AI Suspicious Activity Detection using Human Pose Estimation

[1] YATHESHVAMSI NAIDU K [2] HARIBABU P
[1] Student, Dept of CSE, Raghu Institute of Technology, Vizianagaram,A.P, India
[2] Assistant Professor, Dept of CSE, Raghu Institute of Technology, Vizianagaram,A.P, India
Email id: yatheshvamsinaidu@gmail.com, haribabupogiri@gmail.com

ABSTRACT: Video surveillance plays a central role in today's world. Technologies have become too advanced as artificial intelligence, machine learning, and deep learning invaded the system. Using the above combinations, various systems are in place to help distinguish various suspicious behavior from live tracking images. Human behavior is the most unpredictable and it is very difficult to determine whether it is suspicious or normal. The deep learning approach is used to detect suspicious or normal activity in an academic environment and send an alert message to the appropriate authority when suspicious activity is predicted. Tracking is often done through successive frames extracted from the video. The whole frame is divided into two parts. In the first part, the characteristics are calculated based on the video frames, and in the second part, based on the obtained characteristic classifier, the class is predicted as suspicious or normal.

KEYWORDS: DeepLearning, Activity detection, Tensorflow, Keras, Python, Opencv

I. INTRODUCTION

Numerous real-world environments have applications for human behaviour identification, such as intelligent video surveillance and purchasing behaviour analysis. There are many uses for video surveillance, particularly indoor, outdoor, and public spaces. Security includes surveillance as a crucial component. In today's world, security cameras are a must for both safety and security. One of the key goals of the Indian government's development programme, Digital India, is e-surveillance. Video surveillance is still a part of it. The benefits of video surveillance include effective monitoring, less manpower required, cost-effective auditing capability, and the ability to adopt new security trends. The tracking is currently being done by hand. Because we are dealing with a large amount of video data, it is easy for people to become tired, and manual intervention will result in omissions. It has a significant impact on the system's efficiency. The automation of video surveillance has solved this problem. It is now impossible to manually monitor all events on the CCTV (Closed Circuit Television) camera. Even if the event had already occurred, manually searching for the same event in the recorded video wastes a significant amount of time. The analysis of abnormal events from video is a new topic in the field of automated video surveillance systems.

Human behavior detection in video surveillance systems is an automated method for intelligently detecting suspicious activity. A number of efficient algorithms are available for automatically recognizing human behavior in public areas such as airports, train stations, banks, offices, and exam halls. Video surveillance is an emerging field in the application of artificial intelligence, machine learning and deep learning. Artificial intelligence helps computers think like humans. In machine learning, a key component is learning from training data and making predictions about future data. With the recent availability of GPU processors (Graphics Processing Units) and huge data sets, the concept of deep learning is used.

The use of computer vision in conjunction with video surveillance will ensure public safety and security. The following stages are involved in computer vision methods: modelling of environments, motion detection, and classification of moving objects, tracking, behaviour understanding and description, and fusion of information from multiple cameras. This method necessitates extensive pre-processing in order to extract features from various video sequences. There are two types of classification techniques: supervised classification and unsupervised classification. Supervised classification makes use of manually labeled training data, whereas unsupervised classification is entirely computer-driven and requires no human intervention.

Deep Neural Networks is one of the stylish armatures used to perform delicate literacy tasks. Deep literacy models automatically prize features and builds high position representation of image data. This is more general because the process of point birth is completely automated. From the image pixels, convolutional neural network

(CNN) can learn visual patterns directly. In the case of videotape sluice, long short-term memory (LSTM) models are able of learning long term dependences. LSTM network has the capability to flash back effects.

The proposed system will use CCTV camera footage to monitor human behavior on campus and gently warn when any suspicious event occurs. Intelligent video monitoring relies heavily on event detection and human behaviour recognition. Understanding human behavior automatically is a difficult task. Different areas on a campus are under video surveillance, and various activities are to be monitored. The video footage obtained from campus was used for testing purposes.

