VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering (SCOPE)
B.Tech-Artificial Intelligence and Machine Learning

<u>Fall Semester 2022-23</u>

November, 2023

A project report on

Real-time CCTV Surveillance Footage Violence Detection using Deep Learning

submitted in partial fulfillment for the J-Component project of

CSE4059 – Cognitive Systems

By

ROHAN PAWAR (20BAI1201)

RATNESHWAR (20BAI1192)

HARDIK KATHURIA (20BAI1048)

IBRAHIM AHMED SIDDIQUI (20BAI1189)

Real-time CCTV Surveillance Footage Violence Detection using Deep Learning

Index

Abstract

With the rapid growth of surveillance cameras to monitor the human activity demands such system which recognize the violence and suspicious events automatically. In smart buildings, this technology is mandatory. Violence Detection in real time will prevent any mishappenings from taking place by raising alerts to the security guards. Our model also has YOLO object detection inculcated in it in order to recognize any weapons or tools that an attacker may be using against people and the security guards will act accordingly.

Introduction

Nowadays, there has been a rise in the amount of disruptive and offensive activities that have been happening. Due to this, security has been given principal significance. Public places like shopping centers, avenues, banks, etc. are increasingly being equipped with CCTVs to guarantee the security of individuals. Subsequently, this inconvenience is making a need to computerize this system with high accuracy. Since constant observation of these surveillance cameras by humans is a near-impossible task. It requires workforces and their constant attention to judge if the captured activities are anomalous or suspicious. Hence, this drawback is creating a need to automate this process with high accuracy. Moreover, there is a need to display which frame and which parts of the recording contain the uncommon activity which helps the quicker judgment of that unordinary action being unusual or suspicious. Therefore, to reduce the wastage of time and labor, we are utilizing deep learning algorithms for automated violence detection system. Its goal is to automatically identify signs of aggression and violence in real- time, which filters out irregularities from normal patterns. We intend to utilize different Deep Learning models (CNN) to identify and classify levels of high movement in the frame. From there, we can raise a detection alert for the situation of a threat, indicating the suspicious activities at an instance of time.

Problem Statement

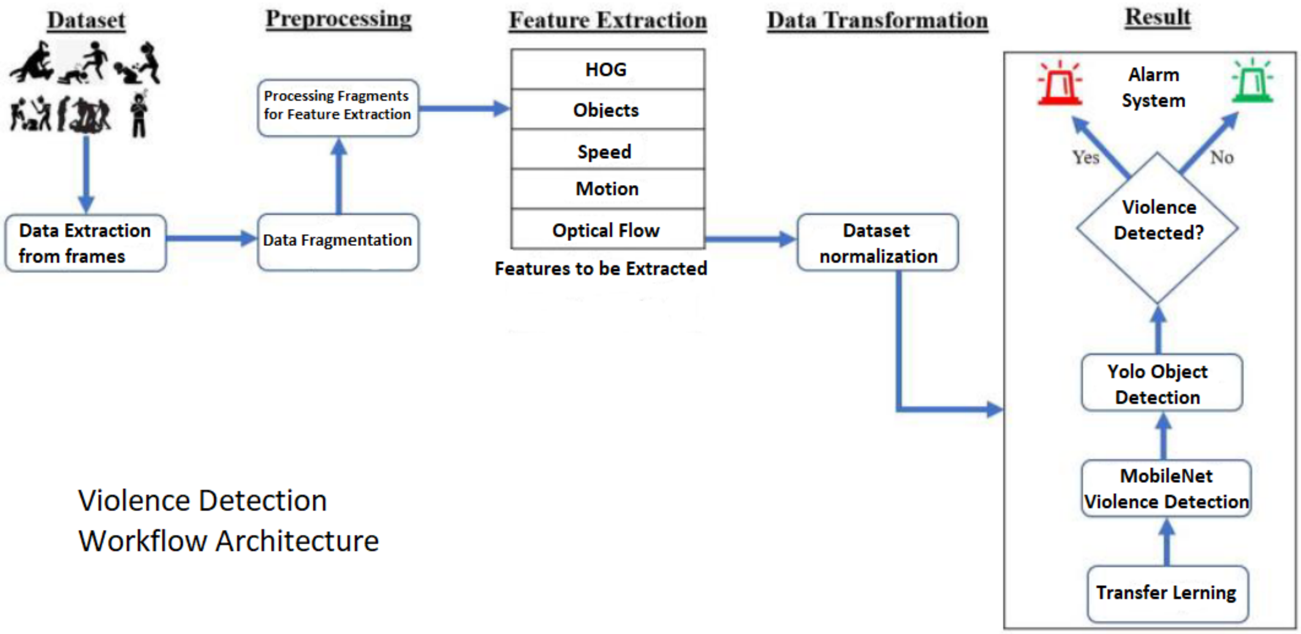
Violence detection has been of significant importance in developing automated video surveillance systems. Public video surveillance systems are common all over the world, being capable of providing accurate and rich information in many security applications. However, the need of watching hours of video footages undermines the chance to take decisions in a short time, which is essential in video surveillance for crime and violence prevention. In this regard, violence detection has its importance in the modern world, with the aim to unburden authorities from the need of watching hours of videos to identify events lasting few seconds. For this purpose, deep learning techniques have been proven effective in extracting spatial-temporal features from videos, i.e., features that represent the motion information contained in a sequence of frames, in addition to the spatial information contained in a single frame.

