A deep learning approach to predict the nature of security camera footage with the help of posenet architecture

Shrayan Banerjee, Vamshi Krishna Reddy, Abhinav Upadhyay, S. Arvinda Krishnan

Student, Student, Student, Assistant Professor,
Department of Computer Science and Engineering,
SRM Institute of Science and Technology, Chennai, India.

Abstract : In the field of surveillance nowadays most of the footage gathered is just stored on the various types of storage media without analyzing the footage in real time and thus the detection and the subsequent analysis is a time consuming and in the most cases yield inconclusive outcomes .

The aim of the project is to provide real time information to the stakeholder about the security footage and warn him beforehand of any kind of anomaly if its taking place at his premises . Neural Networks have been the at the core of solving complex problems. The system obtained is a system that detects unwarranted and abusive behaviour from a video stream such as the one provided by the security camera .

The neural network model utilizes the pose-net architecture to derive the results from the fed in images and then classifies them as violent or non violent according to the training we have provided and the system also have a method to take innew images to add to its database.

Index Terms **–Convolutional Neural Networks, Violence Detection , Surveillance , Machine Learning, Computer Vision, Artificial Intelligence , Classification .**

I.INTRODUCTION

In the paper we first take into account the basic metric to obtain an estimate to the position of people in the video . Using a 2d position estimation or solve the problem of locating the fundamental key-points . It is focused on mapping the individual parts of the body.

The complexity arises when there are multiple people in a stream of image frames. The input may consists of n number of people with different positions . The second challenge would be that different people in the images may introduce some noise such as due to interactions, closure , superimposition or handicapped people in the frame. The third and the most difficult complexity to overcome is the run time time-complexity with increasing with the number of individuals.

The techniques to approach the basic problem for the individual person detection encounters problems such as confidence for certain persons over other persons if they are at close distances to each other the complexity of the problem is directly proportional to the number of persons . The suggested approach tries to separate the runtime time complexity with the number of people in the image instead take help from the global environment variables to estimate the environment conditions faster.

1.1 Aim : To detect violence from camera footage

The aim is to develop a system capable of detection of unwarranted and violent behavior from the camera feed the system will take in the images in a sequence and detects the people concerned in the scene at first as the input and then it segregates each individual in the image then assigns the joint positions of the concerned individual’s and then estimates the position of the joints for all the people concerned in the image then estimate the relative pose co-ordinates for the individuals then identifies if the scene is violent or not .

1.2 Components

For a system to work perfectly in achieving this, there are some prerequisites that must be met. Amongst those are image extraction, Hand position estimation , Head position estimation, Human pose estimation and Naïve Bayes Classification for violence detection . Considering the time complexity involved with estimation of each individual we arrived at the conclusion that naïve Bayes was the most suitable classification algorithm to solve our given problem .

1.2.1 Image Extraction

The technique of image extraction that we have used is through the PIL(Python Imaging Library) that takes in an image of the dimensions of the image as 640 x 480 into our python script and the web app and then applying pose estimation algorithms as discussed below to generate the 2d skeleton of the humans in the context.

1.2.2 Hand Position Estimation

We take the method as discussed in the paper [4] to estimate 2d pose from one depth based map into a voxel possibility our used network estimates each voxel similarity from the voxel processed input. We predict the usability of the I/O representations by evaluating the changes the input and the output. We test our algorithm of the predefined training dataset of HANDS 2017 frame based challenge and obtain a significant degree of accuracy. This can have applications in VR and AR technologies as well.

To generate the 2d likelihood for each of the keypoints we generate a 2d heatmap , where the mean gaussian peak is located at

$$H_n^*(i,j) = \exp(-\frac{(i-i_n)^2+(j-j_n)^2}{2\sigma^2})$$

ground-true location as follows:

$$L1 = \sum_{k=1}^{K}\sum_{p,q}\left\|H_k^*(p,q) - H_k(p,q)\right\|^*$$

We use mean squared loss as our loss function as as mentioned below :

