Violent Human Behavior Detection from Videos using Machine Learning

Mohammad Sadat Hussain Rafsanjani and Ahmedul Kabir*

Institute of Information Technology, University of Dhaka, Dhaka-1000

E-mail: kabir@iit.du.ac.bd

Received on 07 February 2021, Accepted for publication on 13 December 2021

ABSTRACT

Surveillance and security cameras are becoming much more common in city streets, shopping malls, private homes and many other places. In recent years terrorist attacks in shopping malls, schools and public places have increased. Such places, if equipped with a computer vision-based system which can effectively identify abnormal or violent human behavior, can help us in saving many human lives in due time. In this paper, we present such a system which can detect abnormal or violent human behavior in a video using machine learning. The system first identifies every human present in a video, extracts some vital data for that person and then, based on those data points, trains machine learning models to perform the classification. We have found that the system works reasonably well under certain conditions.

Keywords: Computer Vision, Machine Learning, Violent Human Behavior, Violent Behavior Detection.

1. Introduction

Crime rate is on the rise all over the world. Gun crimes, terrorist attacks and public place bombings are fatal and are increasing in an alarming rate. To fight such kinds of acts, numerous countermeasures can be taken. One effective way of fighting such crimes is monitoring human behaviors closely. But it is impossible for a single entity to track and monitor each and every person walking on the street, shopping mall and crowded place. Instead we can build a system that can automatically monitor people's behaviors and provide with feedback for a period of time. If the system senses any unusual activity through data point readings, it will alarm the authority in real time.

Security cameras and Close Circuit TVs are now common in cities, mall, train stations, airports and on roads. CCTV footage can be analyzed to monitor, detect or assess current situation of a place. In case of an unfortunate violent event, authorities are sometimes minutes away from the scene. But lack of proper security measures, such as early detection and alarm system, makes it difficult for them to intercept the attack and save human lives.

In this study, we propose a system that can identify whether a video contains any violent human activity using machine learning. We collected a set of videos, from which our system extracted relevant features for violence detection. Our original contribution is in identifying and processing the features that are fed to the learning algorithms. The system currently works well under ideal conditions. It first identifies the human beings present in the video. For each human, some basic data is collected. Classification of either 'violent' or 'non-violent' is performed using machine learning on the collected data.

A. Objectives

The objective of our system is to identify sudden change in human behavior. This can be captured by following some parameters like the sudden change in movement, direction, speed, accelerations, etc. Our target is to gather as much data as possible to build up a data set and apply some machine learning algorithm to construct a model. Based on that model, we can classify new videos. We can detect human in a previously unseen video and try to make a prediction whether that person is behaving normally or abnormally.

B. Types of Violent Human Behavior

The violent activities can vary based on the nature of events happening in a video involving human subjects. Violent and abrasive human behavior can be categorized in broad range of types. Some of them are described below.

1) *Street Fight:* Street fighting is an act of violence where two or multiple persons engage in a physical confrontation. As they fight among each other, due to the random body movement their body position, directions of running and defensive movements etc. rapidly changes which is unpredictable. There are situation which may look like street fights like military drills, play, movie scene or a football game. But in a street fight, sudden displacement of body, acceleration rate and directions are a bit different.
2) *Running Scared:* When any sort of accidents or attacks happen, people usually run in opposite direction. Events such as sudden explosion or bomb blast, fire accidents, road traffic accidents affects people behavior in a certain way.
3) *Terrorist Activities:* Terror activities means when a group of armed assailants attacks the mass with automatic weapons and other gears. Such incidents are becoming very frequent in Middle East and Far East countries. Terrorists usually attack shopping malls, hotels, train stations and schools. When people are attacked with automatic weapons, they reacts abruptly which is reflected in their body movements and other vitals.
4) *Shooting:* Gunman attacks are on the vibe. Departmental stores, restaurants and shopping malls usually have low security standards thus anyone can enter such premises with firearms. People reacts violently seeing firearms. Such pattern can be used to prepare a data set point.

C. Proposed System

The proposed system captures vital features from a video, analyzes them, and builds a classification model based on them. Since humans behave abruptly at the time of unwanted or disastrous moments, vitals like their speed, direction and distance will changes suddenly and abnormally. Our system works under the rigid assumption that violent activities are the ones that trigger sudden changes in either speed or direction of humans. Therefore, activities that involve little movement - for example, shooting someone while standing at one place - will not be detected as violent. We also do not employ any image processing technique to detect firearms, fire, smoke or bomb blast.

