PHYSICAL HARASSMENT DETECTION USING DEEP LEARNING

[1]P. H. Habeebur Rahman, [2]Navaneeth Jawahar, [3]K. Bushra Sultana, [4]N. Rajendran

[1, 2, 3] Student, Department of Information Technology, B. S. Abdur Rahman Crescent Institute of science and Technology, Chennai, India

[4]Assistant Professor (Senior Grade), Department of Information Technology, B. S. Abdur Rahman Crescent Institute of science and Technology, Chennai, India

Abstract: Deep Learning is a branch of modern technology and machine learning that is fast being used to solve a multitude of problems that affects one, in daily life. Several social problems can be overcome with the proper implementation and usage of Deep Learning. Rape and harassment cases are unfortunately common problems in a country such as India. Identification of such horrors could go a long way in preventing them. With the advent of powerful processing systems, video analysis can be achieved in real time. The approach suggested in this paper involves devising a harassment detection algorithm that can be deployed in real-time for CCTV cameras, thus, enabling the automatic detection of such a critical situation. This approach is novel in that does not require the victim to do any conscious action such as pressing a SOS button via their phones etc. The system proposed involves determining that a circumstance is indeed critical while it occurs to curb any consequences of delay. It follows the development of a deep learning model trained to identify anomalies in voice, object and movement.

***Keywords*: Deep Learning, Harassment detection, real-time, video analysis.**

I. INTRODUCTION

It had been estimated that in India, rape and harassment is the fourth most common crime to be inflicted. In most of the cases, determining the occurrence of such acts is difficult as they tend to occur in scarcely populated locations. Also, due to this inherent difficulty in detection, timely help and alert of authorities becomes absent.

Though many solutions have been proposed for this issue in the past, many of them require a conscious action from the side of victim, such as pressing keys in their smart phones or smart watches. Such actions are not always possible and thus not very effective.

In order to combat such evils, the presence of an automated system to detect the harassment in real time based on a variety of parameters would be welcome. The presence of CCTV cameras has been widespread in each nook and corner of cities in India. Hence, using them as a tool for the purpose of rape and harassment detection would be an ideal choice.

Several technologies and algorithms can help achieve this feat. Artificial Intelligence is an approach to making computers emulate the human-way of thinking. Deep learning and Machine learning are subsets of Artificial Intelligence that lend a compute the powerful ability to make decisions on its own. A machine learning model can be conditioned to learn and make decisions under multitudes of context.

Video analysis is the approach of analyzing a video automatically and determining the events and objects in it. Video analysis of the CCTV footages can help us in detecting anomalous behaviour that corresponds to a threatened victim or an aggressive attacker. The advantage of video input is its ability to pick up on audio as well. Video analysis using deep learning algorithms has found applications in several fields like objection detection, anomaly detection etc. and found to be well-performed.

Video analysis can be done based on several parameters such as voice, trajectory of objects and weapon detection. Audio from the feed can be filtered to detect certain words indicative of danger such as “help”, “police”, and “stop”. Consecutively, the movement of object between frames will help in determining objects exhibiting abnormal behaviour such as running, attacking etc. Identifying certain objects as weapons in the frame would be crucial in determining the criticality of the situation.

Thus a real-time automated system that takes into account such wholesome parameters could help identify and prevent such mishaps in a timely manner.

II. RELATED WORK

Recognizing violence is a tedious task, since it is subjective concept. Research experiments have been obtained for recognition of an object, audio and actions such us walking, jogging, pointing and hand waving. Since, there is no large number of studied datasets of action recognition specific for violent detection. The main aim of large scale CCTV system used in streets is to facilitate with alerting alarms of potentially dangerous scenarios. Since, it reduces the burden of security guard where they monitor large number of cameras with max response time.

Serrano et al. [1] had taken three dataset on different scenarios such as hockey, movie, behave datasets. The feature extraction aims to obtain representative image from each input video sequence. That increases the relevant motion and reduces the irrelevant background and noise. It involved an exact image tracking for clear study of fight detection it is based on Hough forest and 2D convolutional neural network. 2D-CNN is used for image prediction and Hough Forest is used for image classification Fight is defined as a use of physical strength to try to hurt someone. In that author doesn’t introduce audio search aggressive action in surveillance camera. This system won’t predict particular object (weapon) detection.

Zhang et al. [2] advanced to extend IWLD (Improved Weber Local Descriptor) characteristic by adding a temporal component, called MoIWLD (Motion Improved Weber Local Descriptor) to predict violent actions in real video scenes. The authors updated a sparse representation model to minimize the coding coefficient error and classification error and made a supervised classification

using dictionary learning. The work is based on datasets of hockey fight, behave and crowd violence. The false alarms are happen due to waving flag and vigorously clapping hands.