The training of a surveillance system can be divided into three stages: data preparation, model training, and inference. The framework is made up of two applications in smart environments: video surveillance, human-robot interactions, and ambientally supported life, all of which involve the problem of learning agent patterns from sensor data. Deviating is a pattern in the data that either does not match the expected line, resulting in an unusual line, or corresponds to a previously defined non-green, indicating suspicious wear. The goal of the thesis is to identify samples that could be harmful to human health, safety, or other interests according to the agent's state. We encounter many difficulties when using real-world applications. While many real-world applications involve the exercise of detecting actions from unprocessed sensor data, research on the development of plan recognition has relied on the assumption that the agent's fundamental activities are known or can be easily retrieved. The second issue is how to depict a complicated, unstructured design of people who do not behave in predictable ways.

Recognizing activities is a fundamental task in behaviour analysis. The sensor data is transformed into a higher-level description of behavioural primitives. The pipeline for extracting an agent's atomic activities from multidimensional, sequential, spatiotemporal data is presented in this chapter. We first define the issue and talk about the fundamental concepts, then we describe the pipeline's elements, such as noise filtering, feature vector design, model learning, and recognised activity continuity. We also provide methods for identifying compound activities and those that come about as a result of interactions between agents.

II. LITERATURE SURVEY

Related studies have proposed various approaches to detect human actions from videos. The purpose of this work was to identify conspicuous or suspicious incidents during video surveillance.

An unauthorized entrance into a restricted location was discovered using the advance motion detection (AMD) technique. In the first stage, objects were recognized by background subtraction and objects were extracted from the frame sequence. The second phase is suspicious activity detection. The advantage of this system was that the algorithms worked with real-time video processing and had low computational complexity. However, the system is limited in terms of storage services and can also implement high-tech modes for recording video in surveillance areas.

A semantics-based approach was proposed in. The captured video data were processed and foreground objects were identified by background subtraction. After subtraction, a Haar-like algorithm is used to classify the object as alive or dead. Object tracking was done using a real-time blob matching algorithm. Fire detection was also acknowledged in this paper.

Suspicious activity was detected based on movement characteristics between objects using a semantic approach to define suspicious events. Object detection and correlation techniques were used for object tracking. Events are classified based on motion characteristics and time information. The computational complexity of the given framework was low.

Anomalous events in a university are detected by dividing them into regions and estimating optical flow in each area using the Lucas-Kanade method. Next, they generated intensity histograms of the optical flow vectors. Software algorithms are used to analyze video content to classify events as normal and abnormal.

A system designed to distinguish unusual events from normal events based on analysis of motion information from video footage. The HMM method is used to learn the histogram of the optical flow directions of a video frame. It compares captured video images with existing normal images and determines the similarity between these images. The system has been evaluated and validated on various data sets such as UMN and PETS datasets. Observers can detect unusual events in the footage. Humans are detected from the video by background subtraction. Features were extracted using CNN and passed to DDBN (Discriminatory Deep Belief Network). Tagged videos of certain suspicious events are also transmitted to the DDBN and their characteristics are also extracted. Then the comparison of features extracted using CNN and features extracted from tagged video sample of classified suspicious actions performed using DDBN and other suspicious activities each other is detected from the given video.

A real-time violence detection system using deep learning has been developed to prevent violent behavior

by crowds or players in sports. In a sparking environment, images are extracted from real-time video. If the system detects football violence, notify the security guard. To prevent violence upstream, the system detects video actions in real time and alerts security forces. The VID dataset was used and achieved 94.5% accuracy for detecting violence in football stadiums.