2. Backgrounds

A number of state-of-the art commercial video surveillance systems have been available and in use like [1] and [2] in the world. All of these systems are capable of detecting and tracking multiple people in real time. However, most of these systems are not designed or comes with the built-in programs to monitor human behaviors, categories them whether it is normal or abnormal and preserve the data for later use. A number of handful methods have been devised to recognize human action, this includes Hidden Markov Models in case of human action classification [3], where human activities have been represented by a set of postures and velocity vectors [4]. Using the body position and velocity of different limbs, the system can learn from the video. In another research, the learning [5] approach is based on the border information of the blobs is applied. The approach has been proved to be robust in detecting human abnormal behaviors in several experimental demonstrations. In separate study [6] a model is developed with general crowd behavior motion is simulated of a typical population for a specific environment.

Distributed random behavioral model is used to create individual parameters. Relationship between an autonomous virtual humans of a crowd and the emergent behavior originated from it worked as the inspiration for this research. Few concepts from sociology were employed to represent some specific human behaviors and visual output. Thereafter, the model was applied in two different applications: a graphic system called “Sociogram” which visualized the nature of a crowd during the simulation. It also showcased some aspects about the condition of human crowd collision as well.

In recent years it has been observed that, interest about assisted living environments especially for the elderly who live alone has increased due to the incline in senior citizens population growth. To ensure their safety and healthy environment, we need to detect abnormal behavior that may pose as an imminent threat to these people and cause severe harm to them, both mentally and physically. In the paper, a wireless sensor networks based abnormal behavior detection system is proposed. The system models a series of events called episode, where spatial and temporal information regarding the monitored target are included. Later, two different episodes are compared based on similarity scoring functions where temporal aspects are taken inconsideration.

A. Classification Algorithms

Since the nature of this study is to find out whether the human behavior in a video is violent or not, clearly it is a classification problem. To classify human behavior, we prepare a dataset and train it with classification algorithm. Now there are lots of classification algorithm available. For simplicity and practicality, we have chosen to use most common classification algorithms to train the model. Also since our attributes are all numerical, evidence shows that, these does not require much complex algorithm to train and hyper parameter tuning. These algorithm includes:

1) Decision Tree
2) Support Vector Machine
3) K-Nearest Neighbor
4) Logistic Regression
5) Naive Bayes Classifier

1) Decision Tree: A decision tree [7] is a tree where each internal node is a decision node specifying a set of values for a particular feature, and the leaves are values of the target feature. A decision tree is induced using a fairly simple algorithm. When growing a tree, at each step the algorithm decides which feature to split the on. This continues until all the instances are accounted for, or a stopping criteria is fulfilled. Sometimes, pruning is done on the tree to reduce over fitting.

But how to decide which attribute is more important in the decision making? This is done in a process called attribute selective measure. There are two measures available:

- Gini Index
- Information Gain (ID3)

The Gini index specifies the probability of a particular data point being misclassified when randomly chosen [8]. At each decision point, the feature with the minimum Gini index is chosen. Gini index is defined as follows:

$$Gini = 1 - \sum_{n=1}^{n} (p_i)^2$$

Another way is information gain [9] which is also called entropy. Entropy helps to determine a feature that conveys maximum information about a class. At each decision point, the feature with the highest information gain is chosen. Entropy is defined as follows:

$$Entropy = \sum_{n=1}^{n} -p_i \log_2 p_i$$

2) Support Vector Machine: A Support Vector Machine (SVM) [9] is a discriminating classifier. The goal of this classifier is to find the optimal separating hyperplane that can discriminate between two classes. In n-dimensional space, this hyperplane is a (n-1)-dimensional line dividing the space into two parts. Image processing has been one of the major areas where SVM has been applied with relative success [10].