Zaheer et al. [3] proposed a audio surveillance system using the concept of MFCC(Mel-Frequency Cepstral Coefficients) for processing the audio and making it into input vector for analysing the audio and Deep Boltzmann Machine(DBM) for predicting whether it is a scream noise or not. In this the author has proposed 100 % accuracy by creating an own training and testing dataset of 170 and 70 including scream and normal noise. But it is not able to predict noise like car horn, crowd cheering and kids playing loudly.

Bruno et al. [4] developed an approach to break down the concept of violence into various criteria modelling temporal characteristic in still image to prevent kids from seeing inappropriate content in videos using TRoF (Temporal Robust Feature) to classify video sequence into combinations, CNN is used to classify the concept of violence with respect to inputs. Here training dataset consist of 18 movies and test dataset 7 movies. The authors may not able to classify the whole concept of violence into a single framework and more data need to be trained.

A.S. Keceli et al. [5] proposed a violent activity detection using SVM (Support Vector Machine) and SkNN (Subspace K-nearest neighbour) classifier trained with the dataset pre-trained CNN with deep feature of three datasets of hockey, movie and violent flow (ViF) for classification and high computational speed. It shows low accuracy on crowd violent activity.

Nida Rasheed et al. [6] proposed a method for detecting abnormal activity by making difference from normal activity due to the changes happens in the surroundings. Neural networks are used to validate, estimate and train the results of the training dataset for traffic flow. Optical flow is used to detect minute details of the event.GMM (Gaussian Mixture Model) is a technique to remove the noise from the video data. Lucas-kanade and Horn schunck methods are compared for designing a efficient tracking method using Supervised Neural Network Classification.

III. PROPOSED SYSTEM

The main idea of this paper is to propose a system that could effectively detect the occurrence of harassment through CCTV footages in real-time. An attempt is made to build a deep learning model that incorporates audio, trajectory and weapon detection as parameters to conclusively discern if a mishap is taking place or not. The proposed methodology can be summarized into the following modules:

1. Audio detection
2. Weapon Detection
3. Video Anomaly Detection

A. AUDIO DETECTION

This module attempts to identify stress and danger indicative words in the video footage.

1) Data acquisition

Data collection is performed dynamically in the form of video footages off CCTV cameras and stored in hard disks. The python library ffmpeg is used to separate the audio from the video footage. The subsequent audio file is fed as input to Google voice library for python enables the access of Google voice API. This library is used to convert audio into its corresponding text. The text file is stored in local disk.

2) Natural Language Processing

Natural Language processing which can be done using the NLTK library for python, is used to perform pre-processing, Feature extraction and voice intensity and danger word detection.

Pre-processing: This involves filtering through the text data in order to get rid of unwanted and redundant parts. It involves:

1.Tokenization: The text data is broken down into separate using white space as delimiter and as such unwanted non-textual data are removed using split() method of NLTK library.

2. Stop word removal: stop words are those words that hold no emotion. The words include articles, non-lexical etc. Hence, it is favourable that they be removed. This is done by importing stop words package from the NLTK library.

3. Stemming: It is a process of mapping an inflection of a word back to its root word. This highly reduces the complexity of the data. This is done using the Porter Stemmer package available in the NLTK library.

Feature Extraction: Feature extraction is done using the Bag of Words model. The issue with text data is that it is quite messy and most machine learning algorithms prefer well defined, fixed-length inputs and outputs. Since these algorithms cannot work with text directly, they need to be converted into a vector of numbers. This is also done using the bag of words model.

Bag of words (BoW) representation: The Bag of words model is popularly used for feature extraction in natural language processing. The BoW representation describes the occurrences of words within an article or a document. It involves developing a vocabulary of known words and determining the presence of such known words. Initially, all of the unique words are identified and used to design the vocabulary. Now, since the aim is to vectorize the text, the best way to go about, would be to design a vector of length equal to that of the number of unique words and employ a scoring method for each sentence using period as a delimiter. Obviously, the simplest method of scoring would be to assign 1 for the presence of the word and 0 if not. This vectorization process can be done using the Count Vectorizater class available in scikit-learn. The result would be multiple sparse matrices containing larger numbers of 0 counts. In order to battle memory constraints, simple data cleaning techniques can be employed such as stemming and stop word removal. Other alternative methods of representation are the N-gram and Word2Vec representations.

Intensity and Danger word detection: The NLP algorithm may look for words like help, danger or other regional words indicating danger. Sound patterns are also analysed for high pitch shouting noises. The model is trained using the corpus obtained by the using the bag of words (BoW) model. On encountering a danger or trigger word in the text that matches with the words in the corpus, and based on the intensity of those words, the text file is labelled abnormal.