Abnormal event detection includes different modules for processing video data. Deep architectures have been used to detect human behavior. The proposed models based on CNN and LSTM used the UT interaction dataset. One of the system's limitations is that it is difficult to identify similar human behaviors, such as pointing or hitting. Understanding crowd behavior using a deep space-time approach classifies videos into future pedestrian prediction, destination estimation, and overall crowd behavior into three other categories. together. Spatial information in the video image is extracted using a composite layer. The LSTM architecture was used to learn or understand motion dynamics series over time. The data sets used in the proposed system are PWPD, ETH, UCY and CUHK. System accuracy can be improved by using deeper architectures [9]. Daily human activities are recorded from videos and categorize these videos into family, work, care and help. Link sports are made possible through deep learning. CNN was used to take input and RNN features for classification purposes. They used Inception v3 and UCF101 models, Activity net as the data set. Accuracy obtained was 85.9% on UCF101 and 45.9% on Active network.

A system has been developed to monitor students' behavior during exams using a neural network and a Gaussian distribution. It consists of three different phases: Face detection, suspicious detection and anomaly detection. The trained model decides if the student is in a suspicious state, and the Gaussian distribution decides if the student behaves abnormally. The obtained accuracy is 97%. Intelligent video surveillance for crowd analysis was discussed. This is a review article that addresses the relevance of CCTV analysis in today's world, the various deep learning models, algorithms and datasets used for CCTV analysis and video surveillance.

Most of the above works were created with the help of computer vision using various algorithms or via neural networks detecting human behavior analysis from videos. Computer vision methods require a lot of preprocessing to extract trajectories or motion patterns in order to understand the evolution of features in video sequences. Moreover, background subtraction is based on a static background hypothesis, which is often not applicable to real-time scenarios. In the real world, most problems occur in crowds. The above method is not efficient when dealing with crowds. Based on a literature review, we can model a deep architecture for suspicious activity prediction using 2D-CNN and LSTM, thus improving the accuracy of the system. Due to the deep learning approach, most papers only detect suspicious activity. Therefore, we need an efficient mechanism to alert security in the event of suspicious behavior. EHRs are complex because they can contain structured and unstructured data. The latter example is textual information, which may require natural language processing techniques for processing. Additionally, categorical data, such as diagnostic coding, may adopt different coding systems across different institutions.

III.PIPELINE FLOW

The architecture has various phases such as video ingestion, video preprocessing, feature extraction, classification, and prediction. The general layout of the system architecture is shown in Figure 1. The system classifies videos into three classes.

1) Students Using Mobile Phones on Campus - Suspicious Classes
2) Students fight or pass out in Campus Suspect Class 3

(i) Walk, run - normal class video recording

Installing CCTV cameras to monitor footage is the first step in a video surveillance system. Different types of videos are recorded by different cameras to cover the entire surveillance area. The processing in our implementation is done in frames, so the video is converted to frames.

(ii) Dataset Description

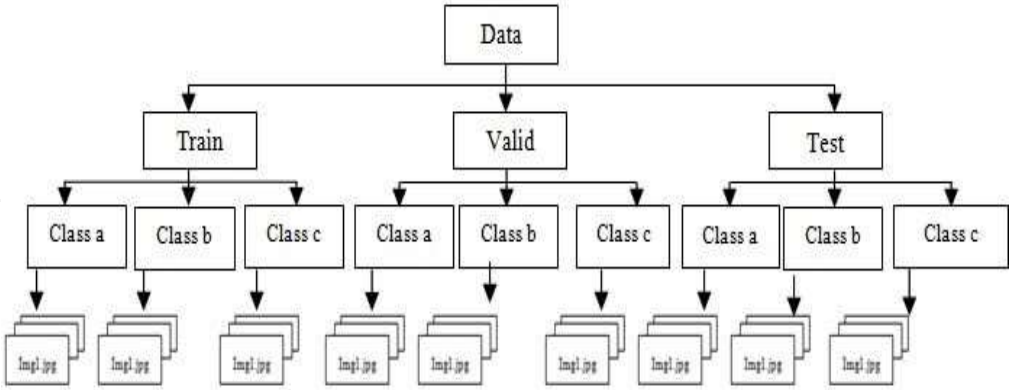
The KTH dataset is a standard dataset containing a collection of sequences representing 6 actions, with 100 sequences for each action class. Each sequence has almost 600 frames and the video is recorded at 25 fps [14]. The model is trained on this data set for normal behaviors (running and walking). Train suspicious behaviors (mobile phone use, fights, fainting on campus) using CAVIAR datasets, campus videos, and YouTube videos. About 7035 frames are extracted from various videos. The entire dataset is manually labeled and split into 80% of the training set and 20% of the validation set. Your record directory structure should look like Figure 2. A combination of KTH, CAVIAR, YouTube videos and videos recorded from the campus are used in our system.