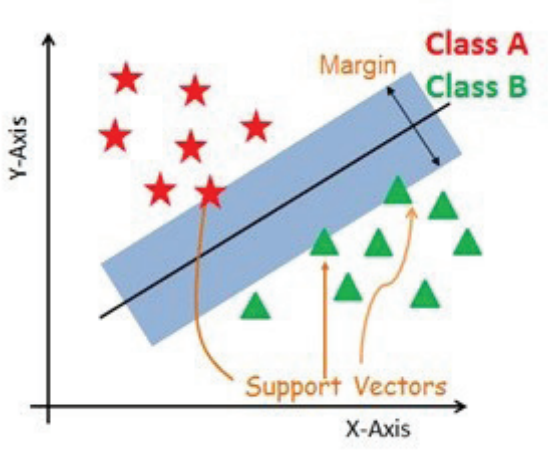

Fig. 1. SVM

A problem with SVM is that it is difficult to find the optimal separating hyperplane in a short amount of time when a huge number of data points are involved. In that case, a number of techniques is used which involves tuning various parameters. This is called regularization. For example, when data points cannot be differentiated in two dimensions, higher level dimension are implied to point out the boundary. There we apply the kernel function.

$W \cdot X + b = 1$

$W \cdot X + b = -1$

3) K-Nearest Neighbor: K nearest neighbors (KNN) [11], [12] is a 'lazy' non-parametric learning algorithm that uses the similarity or dissimilarity between data points in its classification process. In order to classify an instance, its K nearest neighbors are identified, and their labels are checked. The class label that is present in a majority of neighbors is the one that is predicted for the instance.

There are a couple of issues to consider when using KNN. Firstly, there are several distance functions available, the common ones being Eucledian distance, Manhattan distance and Minkowski distance. These distances are defined as:

$$Euclidean \quad Distance = \sqrt{\sum_{n=1}^{k}(x_i - y_i)^2}$$

$$Manhattan \quad Distance = \sum_{n=1}^{k}|x_i - y_i|$$

$$Minkowski \quad Distance = (\sum_{n=1}^{k}(|x_i - y_i|)^q)^{\frac{1}{q}}$$

The other issue is finding the best value of K. If K is too low, the algorithm might overfit, while too high a value might cause under fitting. Usually, many different runs with different values of K are made before choosing the final one.

4) Logistic Regression: The logistic function, also known as the Sigmoid function, is an S-shaped curve that can take a real number and map it into a value between 0 and 1 [13]. The function is defined as follows:

$$Sigmoid(z) = \frac{1}{1 + e^{-z}}$$

Logistic regression is an extension of linear regression, but one to be used for classification rather than regression. The advantage of Logistic regression is that it gives an easy toexplain, for which reason it is a popular algorithm in domains outside computer science.

5) Na¨ıve Bayes: It is a simple algorithm that calculates the probability of a data instance belonging to different classes using Bayes' theorem, and assigns the instance to the class with highest posterior probability [14]. Let x = (x1,...xn) be an instance to be classified represented by a vector with n independent variables, and let c be one of the classes. Then, using Bayes' theorem, the conditional probability of class c given x is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

This algorithm makes the "na¨ıve" assumption that the variables are independent. Nevertheless, it has proven to be an empirically successful supervised learning algorithm, often outperforming more sophisticated algorithms.

3. Methodology

The proposed system has two parts. The model preparation phase acts independently and takes a number of normal and crime related videos as input. Then these videos are analyzed and a number of attributes are collected from them in order to construct a dataset which is used to train multiple prediction models. Our main contribution is in identifying and processing the attributes that should be extracted for constructing the dataset.

In the second part, a previously unseen video is fed to the system upon which the newly trained model is applied. Based on the dataset and data points extracted from the video, the model identifies if there is any abnormalities in the video or not.

A. Overview of the System

The data set consists of several attributes collected from video analysis. These attributes are nominal in type. Basically, we have two set of attributes identified for now. Analyzing the videos, we get the video frame rate and total number of frames in the video. These two attributes will not be part of the data set but will be helpful to prepare it. Thereafter, we detect all the human in a video frame and index it. At first, we identify the position of a target. Subsequently, we collect the speed, distance traveled by the person, direction and acceleration rate of the particular person frame by frame. The direction is calculated in theta scale of 180 degrees as we plan to perform all the classification in 2-dimensional plane. Following this, we prepare a data set that consists of these attributes and label the data accordingly.

