Unusual Activity & Weapon Detection in an ATM using Deep Learning

Shiva Kumar R Naik, Rahul S, Ravi Teja S, R L Varun, Samadhi Pavan Kumar,
School of Computer Science and Engineering, REVA University

Abstract– *As the world is advancing, we've started using automation in almost every aspect of our lives. Also, as we're growing in the 21st century, the real world is getting a little more advanced, and advancement in technology and other things brings drawbacks and some negative impacts on society. ATMs – Automated Teller Machines – are one of the most common locations for criminals to attack people for money or simply steal money from an ATM. Nowadays, almost all banks have started keeping security guards at ATMs, but it's not so cost-effective to keep them there 24 hours a day, and sometimes human error does occur, as a result of which we come across situations where even security guards are injured by criminals trying to loot the ATM. Even installing CCTV cameras has not been able to minimise such incidents.*

In such situations, there is a demand for an automated system that can alert the authorised people regarding any unusual activity inside the ATM room. Taking video as input and taking 3 major points in reference: a. wearing a helmet; b. use of a knife; c. detection of gun system can alert authorities.

Key Words Deep *Learning, Convolutional Neural Network, CCTV- Closed Circuit TV, Safety, Automation.*

I. INTRODUCTION

An unusual activity and weapon detection system for ATMs is a system that finds out if any suspicious activity is happening at the ATM or if anyone has entered with a gun or knife. It helps save precious lives or valuable money for banks or individuals. Many banks depend on CCTV cameras, but there will always be human errors, as humans tend to lose focus after a while, and it's not possible to monitor CCTV footage each moment. With this system, we're trying to solve one of the biggest problems banks are suffering from: not being able to monitor ATMs 24/7. Banks have to recruit security guards, but that is not cost-effective, and it doesn't solve the problem.

Visual/optical surveillance is a well-known field of study with a wide range of applications in human unordinary monitoring systems, public safety in places like banks, shopping malls, and private areas, automated event detection, motion-based identification, human tallying, actuality, mobile robot navigation, and other fields. Rapid advancements in the availability of high-quality, low-cost video recording equipment, supercomputers, and growing demand for analysis of such tapes have sparked widespread interest in and demand for video surveillance in almost every industry.Finding out mobile objects and then identifying those at movies, on the other angle, is very significant and critical. Separating things from the background, on the other hand, is a challenging but necessary task.

As a result, it's critical to comprehend the video's content as well as the context of the items. Items derived from those other background objects have become a significant issue. As anoutcome, interpreting the film and its factors with the shown outlines becomes a very important criteria. The predictable goal of the unforeseen activity detection process is to identify a typical human behaviour strategy. The method is designed at the beginning with respect to a standard dataset of some activity. The useful facts and facts are then matched to the format during confirmation. In the end, it is concluded whether the action is anticipated or not. The demand for a designated regular human activity strategy makes unusual activity detection challenging in real-world security systems.

II. Literature Survey

In this section, we talk about the systems or papers that have been implemented before, which would help us to get the optimum result in the end. In this paper, researchers proposed a unique way of finding unusual human activity in crowded places. Clearly, instead of finding out or subdividing a person, they suggested or proposed a very effective method, what they called a "motion influence map," to constitute a person's activities. The major characteristics of the put-forward motion influence map are that it constructively shows the motion or shifting characteristics of the motion pace, motion aspect, and dimension of the material or

object and those' interactivities inside a frame pattern. By applying the put forward motion influence map, scholars in-depth proposed a simple outline in which they would be able to identify both regional as well as global unusual activities [1].

In a research paper, developers focused on analysing two states that, if disregarded, may put human lives at high risk. Among them are detecting prospective gun-based offences or violations and identifying deserted luggage on frames of observation footage. Developers presented a deep neural network model that can immediately find guns in images. Developers also presented a machine learning as well as a computer vision-based pipeline which would detect deserted baggage so that we could identify prospective gun-based offences and deserted baggage conditions in monitoring footage [2].

