Real Time Scene Analysis for Automated Threat Detection using AI

Kunal Parikh
Department of Information Technology Engineering
A. D. Patel Institute of Technology New Vidya Nagar, Anand, Gujarat, India

Abstract— Robbery crimes are one of the most common crimes. There are many causes of such crime includes poverty. One of the primary cause is inadequate security including lack of alarm systems. The main target of robbers are shops and banks. There is no proper automated system through which measures can be taken against robbery. Proper Security systems is the need of the hour. The system must be capable to work without human intervention as there might be some cases where small mistakes to inform the police could be fatal.

Keywords: Threat Detection, Artificial Intelligence, Image Processing, Scene Analysis, Machine Learning

I. Introduction

Four major causes of robberies are:
1) Inadequate security
2) Poverty
3) Substance abuse
4) Peer Group Pressure

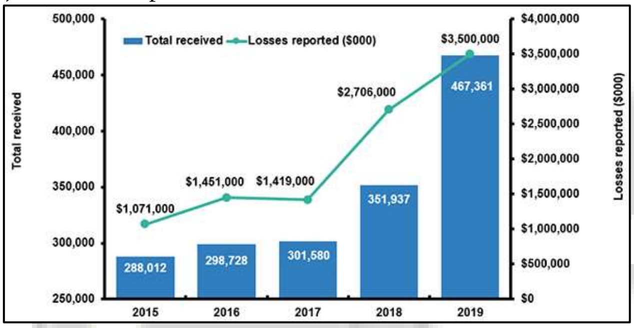

Fig. 1: Annual Loss reported due to robberies/theft

The aim is to develop an automated system that can detect certain threats that normally occurs in shops or banks. Moreover, it can also suggest best measures or actions that can be taken so as to minimize the loss occurred because of such threats. This can be done by analyzing the crime scene in real time and detecting situation which usually occurs in a robbery.

A. Emotion Classification based on Facial Features

The goal is to get a quick baseline to compare if the CNN architecture performs better when it uses only the raw pixels of images for training, or if it's better to feed some extra information to the CNN (such as face landmarks or HOG features). The results show that the extra information helps the CNN to perform better. To train the model, Fer2013 dataset that contains 30,000 images of facial expressions grouped in seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral was used.

The faces are first detected using OpenCV, then to extract the face landmarks using dlib. The HOG features are extracted and the raw image with the face landmarks+hog is inserted into a convolutional neural network.

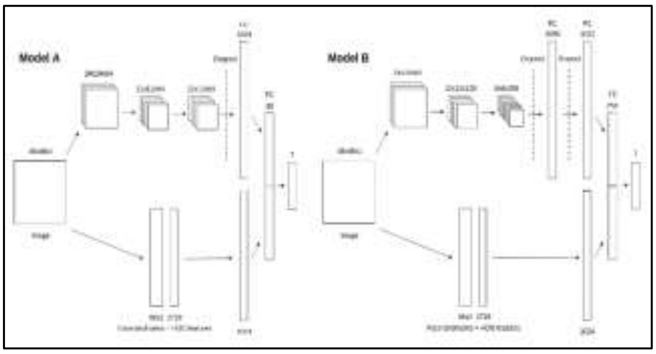

Fig. 2: Working of Emotion Classification

B. Weapon/Threat Object Detection

Object Detection using Haar feature-based cascade classifiers is an effective object detection method proposed by Paul Viola and Michael Jones in their paper, "Rapid Object Detection using a Boosted Cascade of Simple Features" in 2001. It is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. It is then used to detect objects in other images.

Initially, the algorithm needs a lot of positive images (images of faces) and negative images (images without faces) to train the classifier. Then we need to extract features from it. For this, Haar features shown in the below image are used. They are just like our convolutional kernel. Each feature is a single value obtained by subtracting sum of pixels under the white rectangle from sum of pixels under the black rectangle.

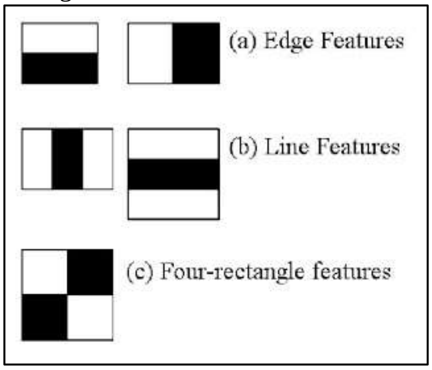

Fig. 3: Haar Cascade

For this, we apply each and every feature on all the training images of different types of weapons and threat objects. For each feature, it finds the best threshold which will classify the faces to positive and negative. Obviously, there will be errors or misclassifications. We select the features with minimum error rate, which means they are the features that most accurately classify the face and non-face images. (The process is not as simple as this. Each image is given an equal weight in the beginning. After each classification, weights of misclassified images are increased. Then the same process is done. New error rates are calculated. Also new weights. The process is continued until the required accuracy or error rate is achieved or the required number of features are found).