We have also developed a GUI-based application that takes video input and extracts vitals for each humans present and identified in it. Then using the previously trained model, we can detect if the person of interest is terrorized or panicked and acts accordingly. Fig. 2 presents the high-level view of the system, whereas Fig. 3 presents the overall working procedure.

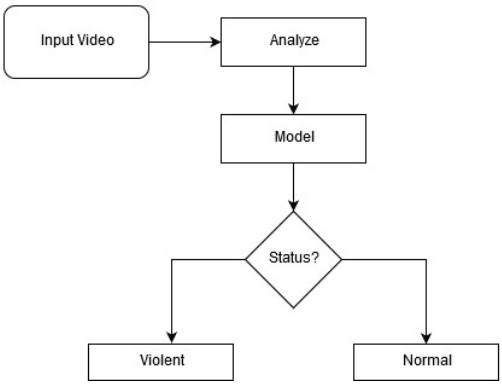

Fig. 2. High-level view of system

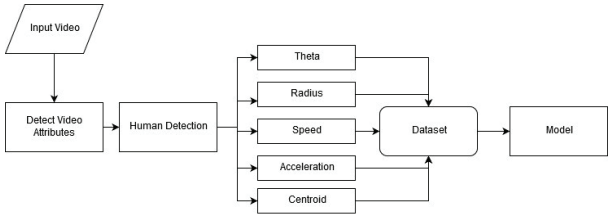

Fig. 3. Working procedure of proposed system

B. Detailed Working Procedure

1) **Video Collection and Pre-processing:** To train the system and produce a model, we need data. Hence, we collected 20 videos from various video sharing sites (Youtube, Metcafe, Vimeo etc.) for processing. These videos contains only violent human behavior like gun shooting, street fighting, bomb blast etc. All of these videos are previously unseen by the system before training started and never been used in any research before.

The type of video we used in this system is very specific in nature. The video we collected are from CCTV footage and security cameras. This gives us two advantages. Firstly, the camera is still, thus we don't have the relative speed issue of a moving subject. Secondly, the direction considered in the video is fixed. Thus the targeted human detected on camera has certain directions to move on relative to camera and computer screen.

When collecting a video, we kept only the part that contains violent activity involving human. Also, videos with border and ads are cropped for consistency. Unclear or low quality videos are not used. The videos we collected has the resolution of 1080x720 (HD) at least to ensure the image quality. Videos where the camera is not stationary are not used, since if the camera is moving on its own, human distance calculation can be compromised.

2) **Human Detection:** For human detection, we are using Google YOLO (You Only Look Once) [15] version 3 API. YOLO is a model for object detection in image. The reason we choose YOLO API for human detection is it is already state-of-the-art hence we don't feel the necessity to develop our own human detection system. As video stream is a collection of lot of images, it can be analyzed by YOLO sequentially, frame by frame. The goal of the object detection task is to determine the location and class of certain objects present on the image.

Using the YOLO API, we analyze each video and detected human in it. We mark each human with a separate number, superimpose the indexing number at their center point which is close to their chest, and track it. Any new human approaching towards the camera is assigned a new value.

Fig. 4. Humans identified using YOLO

YOLO API gives us the location of a human and surrounds it with a box. We calculate the center point coordinates of the bounded box and record it. For dataset preparation, we also convert the coordinate values in polar format for later use.

$$distance = (x_1 - x_2)^2 + (y_1 - y_2)^2$$

3) **Distance Calculation:** The center point pixel value changes over time. Distance travelled by each human across the frame can be calculated by the distance formula. We apply Euclidean distance function to measure the pixel distance. However, we need to be careful about what amount of changes are considered as movement. For example, if a person is standing on the same place but moving or shaking his upper part of body, circling around in a small region, sitting down from standing position or standing up from a chair, in these cases the center point value will change. We need to avoid such incidents that is why when detecting the center point, we depend on a threshold value. When we calculate the center point frame by frame, we check if the distance between the points are greater than the threshold value. Otherwise we discard it.