Background Study

Violent activities can be recognized by analysing human actions. Such activities usually involve specific motion patterns and take place in relatively low time intervals. The most important information needed to recognize and characterize human physical actions and interactions is motion. A major advantage of DNN based solutions is that these networks discover automatically optimized features needed to solve the problem at hand. However, there are two drawbacks of this approach. One is the need of very large databases for proper training. The other one is the hardware resource needed for real time operation. Both are critical in the case of our application. A good survey of the work in activity recognition, which focuses on DNN solutions for recognition of human based gestures.

One possible approach to capture the time domain evolution of activities is to feed the DNN with a set of consecutive frames.

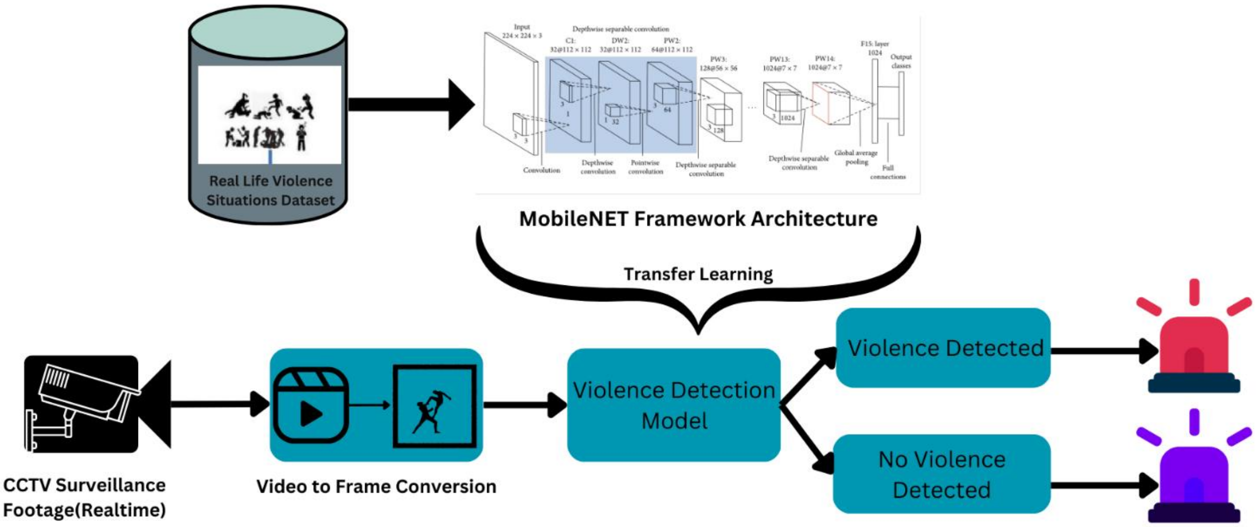
Methodology



Violence Detection
Workflow Architecture

MobileNet is a streamlined architecture that uses depth wise separable convolutions to construct lightweight deep convolutional neural networks and provides an efficient model for mobile and embedded vision applications. Transfer learning will be applied to this to train the model for Violent activities detection as well as object detection.

Architecture



CNN- CNN stands for Convolutional Neural Network, which is a type of neural network commonly used in deep learning tasks such as image recognition, object detection, and segmentation. CNNs are designed o automatically learn spatial hierarchies of features from raw input data. The main building blocks of a CNN are convolutional layers, pooling layers, and fully connected layers. After several convolutional and pooling layers, the output is passed through one or more fully connected layers, which perform classification or regression based on the learned features. These layers are similar to those in a traditional feedforward neural network, where each neuron in the layer is connected to all the neurons in the previous layer.

- **Convolutional Layers-**

Convolutional layers use filters or kernels to perform a convolution operation on the input data and extract features such as edges, corners, and textures.For advanced feature extraction, some people also prefer to use HOGs, SIFT and SURF methods.