In the concept of Cascade of Classifiers, instead of applying all 6000 features on a window, the features are grouped into different stages of classifiers and applied one-by-one. (Normally the first few stages will contain very many fewer features). If a window fails the first stage, discard it. We don't consider the remaining features on it. If it passes, apply the second stage of features and continue the process. The window which passes all stages is a face region.

C. Speech based Emotion Classification

This module uses machine learning that could detect emotions from the speech we have with each other all the time. Nowadays personalization is something that is needed in all the things we experience every day.

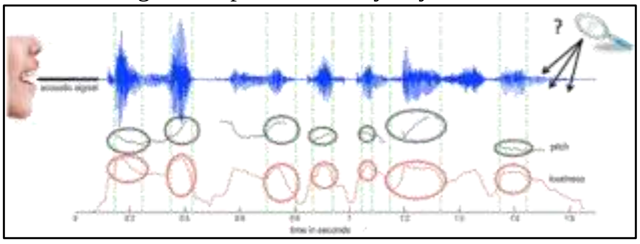

Fig. 4.1: Audio Feature extraction

RAVDESS. This dataset includes around 1500 audio file input from 24 different actors. 12 male and 12 female where these actors record short audios in 8 different emotions i.e. 1 = neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7 = disgust, 8 = surprised. Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

SAVEE. This dataset contains around 500 audio files recorded by 4 different male actors. The first two characters of the file name correspond to the different emotions that portray.

The next step involves extracting the features from the audio files which will help our model learn between these audio files. For feature extraction we make use of the LibROSA library in python which is one of the libraries used for audio analysis.

Since the task is an order issue, Convolution Neural Network appears to be the obvious decision.

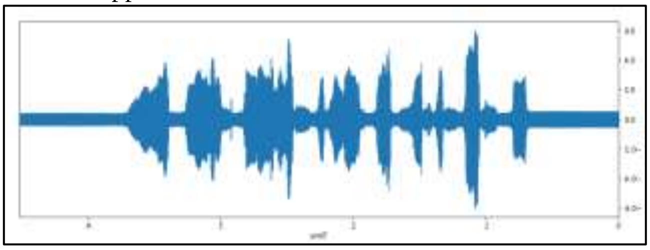

Fig. 4.2: Audio Waveform captured using LibROSA

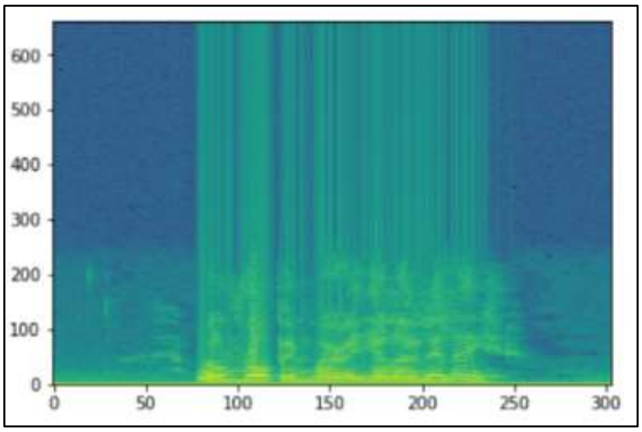

Fig. 4.3: Audio Spectrogram

D. Pose Estimation

Human Pose Estimation is defined as the problem of localization of human joints (also known as key points - elbows, wrists, etc.) in images or videos. It is also defined as the search for a specific pose in space of all articulated poses.

Fig. 5.1: Pose Estimation

PoseNet is a vision model that can be used to estimate the pose of a person in an image or video by estimating where key body joints are. The PoseNet model is image size invariant, which means it can predict pose positions in the same scale as the original image regardless of whether the image is downscaled. This means PoseNet can be configured to have a higher accuracy at the expense of performance. Pose Estimation model uses Tensor flow.

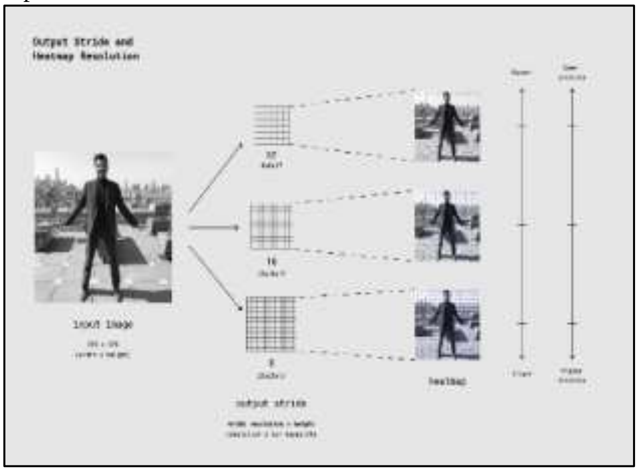

Fig. 5.2: Pose Estimation mechanism

II. WORKFLOW

In an ideal scenario, the thief/robber usually enters the premises with a gun or a knife with their face covered with helmet or facemask. The robber usually targets the cashier at the cash counter. The setup required for this system includes a dash camera (wide angle) with microphone which is responsible for the surveillance of the cash counter area and record the conversation of the robber and people around and CCTV cameras to detect pose estimation along with weapon detection.