4) **Speed & Acceleration Calculation:** As we get the coordinates of center point, we start to record the timestamp of each center point as it rapidly changes in consecutive frames. Speed is calculated of each center point in separate but successive frames.

$$speed = \frac{distance\ travelled\ between\ frames}{time}$$

Once the speed and distance between the frames are calculated, we calculate the acceleration value by taking the difference between speeds divided by the timestamp in consecutive frames.

$$acceleration = \frac{speed\ calculated\ between\ frames}{time}$$

5) **Dataset Preparation:** The dataset is prepared with all the values we recorded and calculated previously, center point coordinate in both polar and Cartesian format, speed, acceleration rate and time. But the dataset lacks label value till now. For that, we apply some heuristics on speed and theta values. Initially, the label is marked false. If the dataset attribute values are increasing or decreasing abruptly, we mark consecutive labels into the opposite value. We analyzed a large number of normal and abnormal incident videos to build a large dataset.

6) **Training the Model:** After the dataset is prepared, we use five separate supervised machine learning algorithms, namely Decision tree, SVM, K-Nearest Neighbors, Logistic Regression and Naive Bayes, to train the model. Nominal values are normalized for smooth processing. We individually train the model and test it on a new video.

7) **Detection:** A new video is fed to the system and the trained model is applied on it. Each of the five models analyzes each frame separately and detects anomalous human behavior on it. Later on, we combined results for each frame to a single place and determine the result based on a majority vote. At every 10 frame interval, the system outputs a detection.

4. Result

In this section, we present the detailed results of two example videos. To get some insight from the data points, we plot the direction, speed and acceleration graphs of each video. In Fig. 5, we show snapshots from a video (Video 1) where a person holds on a gun, comes near a car, fires some shots and goes to back to the direction he came from. This sudden change of direction and displacement is suspicious.

If we observe Fig. 6 where theta is plotted against the time frame by frame, we see the curve at first goes lower in a constant rate. Then suddenly it grows higher, changing the direction suddenly.

Fig. 5. Snapshots from Video 1, a video showing violent behavior

Such fluctuations are abnormal even in the naked eye of an observer.

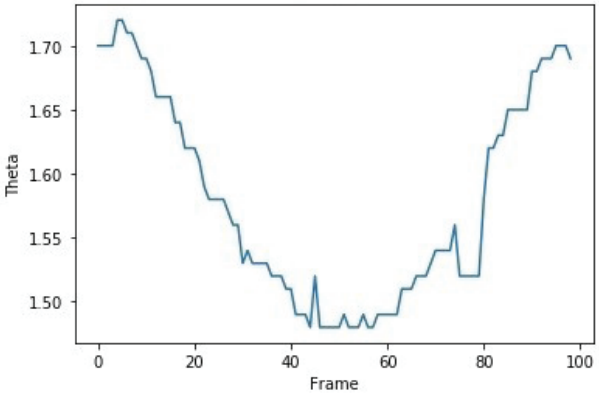

Fig. 6. Direction Graph for Video 1

This is also observed in the speed graph of Fig. 7, which is also plotted against time. The speed grows suddenly and then falls. In respect to the frames, the speed changing rate is very high.

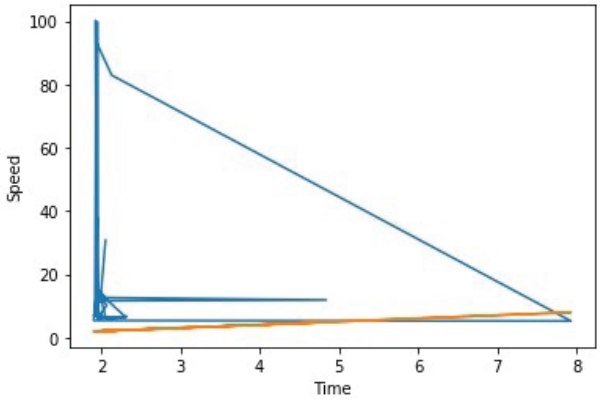

Fig. 7. Speed Graph from Video 1

Now we analyze a separate video (Video 2), where a group of people are standing and having a chat. When analyzed, our system did not raise any issue. The direction graph (Fig. 10) and speed graph (Fig. 11) are also normal. Although

some people moved around a bit, but due to the threshold value, it is deemed negligible during processing.

Note that, compared to abnormal behavior direction graph, the theta value here is moving very slowly which is negligible and very gradual. The speed change over time is very slow and gradual. This is a clear indication of stable human behavior. The acceleration graph also shows the same.