- **Pooling Layers-**

Pooling layers reduce the spatial dimensions of the output by downsampling the input, which helps to reduce the number of parameters and computational cost of the network.

- **Fully Connected Layers/Dense Layers-**

Fully connected layers are used to make predictions based on the learned features and are typically placed at the end of the network. CNNs are trained using backpropagation with gradient descent to minimize a loss function and optimize the network's parameters.

MobileNet- MobileNet is a convolutional neural network architecture that is designed for efficient and lightweight image classification tasks on mobile and embedded devices. It was first introduced in a 2017 research paper by Google researchers Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNet is designed to be highly efficient by using depthwise separable convolutions, which are a type of convolutional operation that splits a standard convolution into two separate operations: a depthwise convolution and a pointwise convolution. In the depthwise convolution, a single filter is applied to each input channel separately, while in the pointwise convolution, a 1x1 convolution is applied to combine the output of the depthwise convolution.

By using depthwise separable convolutions, MobileNet is able to significantly reduce the number of parameters and computations required for violence and non-violence classification, while still maintaining a high level of accuracy.

Hardware and Software Components

HARDWARE COMPONENTS

1. **Laptop Camera-**

For capturing realtime video, laptop camera was used.

2. **Speaker-**

The alarm system consists of beeps of 1000 milliseconds burst with a frequency of 400 Hertz, which was played through the speakers.

3. **GPU-** GPU used to run the project was Nvidia GeForce GTX 1650 TI.

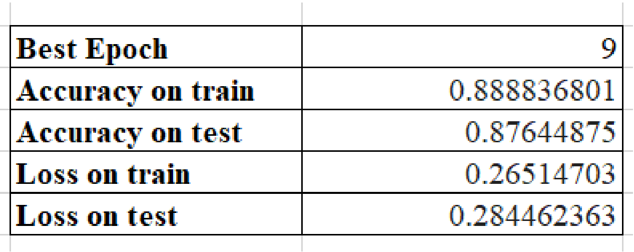
SOFTWARE COMPONENTS

1. **Python**
2. **Jupyter Notebook**
3. **Google Collab**

□

Results

Because of limited space allotted on Google Collab, I was only able to train the model with 700 videos (350 violence, 350 non-violence). So, the accuracy of the model is 89% but it can be 95-99%. If we are able to train it with 2000 videos.



Best Epoch	9
Accuracy on train	0.888836801
Accuracy on test	0.87644875
Loss on train	0.26514703
Loss on test	0.284462363

Table 1- Evaluation Matrix

The training loss indicates how well the model is fitting the training data, while the validation loss indicates how well the model fits new data. The training and validation loss and training and validation accuracy curves are given below:

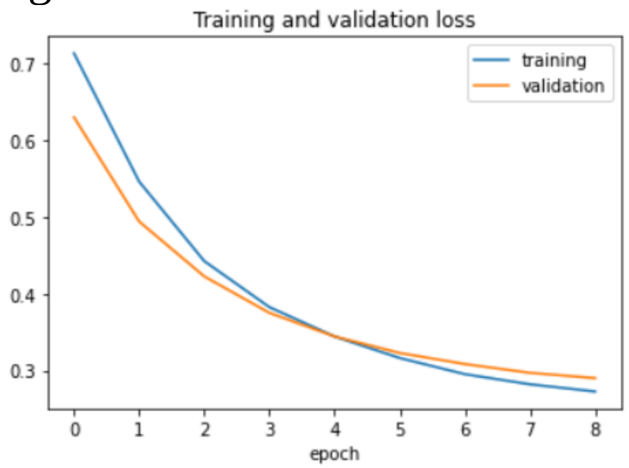


Figure 4- Training and Validation Loss Curve

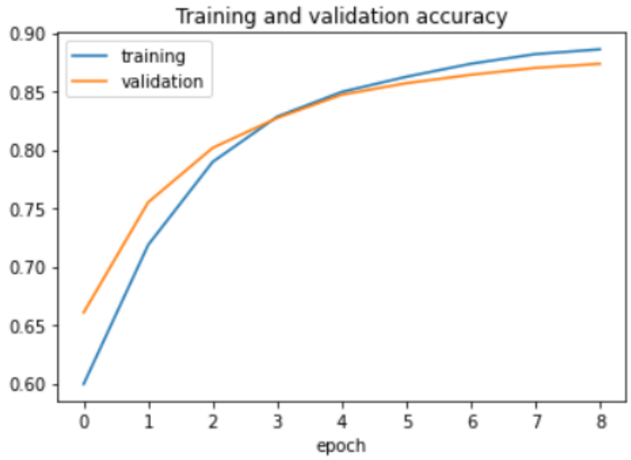
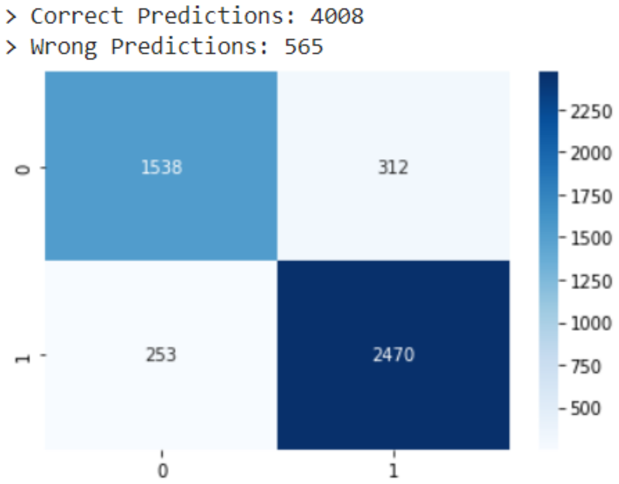


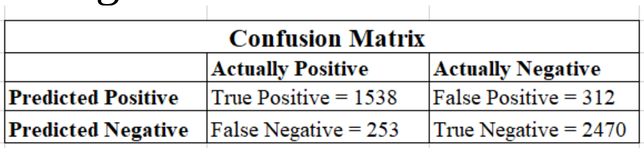
Figure 5- Training and Validation Accuracy Curve

Confusion Matrix below depicts the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes. The confusion matrix of the violence detection model prediction is given below:



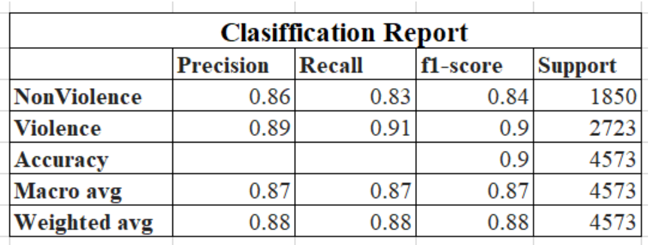
```
> Correct Predictions: 4008
> Wrong Predictions: 565
```

Figure 6- Confusion Matrix



Confusion Matrix		
	Actually Positive	**Actually Negative**
Predicted Positive	True Positive = 1538	False Positive = 312
Predicted Negative	False Negative = 253	True Negative = 2470

Table 2- Confusion Matrix Values



Clasiffication Report

	Precision	Recall	f1-score	Support
NonViolence	0.86	0.83	0.84	1850
Violence	0.89	0.91	0.9	2723
Accuracy			0.9	4573
Macro avg	0.87	0.87	0.87	4573
Weighted avg	0.88	0.88	0.88	4573