1.2.3 Head Pose Estimation

We use convolutional neural networks to solve the problem of head pose estimation , availability of a huge training set of images is a pre-requisite for such a network to perform well . We fortunately now have the pre-annotated MPII2 dataset that contains various labels for various types of activities as the labels. In the last layer of the neural net we use a softmax layer to estimate the output of the neural network as discussed in [5]. The softmax loss function maximizes the effective probability between the estimated values of the network and the base truth that is given by the labelled dataset. Therefore a convolutional neural network is used to train the head pose estimation in the network. We have a given moment X with a distribution of labels y,x=φ(X;0) would be the activation function in the last layer in a deep convolutional neural network. We use a softmax activation function to turn this

$$\hat{y} = \frac{\exp(x_j)}{\sum_t \exp(x_t)}$$

into a probability distribution y^ that is similar to y.

As we have a given dataset D the goal is to find out θ to generate a probability distribution to that of y . The difference is

$$\theta^* = \operatorname*{argmin}_{\theta} \sum_k y_k \ln \frac{y_k}{\hat{y}_k} = \operatorname*{argmin}_{\theta} - \sum_k y_k \ln \hat{y}_k$$

$$\hat{y}$$

calculated as the measurement of similarity between the best parameter θ* is determined by :

$$T = -\sum_k y_k \ln \hat{y}_k$$

The loss function is as follows:

The we use stochastic gradient descent for our problem:

1.2.4 Human Pose Estimation

We take the inputs of the above mentioned methods to estimate the human pose based on the given key-points and construct the 2d skeleton of the humans in the picture as per the paper[3] and then estimate the position according to the labels as mentioned in the mpii2 dataset as violent for activities like boxing , archery etc and assign them a score of 1 and for others activities we assign them a score of 0.

$$d = (.., d_i^t, ...)^t, i \in (1, ..., n),$$

To represent human position, we code the coordinate points of all n body joint co-ordinates in position vectors represented as where di contains the c and d coordinates of the ith joint. A image is labelled by (c, d) where c represents image data and d is the ground truth pose vector.

$$N(d_i; b) = \begin{pmatrix} \frac{1}{b_w} & 0 \\ 0 & \frac{1}{b_h} \end{pmatrix} (d_i - b_c)$$

As the human joint positions are in final image points , the points need normalization. Following which the joint d_i could be transformed using box center and scaled by the length of the box which can be referred to as normalization by b:

1.2.5 Gaussian Naïve Bayes For Violence Detection

Naive Bayes process is a set of supervised machine learning technique based on using Bayes theory with a naive belief of context based independence between any two features considering the value of the class variable. Bayes theorem is the following

Likelihood Class Prior Probability

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

Posterior Probability Predictor Prior Probability

$$P(c\,|\,X) = P(x_1\,|\,c) \times P(x_2\,|\,c) \times \cdots \times P(x_n\,|\,c) \times P(c)$$

relationship, dependent feature vector through x and given class variable y;

When working on time series data, a common belief is that the time series data corresponding to every distinct class is divided into each considering a Gaussian probability distribution. For example, consider the dataset being trained consists of continuous variables, x. Initially separate the dataset by classes, then evaluate variance as well as average of x in distinct classes. Class C_k is associated by μ_k, the average of x, and σ^2_k,the variance of x. The probability distribution of observation values v given a class C_k,

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2_y}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma^2_y}\right)$$

$p(x = v=C_k)$, can be evaluated using Normal distribution function represented by μ_k and σ^2_k. i.e.,

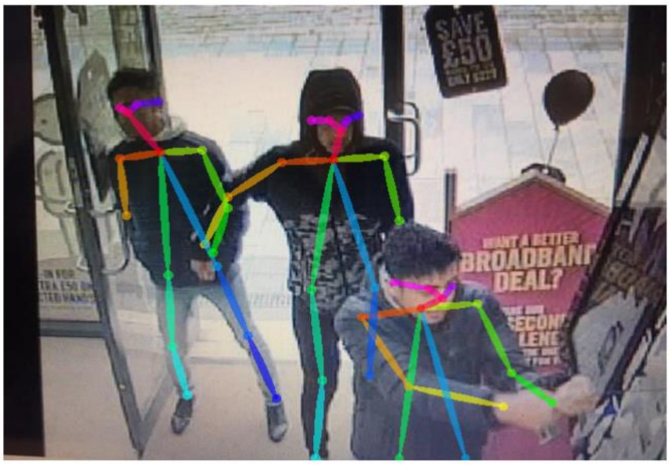
In this project, we used Gaussian naive bayes because the features in the data were continuous and they are independent of each other. For example, the hand pose values are independent of head pose values.