Overcrowded visuals pose new challenges to livestream assessment that conventional approaches cannot address. In this paper, researchers present a novel statistics based working outline for modelling the regional spatio-temporal movement sequence nature of overly crowded places. Their key perception is to make the most of the heavily dense interest of the overcrowded visuals by designing the major movement sequences in regional regions, thus constructively holding the fundamental innate constitution they appear in on the screen. In simple terms, they modelled the motion difference between regional space-time capacity and their spatial-temporal demographical nature to distinguish the overall nature of the given moment in any video. They demonstrated that by capturing the stable state movement nature with the indicated spatio-temporal movement pattern representation, we could easily find an unusual activity such as statistical or demographical divergence. Their experiment demonstrates that regional spatio-temporal movement trend representation yields pledging real results with activities that are difficult to perform and that are difficult to analyse even for human speculation[3].

The challenges in detecting new-born are unique and challenging. In this paper, scholars identified a major problem of jaw contour being less distinct in this particular case. They proposed a multi-label CNN based system which detects facial action unit usual habits of new-born. They used an extension of FACS for new-born. They tried to give an alternative to manual coding which isautomatic AU detection to find new-born expressions [4].

III. OBJECTIVES:

a) To design a model for implementing the abnormality detection algorithm to detect unusual activities.
b) To record the video from the camera when any unusual activity is detected while the absence of the operator.
c) To provide an instant alert of video containing the unusual activity.

IV. METHODOLOGY

A. System Design and Architecture:

Modern society's biggest security problem has a solution now. The system is one step towards safeguarding precious lives and hard-earned money.
The proposed system is based on CNN and Deep learning.

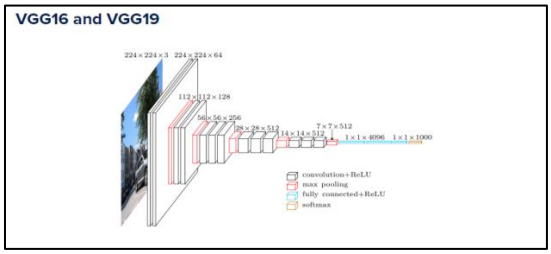

Fig1. Visualization of VGG Architecture

Every layer of a CNN puts in a well-defined set of filtrations, generally, thousands or multitudes of them, after which it puts together the results before passing the outcome into the next upcoming layer of the network. While training, CNN tries to learn the value systems for these filtrations instantly.
In terms of image classification, our CNN might learn to:

- In the first layer, the image is determined from raw pixel data.
- In the second layer, use these corners to identify shapes (i.e., "blobs").

- In the network's highest layers, use these forms for inferential analysis characteristics such as facial characteristics, car parts, and so on.

The final layer of a CNN employs these higher-level features that make forecasts about the image's contents.
An (image) convolution in deep learning is an element-wise linear combination of two matrices accompanied by a sum.

1. Consider two matrices (both of themhavingthe same dimension).
2. Multiply them element by element, (not by the dot product, but by a simple multiplication).
3. Add all the elements together.

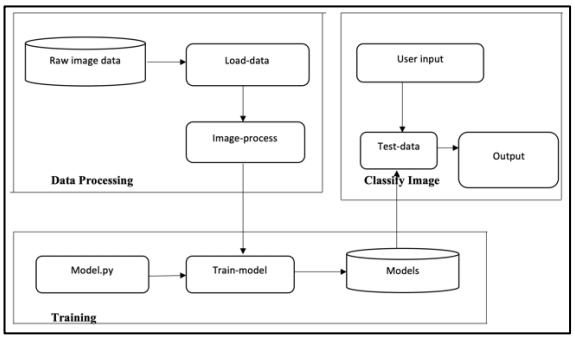

Fig2. System Architecture

B. DATAPROCESSING

Data processing refers to the process of transforming data from one form to a much more usable and preferred shape, i.e. making it much more relevant and instructive. This entire process could be automated with the help of machine learning methodologies, computational analysis, and statistical data. Based on the job at hand and the machine's requirements, this entire process can produce charts, films, diagrams, tables, pictures, and a variety of other file types.

a. Raw Image Data: We collect image data/datasets for training and testing purposes.
b. Input: Now that the data has been prepared, it may be in a format that is not machine-readable, so some conversion methods are required to transform it into a readable format. Excellent calculation and precision are required to complete this task. Data can be gathered from a variety of sources, including MNIST Digit data (pictures), Facebook posts, music recordings, and video clips.
c. Processing: This is the phase when algorithms and machine learning approaches are needed to accurately and efficiently execute the commands given over a vast volume of data.
d. Output: At this level, the machine obtains results in a substantial way that the user may easily deduce.Reports, graphs, movies, and other forms of output are possible.
e. Training: We can now use the photos in our training set to train our network.The purpose of this task for our infrastructure is to identify how to recognise every one of the groups in our data sets. If the model shows an error, it learns from it and tries to do better the next time.
f. Evaluate: Finally, we should assess the performance of our perfectly-trained squad.This project demonstratesevery oneof the photographs in our trial sample to the system and communicates it to assume whatever the picture’s tag is. The model's forecast for a photograph in the trial or test data isfinally calculated.