Table 3 – Classification Report

Conclusion

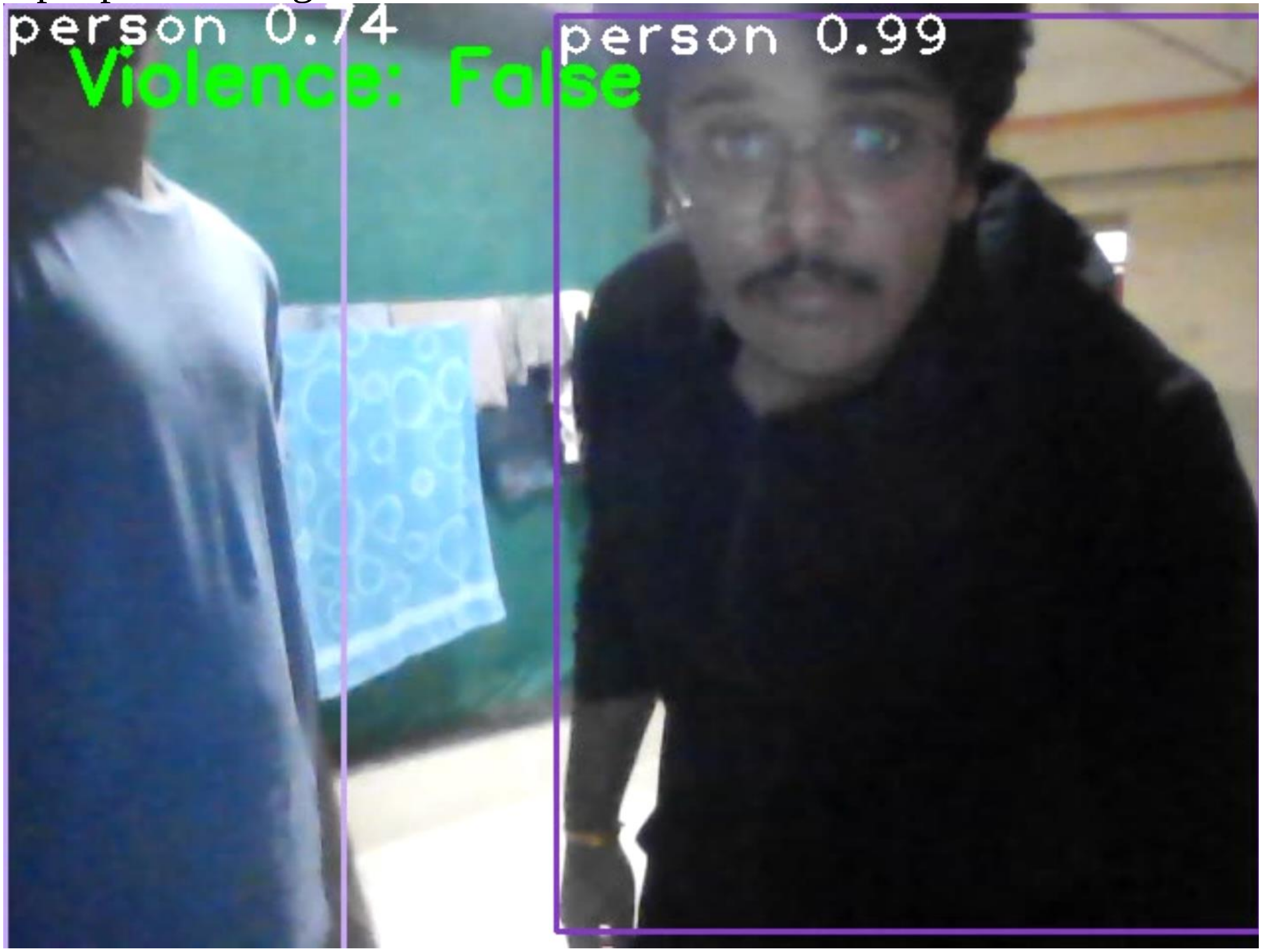
Artificial intelligence is crucial for spotting violence in video footage. Violence detection is still an issue even though it appears to be complicated at the moment.

With this paper, we aimed to demonstrate how violence detection is possible and how it can be easily implemented using the most basic techniques currently available, all due to the ongoing developments in deep learning and AI. These models have learned how to extract dynamic and practical features from everyday images and use them as a jumping-off point for learning new tasks and also alerts when violence is detected and produce the beep sound, so that the nearest authority can take the action according to the situation. After being trained on more than a million images, these networks can categorize images into violent and non violent classes. Transfer learning was used with pre-trained networks because it is frequently simpler and faster than creating or training a network from inception.