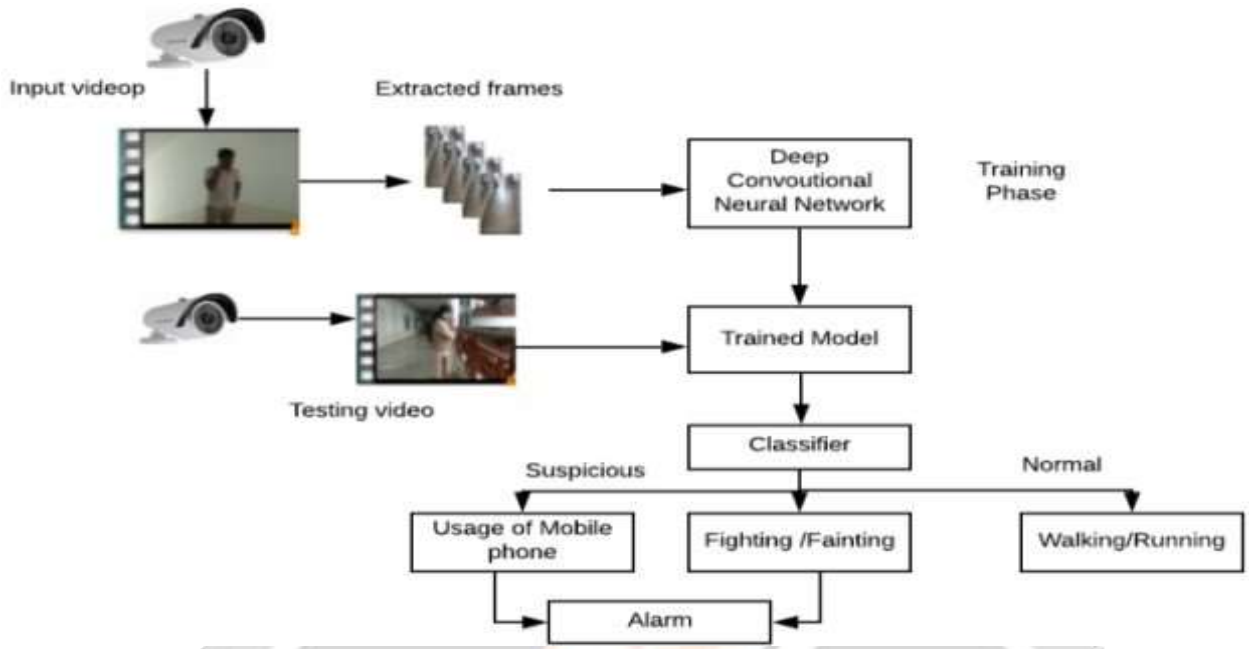
(iii) Video pre-processing

A deep learning network is used in our proposed system to detect suspicious activity from CCTV. Thanks to the deep learning architecture, the accuracy achieved can be higher and it also works better with large data sets. A detailed overview of the design is shown in Figure 3.

The input video is taken from existing and generated datasets. As part of pre-processing, images are extracted from recorded videos. Based on the videos, three labeled folders are created and images are stored in them. The entire video is converted to 7035 frames and the frames are stored as jpg format images. Then each image was resized to 224 × 224 to fit the 2D CNN architecture and stored them. The test video was also converted to frames and resized to 224 × 224 and stored in a folder. The OpenCV library in python is used for video preprocessing.