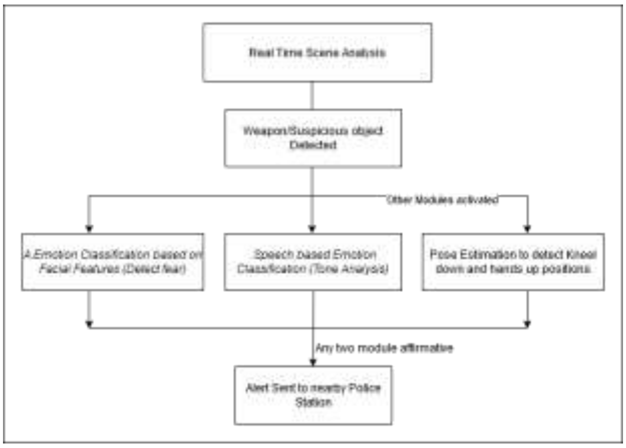

Fig. 6.1: Flow of system

When the robber enters with a weapon, the CCTV camera at different angles captures the weapon. If a threat object is detected, other modules are activated. These modules only get activated for a specific time to avoid false alarms as there can be cases of inaccuracy or guns found on people outfits.

Cascade Classifier is initially implemented to train and test the initial dataset. The image classifier will be initially trained using some train datasets which is collection of positive images (which contains a weapon, helmets, face masks) and negative images (frame which does not contain these).

For the preprocessing part, each frame (RGB) converted into Grayscale image using OpenCV to reduce the complexity of the model. Furthermore, Gaussian Blur to used to reduce the level of noise in the image, which improves the result of the following edge-detection algorithm.

OpenCL, a low-level API for heterogeneous computing that runs on CUDA-powered GPUs is used as a runtime environment.

Below snapshots depicts the weapon/suspicious object detection algorithm in action.

Fig. 6.2: Detection of Threat Object/weapon

Fig. 6.3: System activates other Modules

Once the threat weapon/suspicious object is detected, other modules get activated. The dash camera captures the facial emotions of the cashier or the person near the area where the valuable are kept. The ML model looks for the expression of fear on the cashier's face. Furthermore, the microphone captures the audio and the tone of the conversation around is analyzed in real time. The speech classification feature also uses Natural Language Processing and sentiment analysis to understand the conversation and look for threat/robbery scenarios. The CCTV camera analyzes the whole scene and detects kneeling down and hands up positions if any.

There might be scenarios where all the above conditions are not detected, so even if two conditions are stated positive by the system, an alert is sent to the nearby police station in real time without the robber being aware of it. This make the system autonomous and relative safer.

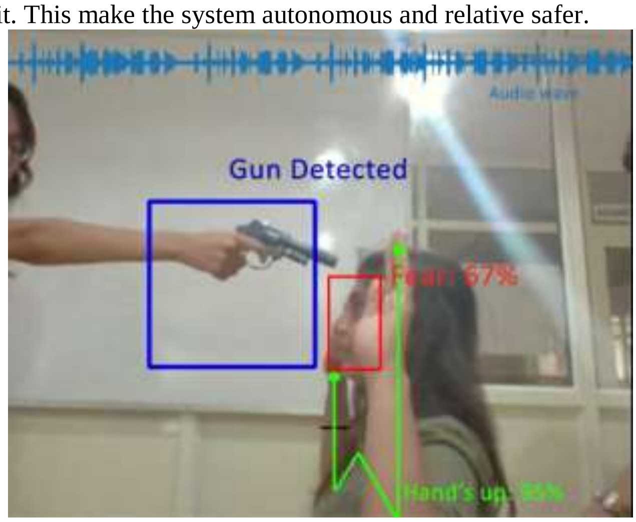

Fig. 6.4 Waveform Analysis, fear detected, hands up position

III. CONCLUSION

By using this system the threats can be detected automatically. Moreover, this system is quick and responsive which makes easy to take fast actions as that is at the top priority. Also this system is more accurate as object detection technique is used followed by facial emotion detection, pose estimation and tone analysis to make precise decision.

As this system is automated there is no need of human operation which help to reduce fatality. To illustrate it briefly, if a person calls to the police then it may be possible that the robbery may take some responsive actions and that can more dangerous. So by eliminating human operation this system is safer and fast as well.

REFERENCES

[1] Rafael Padilla, Cicero Ferreira Fernandes Costa Filho ,Marly Costa "*Evaluation of Haar Cascade Classifiers for Face Detection*."
[2] Alexander Toshev, Christian Szegedy "*Human Pose Estimation via Deep Neural Networks*."
[3] Nithya Roopa.S "*Emotion Recognition from Facial Expression using Deep Learning*"
[4] Pawan Mishra "*Emotion Recognition through Speech Using Neural Network*".
[5] Ayush Jain, Mohana "*Real Time Object Detection and Tracking Using Deep Learning and OpenCV*."
[6] Hua-mei Chen, Raghuveer Rao "*Imaging for concealed weapon detection*."
[7] Paul Viola, Michael Jones "*Rapid Object Detection using a Boosted Cascade of Simple Features."*