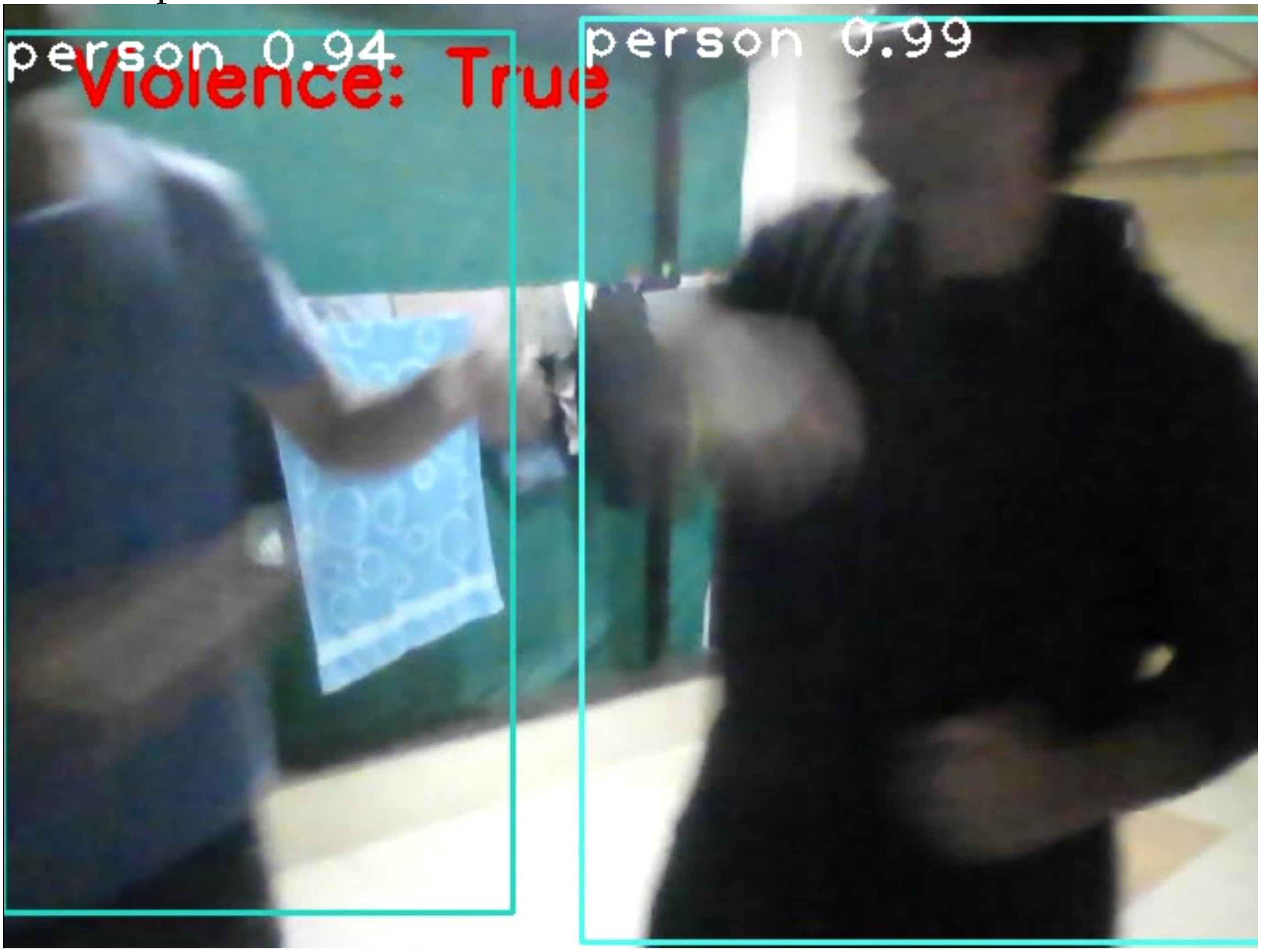
Appendix

Screenshots-

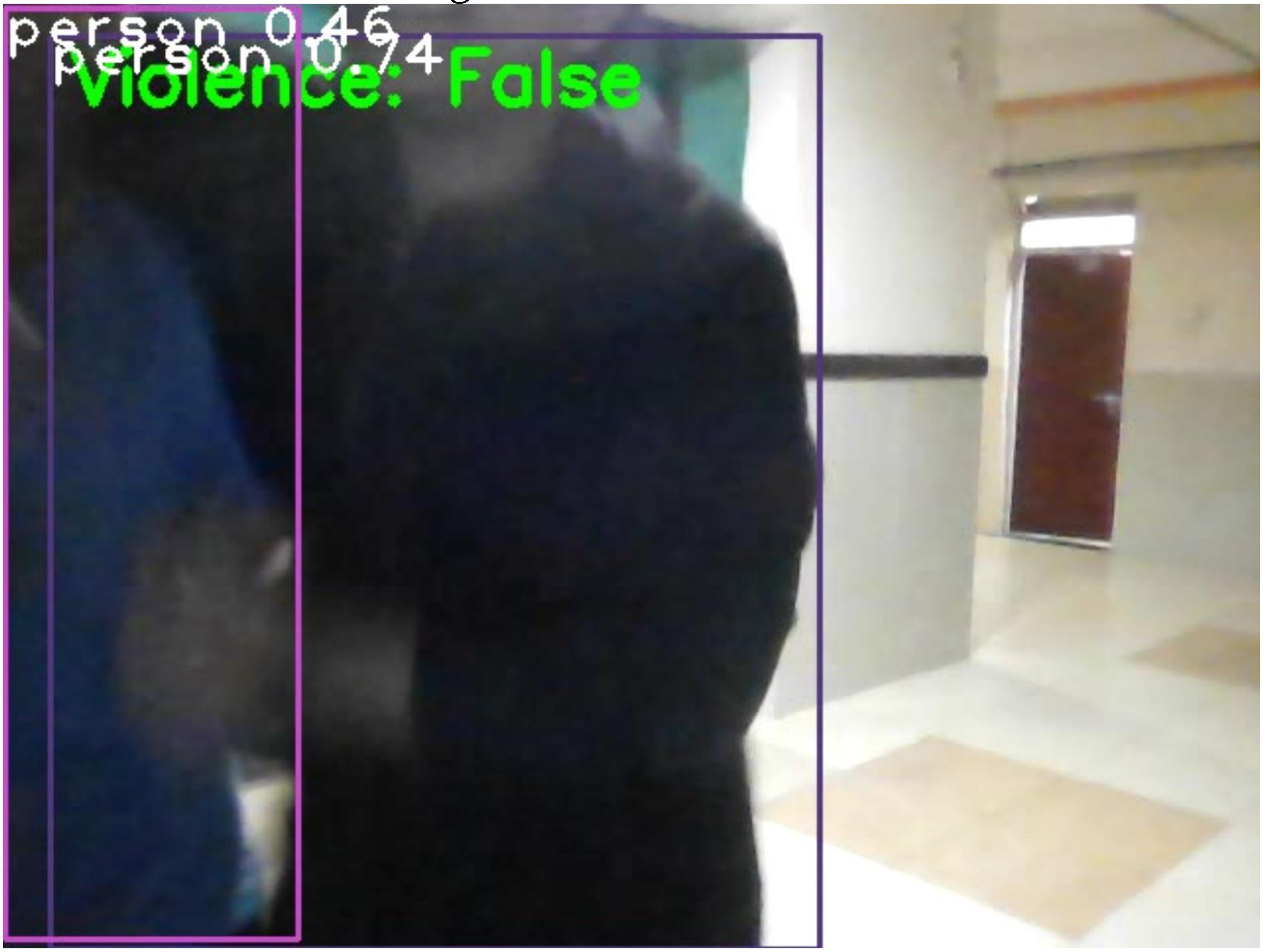
2 people standing-