In image feature extraction, a pre-trained CNN model called VGG-16 is trained on the ImageNet dataset. The VGG-16 architecture is shown in Figure 4. The VGG-16 neural network has convolutional layers with size 3 × 3, maximum pooling layers of size 2 × 2, and connected layers fully at the end, making a total of 16 layers in the neural network used. here. The input image should be in RGB 224 × 224 × 3 format. The representation of different layers includes convolutional layers, ReLU (rectified linear unit) layer i.e. activation function, clustering layers max, fully connected dense layers and normalized layers. The model can be tailored to our needs and the last layer of this model is removed. Then the model is trained on the LSTM architecture. LSTM network is a type of RNN capable of learning order dependence in sequence prediction problems. We have ReLU enabled, dropout, and fully connected dense layers. The number of neurons in the last layer is equal to the number of layers we have and so the number of neurons here is three.





IV. RESULT ANALYSIS

The goal of this project is to use CCTV footage to monitor suspicious activity on campus and alert security when suspicious events occur. This was done by extracting features from the frames using CNN. Once the extraction is complete, we use the LSTM architecture to classify frames as suspicious or normal classes. Figure 5 shows a suspect video sequence and a normal video sequence.

The steps to build a complete system are acquisition of a video sequence from CCTV footage, extraction of frames from the video, image preprocessing, preparation of training and validation sets from the dataset, training and testing. In the event of suspicious activity, the system will send her SMS to the responsible agency. This system was developed on an open source platform using Python. To send SMS, create an account with Twilio and install the Twilio library in Python. Twilio allows you to programmatically make and receive phone calls, send and receive text messages.

(iv)Training and Testing

Input videos come from the CAVIAR dataset, KTH dataset, YouTube videos, and campus videos. We have collected about 300 videos of various suspicious and normal behaviors. As part of preprocessing, frames are extracted from the recorded video. The pre-trained model used in the system is VGG-16 and we use its insights to solve the problem. The last layer of this model is removed based on requirements and LSTM architecture is used for classification. Our dataset is trained on this.

CCTV video footage of various scenarios was taken from our premises for testing and converted into footage. The stored images are fed to the trained model and finally the classifier classifies the video as suspicious or normal behavior.

V. RESULTS

The training phase accuracy was 76% for the first 10 epochs. Model accuracy can be improved by increasing the number of iterations. Images are extracted from the video and stored in a single folder for testing purposes. Using our trained model, the system predicts whether the frame is suspicious (cell phone used on campus, fighting or fainting) or normal (walking, running). In the event of suspicious activity, a notification will be sent to the authority corresponding to the predicted class. The obtained accuracy is 87.15%. The confusion matrix is presented in Table I.

VI. CONCLUSION AND FUTURE WORK

In today's world, most people are aware of the importance of CCTV footage, but most of the cases, CCTV footage is used for investigative purposes after the crime/incident has occurred. . The proposed model has the advantage of stopping crime before it happens. Real-time CCTV footage is monitored and analyzed. The result of the analysis is an order for the respective competent authority to take action if, in the event that the results indicate that an untoward incident will occur. So it can be stopped.

Although the recommendation system is limited to academia, it can also be used to predict more suspicious behaviors in public or private places. The model can be used in any situation where training with suspicious activities is appropriate for that situation. The model can be improved by identifying suspicious individuals from suspicious activity.

Detecting unusual and suspicious behaviors becomes more difficult as triggers are observed over a longer period of time. In many fields, no single event is sufficient to identify deviant behavior. Instead, it is recommended to combine multiple reviews. This is in contrast to previous chapters, which focused on detecting a single suspicious event. This chapter proposes a two-step detection system that first detects trigger events in behavioral traces and then combines the evidence to provide a level of suspicion. This chapter specifies the conditions that any reasonable detector must meet, analyzes the three detectors, and proposes a new detector that generalizes utility-based plane recognition with arbitrary utility functions.

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

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

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

5.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.

6.K. Kavikuil and Amudha, J., "Leveraging deep learning for anomaly detection in video surveillance", Advances in

Intelligent Systems and Computing,2019.

7.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.

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.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.

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.