting features from the frames using CNN. After the extraction was done, LSTM architecture is used to classify the frames as suspicious or normal class. Fig.5 shows the Suspicious and Normal videos sequences.

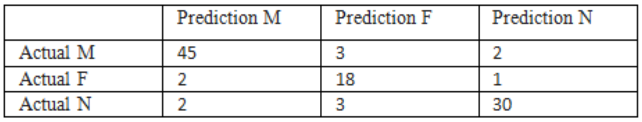
The steps for building the complete system are collect video sequences from CCTV footage, extraction of frames from videos, pre-processing of the images, and preparation of training and validation sets from the datasets, training and testing. In the case of suspicious activity, the system sends an SMS to the respective authority. The system has been developed in an open source platform using python. Sending of SMS is done by creating an account in Twilio and installed the twilio library in python. Twilio allows programmatically make and receive phone calls, send and receive text messages.

A. Training and Testing

The input videos are taken from CAVIAR dataset, KTH dataset, YouTube videos and videos taken from campus. Around 300 videos of different suspicious and normal behavior videos are collected. As part of pre-processing, frames are extracted from the captured videos. The pre-trained model used in our system is VGG-16 and take its learnings to solve our problem. The last layer of this model is removed based on our requirement and LSTM architecture is used for classification. Our dataset is trained on it. CCTV video footages of different scenarios are taken from our campus for testing and it is converted into frames. The stored frames are given to the trained model and finally the classifier classifies the video into suspicious or normal behavior.

B. Results

The accuracy of the training phase is 76% for the initial 10 epochs. The accuracy of the model can be improved by increasing the number of iterations. The frames are extracted from videos and stored in a single folder for the purpose of testing. Using our trained model, the system predicts the frames as suspicious (mobile phone using inside the campus, fighting or fainting) or normal (walking, running) class. In the case of suspicious activity, a message will be sent to the corresponding authority with the predicted class. The accuracy achieved is 87.15%. The confusion matrix is as shown in Table I.



	Prediction M	Prediction F	Prediction N
Actual M	45	3	2
Actual F	2	18	1
Actual N	2	3	30

V. Conclusion AND FUTURE work

In present world, almost all the people are aware of the importance of CCTV footages, but most of the cases these footages are being used for the investigation purposes after a crime/incident have been happened. The proposed model has the benefit of stopping the crime before it happens. The real time CCTV footages are being tracked and analyzed. The result of the analysis is a command to the respective authority to take an action if in case the result indicates an untoward incident is going to happen. Hence this can be stopped.

Even though the proposed system is limited to academic area, this can also be used to predict more suspicious behaviors at public or private places. The model can be used in any scenario where the training should be given with the suspicious activity suiting for that scenario. The model can be improved by identifying the suspicious individual from the suspicious activity.

References

[1] P.Bhagya Divya, S.Shalini, R.Deepa, Baddeli Sravya Reddy,"Inspection of suspicious human activity in the crowdsourced areas captured in surveillance cameras",International Research Journal of Engineering and Technology (IRJET), December 2017.

[2] Jitendra Musale,Akshata Gavhane, Liyakat Shaikh, Pournima Hagwane, Snehalata Tadge, "Suspicious Movement Detection and Tracking of Human Behavior and Object with Fire Detection using A Closed Circuit TV (CCTV) cameras ", International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue XII December 2017.

[3] U.M.Kamthe,C.G.Patil "Suspicious Activity Recognition in Video Surveillance System", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

[4] Zahraa Kain, Abir Youness, Ismail El Sayad, Samih Abdul-Nabi, Hussein Kassem, " Detecting Abnormal Events in University Areas ", International conference on Computer and Application,2018.

[5] Tian Wanga, Meina Qia, Yingjun Deng, Yi Zhouc, Huan Wangd, Qi Lyua, Hichem Snoussie, "Abnormal event detection based on analysis of movement information of video sequence" ,Article-Optik,vol-152,January-2018.

[6] Elizabeth Scaria, Aby Abahai T and Elizabeth Isaac, "Suspicious Activity Detection in Surveillance Video using Discriminative Deep Belief Netwok", International Journal of Control Theory and Applications Volume 10, Number 29 -2017.

[7] Dinesh Jackson Samuel R,Fenil E, Gunasekaran Manogaran, Vivekananda G.N, Thanjaivadivel T , Jeeva S , Ahilan A, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM",The International Journal of Computer and Telecommunications Networking,2019.

[8] Kwang-Eun Ko, Kwee-Bo Sim"Deep convolutional framework for abnormal behaviour detection in a smart surveillance system."Engineering Applications of Artificial Intelligence ,67 (2018).

[9] Yuke Li "A Deep Spatiotemporal Perspective for Understanding Crowd Behavior", IEEE Transactions on multimedia, Vol. 20, NO. 12, December 2018.

[10] Javier Abellan-Abenza, Alberto Garcia-Garcia, Sergiu Oprea, David Ivorra-Piqueres, Jose Garcia-Rodriguez "Classifying Behaviours in Videos with Recurrent Neural Networks", International Journal of Computer Vision and Image Processing,December 2017.

[11] Asma Al Ibrahim, Gibrael Abosamra, Mohamed Dahab "Real-Time Anomalous Behavior Detection of Students in Examination Rooms Using Neural Networks and Gaussian Distribution", International Journal of Scientific and Engineering Research, October 2018.

[12] G. Sreenu and M. A. Saleem Durai "Intelligent video surveillance: a review through deep learning techniques for crowd analysis" , Journal Big Data ,2019.

[13] Radha D. and Amudha, J., "Detection of Unauthorized Human Entity in Surveillance Video", International Journal of Engineering and Technology (IJET), 2013.

[14] K. Kavikuil and Amudha, J., "Leveraging deep learning for anomaly detection in video surveillance", Advances in Intelligent Systems and Computing,2019.

[15] Sudarshana Tamuly, C. Jyotsna, Amudha J, "Deep Learning Model for Image Classification", International Conference on Computational Vision and Bio Inspired Computing (ICCVBIC),2019.