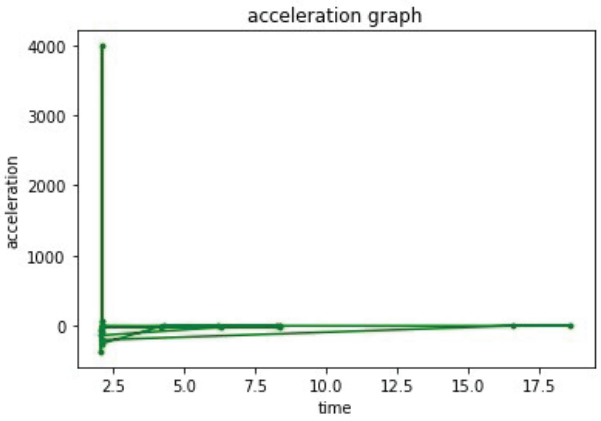

Fig. 8. Abnormal acceleration rate in Video 1

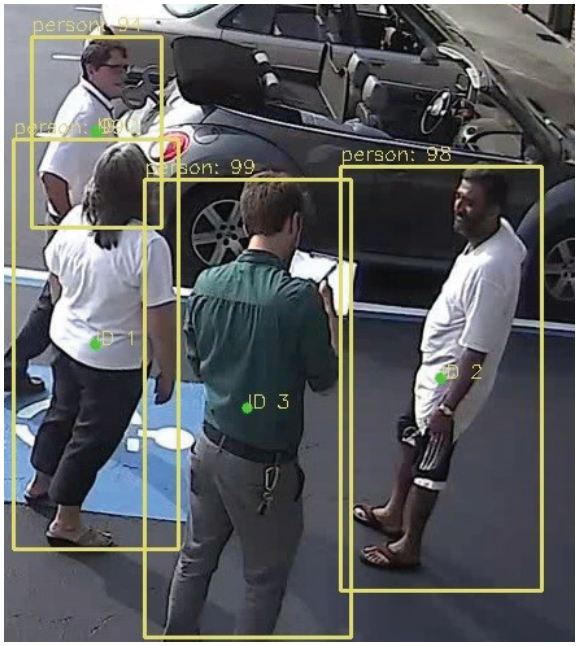

Fig. 9. Snapshots from Video 2, a video showing normal behavior

5. Conclusion

Our system can detect the human in a time frame, record the relevant attributes and based on that data can detect whether there is any abnormal behavior representing violence present in the frame or not. We have identified the different parameters that are useful in identifying the movement of humans in a video. We have then collected a dataset that uses several points to detect the abnormalities. The system works well only under conditions where the movement is of a certain type.