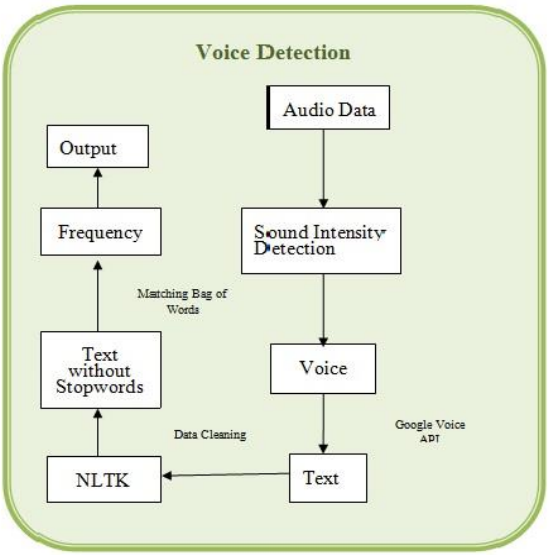

Fig. 1 Module diagram of Audio Detection

B. WEAPON DETECTION

This module aims at detecting the presence of weapons at the scene. It involves:

1) Data collection and pre-processing

Data for training the classifier model is from Google as an extensive dataset of images of various kinds of weapons. These images are then categorized according to their types. For example: guns, knives etc. The data set is split as 80 % for training and 20 % for testing.

2) Model Training

The model is trained using the image data set using Convolution Neural Networks (CNN) which is implemented using Tensorflow packages. Tensorflow gpu is employed. ssd_mobilenet v1_coco has a decent accuracy and speed, so that algorithm is used. Each image goes through a series of convolution layers It involves the following techniques:

Convolution: This first layer is used to understand the relationship between pixels. It is primarily used extract features from the image.

Pooling: Pooling is a technique used to reduce the number of features when the images are large. It is also called sub sampling, reducing dimensions while still retaining important information.

Flattening: This operation converts the output of convolution into a 1-dimensional array which is used by the dense layer for final classification.

3) Classifier model

A classifier is built in order to detect weapon and its type for the incoming video input in real time. The SVM classification algorithm is used. The numpy arrays of the objects detected is compared with the trained and classified arrays in order to detect if it is a weapon and what type of weapon it is.

This “weapon detection” parameter has a heavy weightage in the final result because a weapon will definitely have a negative impact in the situation.

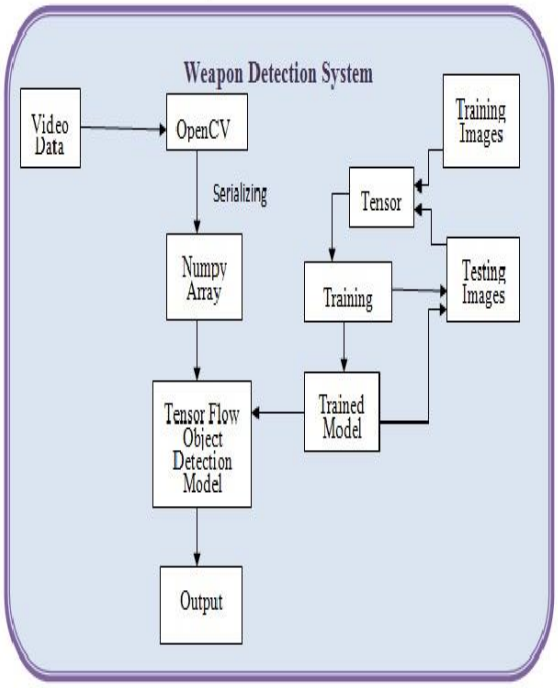

Fig. 2 Module diagram weapon diagram

C. Video Anomaly Detection

Video input is obtained real time and split into frames. The optical flow for each object is calculated to find anomalous behaviour. Optical flow is the process of determining the pattern of movement of object between two consecutive frames. The optical flow computation can be used to derive the structure of motion of an object. The Lucas- Kanade method of optical flow that works on the assumption that all neighbouring pixels have the same motion is implemented. After determining the optical flow of each object in the frames, the objects with unusual behaviour and structure is identified.

The Lucas-Kanade optical flow is implemented as a set of methods available in the OpenCV library. The module requires the use of 3 functions:

cv2.calcOpticalFlowPyrLK () - It is a pretty simple application to track certain points in the video.

cv2.goodFeaturesToTrack () - used to decide which points to track. This is done by taking the first frame, detecting some Shi-Tomasi corner points in it. These points are tracked iteratively using the Lucas-Kanade optical flow.

cv2.calcOpticalFlowPyrLK () - The previous frame is passed, its previous points and next frame. It returns next points along with some status numbers which has a value of 1 if next point is found, else zero. This is iteratively done till the last frame.