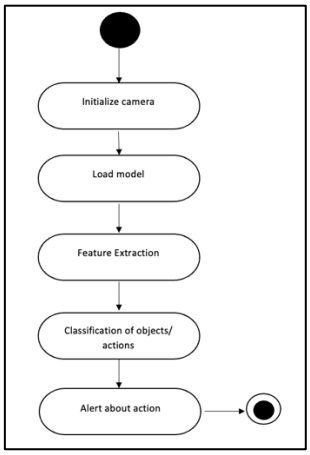

Fig3. Activity Diagram

C. PROPOSED SYSTEM

Using CNN, the proposed method can distinguish human movementsin anovercrowded situation and determine if the activity or movements is normal or odd. We'll create a deep teaching method to classify the dataset's behaviour. A picture is fed into the system. It categorizes this data into one of several categories (usual and unusual s). An ensemble of CNNs, as well as image preparation processes and neural networks (NNs) which mix the image characteristics from the CNNs along with the image, construct the system. The ensemble uses unweighted averaging to combine the NNs' outputs into a set of forecast possibilities for the classes. The classification is based on the maximum likelihood.

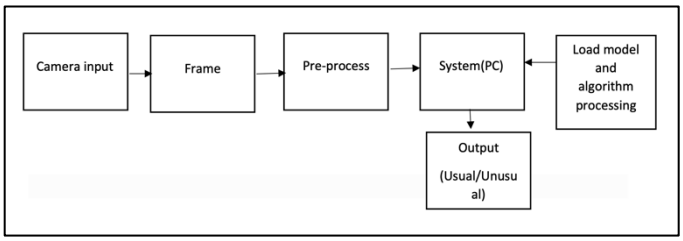

Fig4. Block Diagram

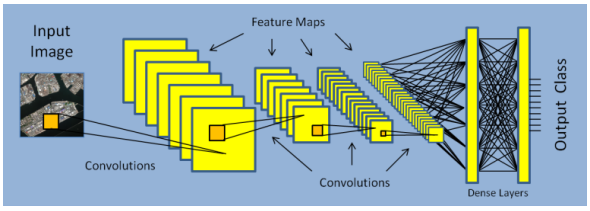

Fig5. The structure of aCNN. The input image is passed via a set of image feature detectors.

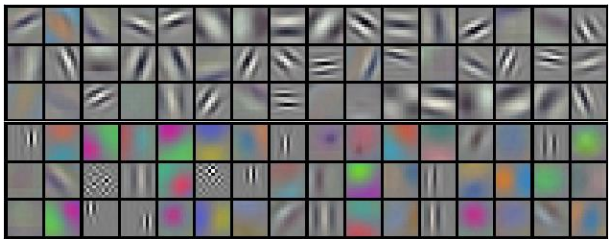

Fig6. The above image shows the Example of the image feature detectors that a CNN might "learn" during its training.

As shown in Fig. 5, a CNN is made up of multiple processing layers. Each layer is made up of a group of convolutional filters that identify visual details. The Gabor-like and colour blob filters depicted in Fig. 6 are feature detectors in the early stages.

Higher-level feature detectors are formed when layers are added. The CNN mixes the sensor outputs in fully linked "dense" layers near the conclusion of the series, satisfaction was found in a set of possibilities, one for each class. CNNs, dissimilar toprior methods such as SIFT and HOG, do not necessitate feature detectors being designed by the algorithm designer. As it trains, the network learns which characteristics to look for and how to recognize them.