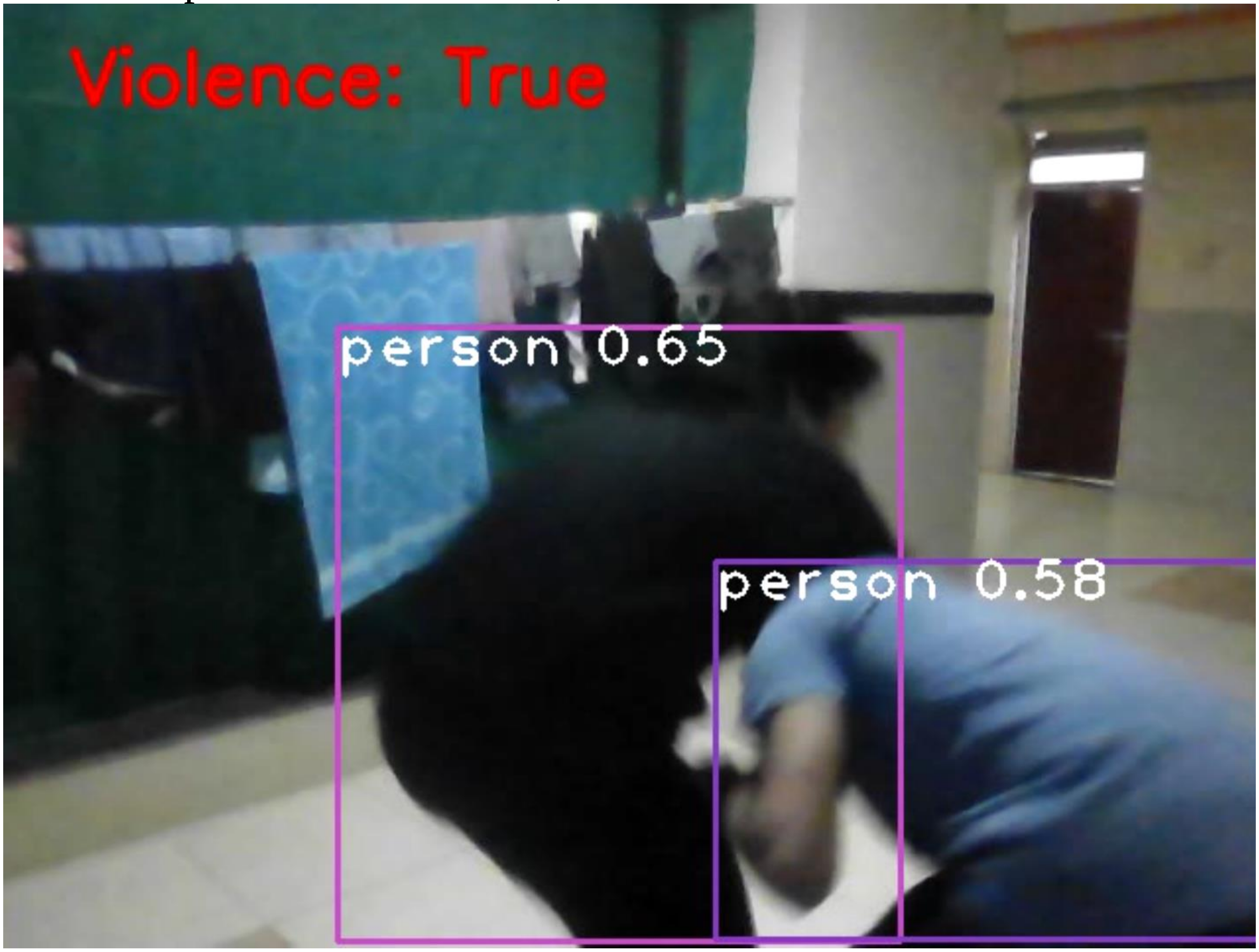
After slap-



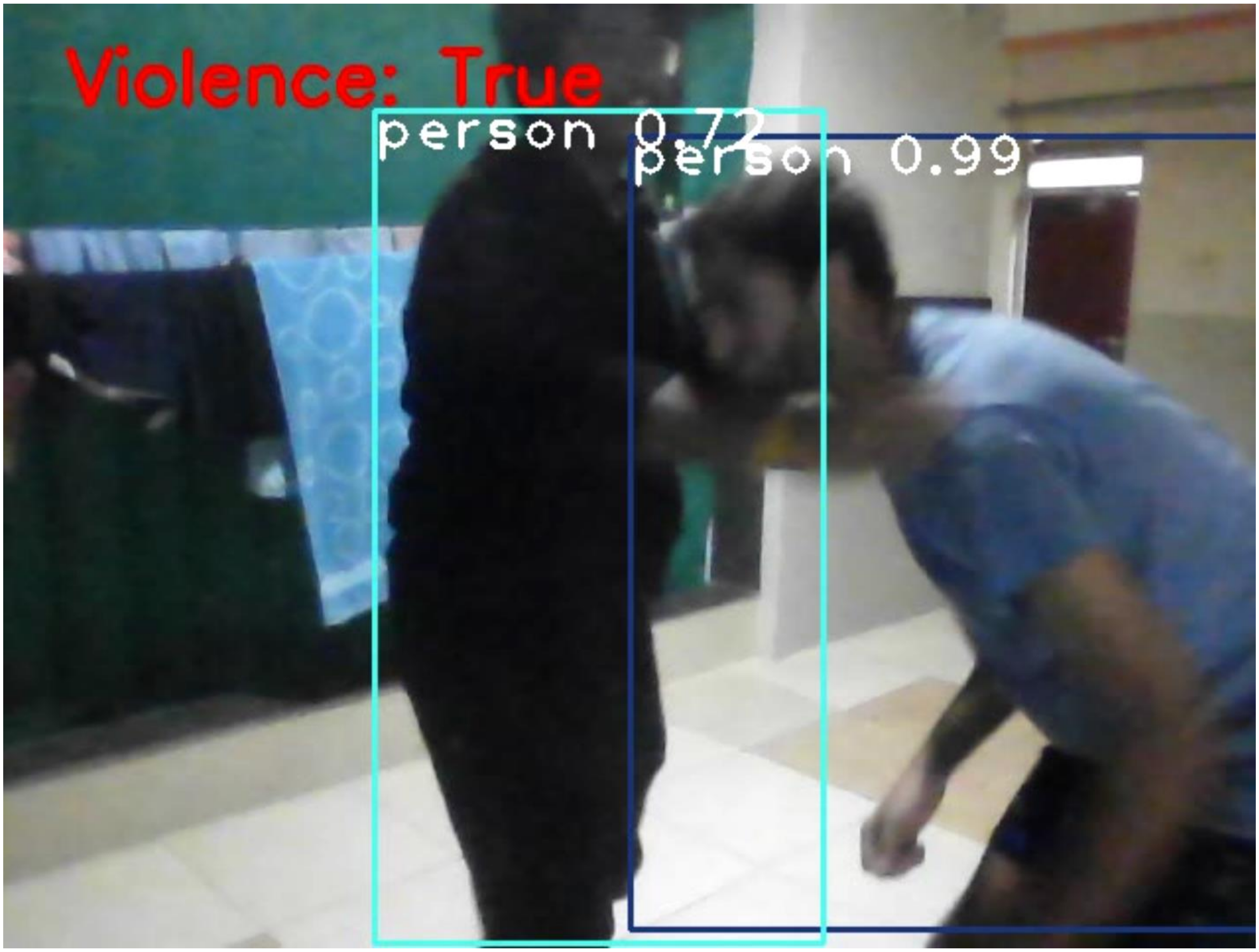
Model does not detect hug as violence-



When the person is taken down, violence is detected-



Choke-



When the person is hit with an object, in this case a bottle, violence is detected-