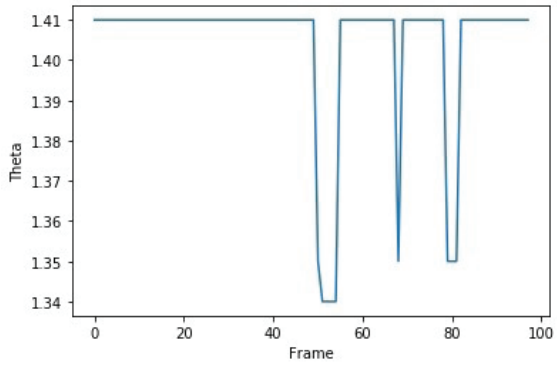

Fig. 10. Direction graph for Video 2

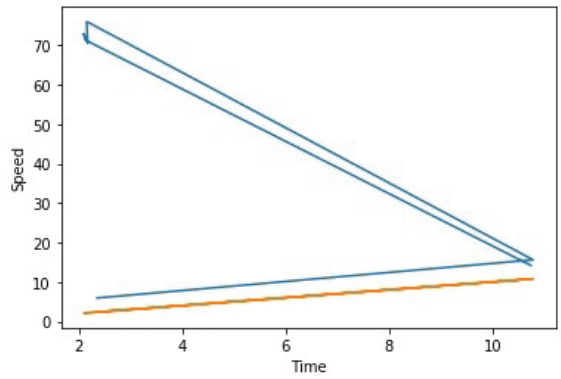

Fig. 11. Speed graph for Video 2

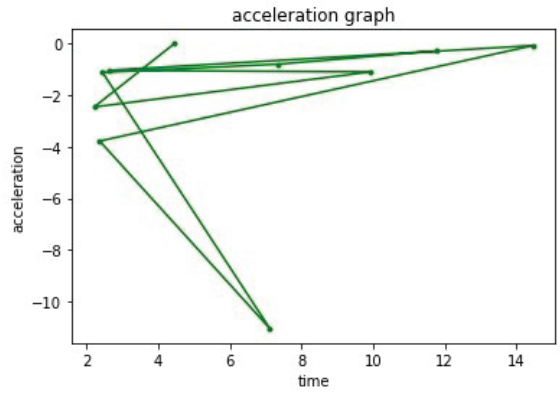

Fig. 12. Acceleration graph for Video 2

Fig. 13. GUI Application

For example, it can detect human coordinates (x,y) but if the person is moving towards the camera or moves back in the same axis z, our system fails to recognize it. Also, our system cannot recognize violence that involves minimal movement. In the future, our plan is to expand this work into a robust system where different types of motions will be addressed, and various forms of violent behaviors will be identified. We will also perform a thorough comparison of our method with other existing methods on the video dataset that we collected.

References

1. K. Park, Y. Lin, V. Metsis, Z. Le, F.S. Makedon, "*Abnormal Human Behavioral Pattern Detection in Assisted Living Environments*", PETRA '10: Proceedings of the3rd International Conference on Pervasive Technologies Related to Assistive Environments Article No.: 9 Pages 1–8, 2010, DOI: https://doi.org/10.1145/1839294.1839305
2. X. Wu, Y. Ou, H. Qian, Y. Xu, "*A Detection System for Human Abnormal Behavior*", IEEE RSJ International Conference on Intelligent Robots and Systems, 2005, DOI:10.1109/IROS.2005.1545205
3. P. Afsar, P. Cortez, H. Santos, *"Automatic visual detection of human behavior*", Santos, Expert Systems with Applications,Volume 42, Issue 20, Pages 6935-6956, 2015
4. B. Krausz, C. Bauckhage, "*Automatic Detection of Dangerous Motion Behavior in Human Crowds*", 8th IEEE InternationalConference on Advanced Video and Signal Based Surveillance (AVSS),2011, DOI: 10.1109/AVSS.2011.6027326
5. S. R. Musse, D. Thalmann,"*A Model of Human Crowd Behavior: GroupInter-Relationship and Collision Detection Analysis*", Inter-Relationship and Collision Detection Analysis. Computer Animation and Simulation'97. Eurographics. Springer, Vienna, 1997
6. M. J. Roshtkhari, M. D. Levine, "*Online Dominant andAnomalous Behavior Detection in Videos*", IEEE Conference on Computer Vision and Pattern Recognition, 2013
7. J.R. Quinlan, "*Simplifying Decision Trees*", International Journal ofMan-Machine Studies, Volume 27, Issue 3, Pages 221-234, 1987
8. C. Jin, L. De-lin, M. Fen-xiang, "*An Improved ID3 DecisionTree Algorithm*", 4th International Conference on Computer Science& Education, 2009
9. O. L. Mangasarian, D. R. Musicant, "*Active Support Vector Machine Classification*", Advances in neural information processingsystems, 2000
10. A. Tzotsos, D. Argialas, "A Support Vector Machine Approach for Object Based Image Analysis", 1st International Conference on Objectbased Image Analysis, 2008
11. V. Garcia, E. Debreuve, M. Barlaud, "*Fast k Nearest Neighbor Search using GPU*", 2008, arXiv:0804.144824
12. J. M. Keller, M. R. Gray, J. A. Givens, "A Fuzzy KNearest Neighbor Algorithm", IEEE Transactions on Systems, Man, and Cybernetics, Volume: SMC-15 , Issue: 4, 1985
13. C. J. Peng , K. L. Lee, G. M. Ingersoll, "*An Introduction to Logistic Regression Analysis and Reporting*", The Journalof Educational Research Volume 96, - Issue 1, 2002
14. I. Rish "*An empirical study of the naive Bayes classifier*", abnormal behavior, IJCAI Work Empirical Methods Artif Intell, 2001
15. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "*YouOnly Look Once: Unified, Real-Time Object Detection*", 2015,arXiv:1506.02640