V. IMPLEMENTATION

A. IMAGE FUNDAMENTALS

Pixels: The Building Blocks of Images

Pixels are indeed the basic building blocks of an image. Each image is composed of pixel resolution. The pixel is the most fine-grained granularity obtainable. A pixel is commonly regarded as the "colour" or "intensity" of illumination that would seem in a particular location in our image. When regarded as a grid, every square in an image contains a single pixel. Take a look at Figure 7 for an example. Figure 7 depicts a picture with a reliability or pixel density of 2500, which also means it is 500 pixels wide by 500 pixels tall.A picture can be thought of as a (multidimensional) matrix. Our matrix contains 500 columns (width) and 500 rows in this example (height). In total, our image contains 500 X 500 = 2500 pixels.
Most pixels are denoted or understood in the following ways:

a. Grayscale/single channel
b. Colour

In any grayscale picture, each pixel has a numeric value which falls between 0 and 255, with zero relating to "black" and 255 relating to "white." There are several shades of grey between 0 and 255, with countnearer to zero becoming darker and countnearer to 255 being lighter. Picture7 depicts a grayscale gradient picture, with darker pixel resolution on the left side and shinier pixel resolution on the right side. Colourful pixels, on either side, are normally represented using the RGB colour space.

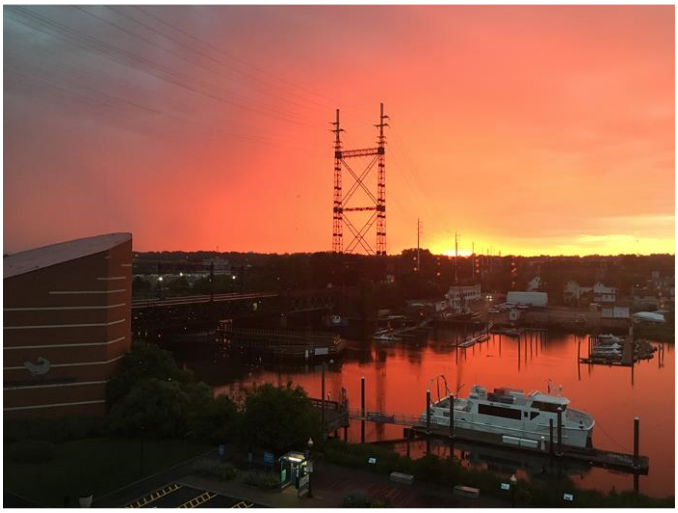

Fig6.Image having 1000px width and 750px height makes total of 75,000 Pixels(Image for demonstration only, actual size may vary)

Fig7. Image gradient showing pixel values going from very dark: black (0) to lighter: white (255).

B. IMPORT VIDEO AND INITIALISE FOREGROUND DETECTOR

The goal of mobile person identification is to separate people-related regions via the remainder of an image pattern. For motion segmentation, background prediction is an especially well-liked method. Even though the background is not stationary, the background scene modelling is analytically learnt utilizing the redundancy of image intensity in a training stage. Each pixel's redundancy information is saved individually in a history map that depicts the intensity variations at different pixel locations. Instead of analyzing the whole video right away, the specimen begins by collecting abeginning frame buffer in which the

movable things are separated from the backdrop.This aids in the progressive introduction of the video processing stages. To initialize the Gaussian mixture model, the foreground identifier needs a fixed amount of clip frames. The first 50 frames in this example are used to begin three Gaussian modes in the probabilistic model. Following the learning, the identifier starts to produce increasingly accurate segmentation results.

C. PROCESS THE REST OF THE VIDEO FRAMES

The leftover video frames are processed in the last stage. The differential approach for motion estimation is a frequently used method in computer vision. It works by assuming that the flow in a local neighbourhood of the pixel under examination is largely constant and then solving the fundamental optical flow model for every single pixel in that localityutilizingthe basis of the minimum square. The approaches can typically overcome the inherent uncertainty of the motion equation by merging the data from many neighbouring pixels. It's also less affected by image noise than point-based approaches. However, because it is a strictly local approach, it is unable to offer flow information inside uniform parts of the image.

D. ACTION RECOGNIZATION

The purpose of activity recognition is to analyze ongoing occurrences from video/live data in an automatic manner. This part of the project recognizes actions depending on objects, preferableto a human model, and successfully identifies several fundamental actions. Wearing a helmet, carrying a knife and a rifle, and so on.