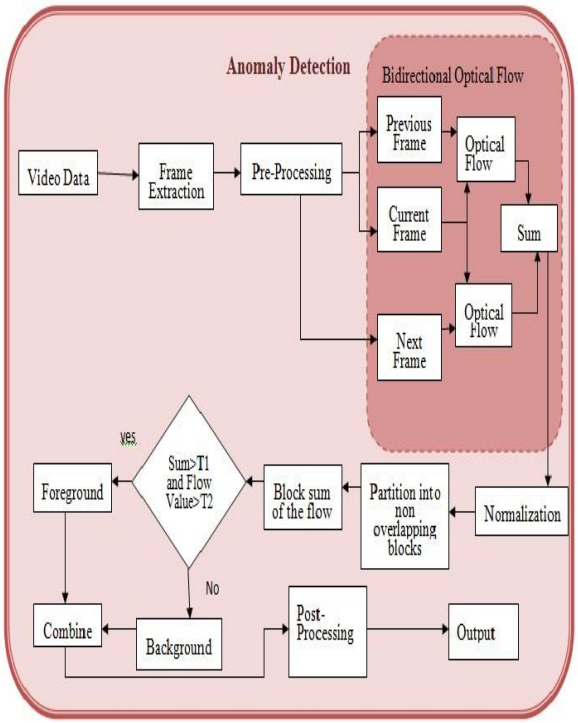

Fig 3: Module diagram of video Anomaly Detection

IV. RESULTS AND DISCUSSION

An overview of the performance of the classifiers is discussed. Also, the accuracy the models is also mentioned.

V. CONCLUSION

Thus, the successful implementation of the above proposed system could ensure the safety across all areas and neighborhoods and eliminate several societal threats. The absence of the need for human operation ensures all round the clock monitoring. This efficient system can be implemented across several social scenarios, creating a safe and watchful environment for the public

VI. FUTURE WORK

This is system can be further extended to include posture detection module to enhance the precision of harassment detection. This system can be incorporated in all environments where CCTV cameras find application. This system can be implemented, with a few added parameters, for the detection of thefts and muggings. The system can be further enhanced to dynamically send alerts to the relevant authorities regarding the incidents thus prompting quick action. With increasing power of processing systems, the proposed system can be implemented on a much larger scale while being economically efficient.

VII. REFERENCES

[1] Ismael Serrano, Oscar Deniz, Jose L. Espinosa-Aranda and Gloria Bueno, Flight Recognition in video using Hough Forest and 2D convolutional Neural Network, IEEE Transaction on image processing, 2018, Volume 27, Issue 10, pg-4787-4797

[2] Tao Zhang, Wenjing Jia, Xiangijan He and Jie Yang, Discriminative dictionary learning with motion Weber local descriptor for violence detection, IEEE Transactions on Circuits and Systems for Video Technology, 2017, Volume 27, Issue 3 , pg:696– 709.

[3] Md.Zaigham Zaheer, Jin Young Kim, Hyoung-Gook Kim and Seung You Na, A Preliminary Study On Screaming Sound Detection, IEEE, International Conference on IT Convergence and Security(ICITCS), 2015, 4 pages.

[4] Bruno Malveria Peixoto, Sandra Avila, Zanoni Dia and Anderson Rocha, Breaking Down Violence- A Deep Learning Strategy to Model and Classify Violence in Video, ACM, International Conference on Availability, Reliability and Security, 2018, 7pages

[5] A.S.Keceil and A. Kaya, Violent Activity Detection with Transfer Learning Method, IEEE – Electronic Letter, Volume 53, Issue 15, pg- 1047-1048, 2017.

[6] Nida Rasheed, Dr. Shoab A. Khan, Adnan Khalid, "Tracking and Abnormal Behavior Detection in Video Surveillance using Optical Flow and Neural Networks", IEEE Computer Society, 28th International Conference on Advanced Information Networking and Applications Workshops, 2014.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatio temporal features with 3d convolutional networks, IEEE international Conference on Computer Vision (ICCV), 2015, pg 4489-4497.

[8] Chetana D Patil and Bharathi V K , "Event Detection in Video Using Saliency Value and Histogram of Optical Flow", International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2016,Volume 5, Issue 5, pg- 4201-4205

[9] Malay Shah and Prof. Rupal Kapdi, "Object Detection Using Deep Neural Networks", IEEE, International Conference on Intelligent Computing and Control System(ICICCS), 2017, pg:787-790.

[10] Robert Leyva, Victor Sanchez and Chang-Tsun Li, "Video Anomaly Detection With Compact Feature Sets For Online Performance", IEEE Transactions On Image Processing, Volume 26, Issue 7, July 2017.

[11] Y. Yuan, J. Fang and Q. Wang, "Online Anomaly Detection in Crowd Scenes via Structure Analysis", IEEE Transaction on Cybernetics, Volume 45, Issue 3, pg: 548-561, 2015.