VI. OUTCOMES

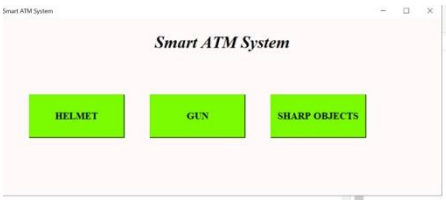

Fig8. GUI

Fig9. Knife Detected by Camera

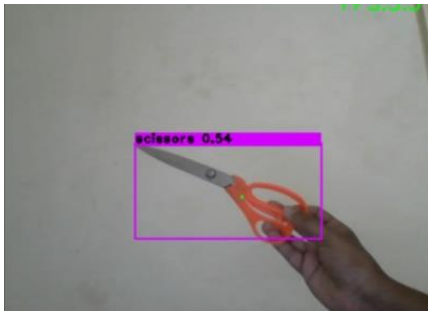

Fig10.A sharp object like Scissors Detected by the camera

Fig11. Helmet Detected by Camera

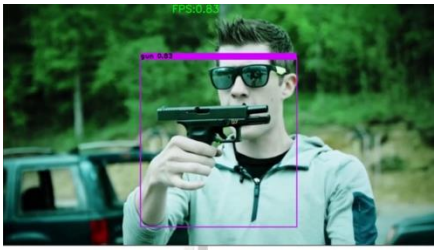

Fig12. Gun detected by Camera

VII. CONCLUSION

In our daily life when we see human-to-human interactivity andmutual interactions, recognizing unusual conduct is critical. Because it comprises information about just a person's identity, character, and psychological condition, it is difficult to draw out.

One of the main topics of research in the science topics of machine learning and computer vision is the human ability to recognize other people's behaviours. As a result, multiple activity recognition systemsare required for many multimedia applications such as surveillance systems, human contact, and robotic systems for human behaviour portrayal.

Suspicious behaviour in public spaces is harmful and can have major ramifications. Different methods are based on the acquisition of video sequences that detects motion or pedestrians, but computers are not clever enough to identify suspicious conduct in thereal world.

Thus, a system built is helpful in real-life scenarios. Large scale implementation is easy and low cost since all the ATMs will have CCTV connected already. Recognition of unusual activitywith a weapon or sharp object detection from live video input along with an automated messaging system makes this unique from existing ones.

REFERENCES

[1] D. Lee, H. Suk, S. Park and S. Lee, "Motion Influence Map for Unusual Human Activity Detection and Localization in Crowded Authorized licensed use limited to: Cornell UniversityLibrary. Downloaded on August 22,2020 at 13:47:08 UTC from IEEE Xplore. Restrictions apply. Scenes," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 10, pp. 1612-1623, Oct. 2015.

[2] Sathyajit Loganathan,GayashanKariyawasam, Prasanna Sumathipala,"Suspicious Activity Detection in Surveillance Footage" in IEEE International Conference add in 2020, Electronic ISBN 978-1-7281-553.-6.

[3] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1446-1453, Jun. 2009.

[4] Zakia Hammal1, Wen-Sheng Chu1, Jeffrey F. Cohn1,2, Carrie Heike3, and Matthew L. Speltz4, "Automatic Action Unit Detection in Infants Using Convolutional Neural Network", 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE.

[5] Mrs. Varsha ShrirangNanaware, Dr. Mohan Harihar Nerkar and Dr. C.M. Patil, "A Review of the Detection Methodologies of Multiple Human Tracking & Action Recognition in a Real Time Video Surveillance", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017).

[6] Jiahao Li†, Hejun Wu∗ and Xinrui Zhou‡, “PeMapNet: Action Recognition from Depth Videos Using Pyramid Energy Maps on Neural Networks”, 2017 International Conference on Tools with Artificial Intelligence, IEEE.

[7] Nour El Din Elmadany, Yifeng He and Ling Guan, “Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis”, IEEE TRANSACTIONS ON IMAGE PROCESSING, 2018.

[8] Soumalya Sen, Moloy Dhar and Susrut Banerjee, “Implementation of Human Action Recognition using Image Parsing Techniques”, 2018 Emerging Trends in Electronic Devices and Computational Techniques (EDCT), IEEE.

[9] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *NIPS 2015 - Advances in Neural Information Processing Systems 28*, 2015.

[10] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus and Yann LeCun, OverFeat: Integrated Recognition Localization and Detection using Convolutional Networks, New York University:Courant Institute of Mathematical Sciences, [online] Available: https://arxiv.org/pdf/1312.6229.pdf.

[11]. Anwarbasha, H., S. Sasi Kumar, and D. Dhanasekaran. "An efficient and secure protocol for checking remote data integrity in multi-cloud environment." *Scientific Reports* 11, no. 1 (2021): 1-8.