II.EXPERIMENTS

2.1 Image Extraction

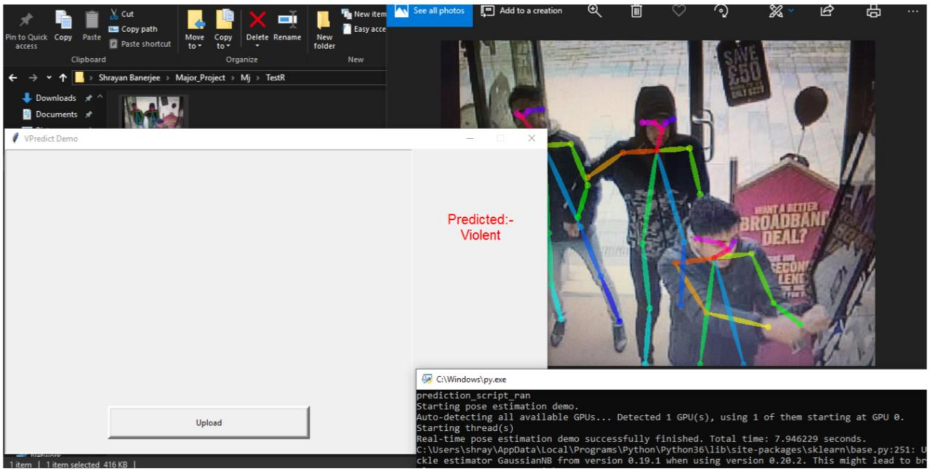
THE FRAMES ARE EXTRACTED FROM A VIDEO AND CONVERTED THEM TO PNG FILES. AND THEN THE IMAGES ARE USED TO GENERATE POSE ESTIMATION.

2.2 2D Skeleton Generation



The images extracted are trained through a open pose model and it renders a skeleton image and coordinates of that skeleton.

2.3 Result



Predicted:-
Violent

C:\Windows\py.exe

```
prediction_script_ran
Starting pose estimation demo.
Auto-detecting all available GPUs... Detected 1 GPU(s), using 1 of them starting at GPU 0.
Starting thread(s)
Real-time pose estimation demo successfully finished. Total time: 7.946229 seconds.
C:\Users\shray\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\base.py:251: U
ckle estimator GaussianNB from version 0.19.1 when using version 0.20.2. This might lead to br
```

The rendered skeleton coordinates are classified as a violent or non violent using naive bayes classifier.

III.ADVANTAGES

The advantages of such a system are vast. Most noticeable among these are:

1. The proposed system doesn’t need human intervention.
2. Can reduce the number of Security Personnel required .
3. Can log in reports anonymously and send the data across.
4. Automatic system to get alerts whenever a security threat is encountered
5. Gives the concerned individuals sufficient time to react to the situation.

IV.CURRENT LIMITATIONS

The biggest limitations to this project have been from the context of time-complexity. The time taken for processing an image on a laptop with a decent graphical processing unit is about averaging at 10 seconds therefore for a stream of images such as in a video captured at about 30 frames per second will increase the time taken exponentially.

From the context of storing footage and readily processing it would be mostly done on the cloud where we can exclusively dedicate memory and a cluster of graphical processing unit’s the chances for latency can be dealt with. And having a high core count of the cpu cores of server processors and further divide the workload for individual cores hence increasing performance .

V.ACKNOWLEDGMENT

I would like to thank Mr. S Arvinda Krishnan for being a guide and taking the time from his schedule to help us when we need it. Also, I’d like to thank SRM Institute of Science and Technology for giving us the opportunity to understand how to work in a research domain, which otherwise, we would have remained oblivious of.

REFERENCES

[1] Christian Schüldt ; Ivan Laptev ; Barbara Caputo. (2004). "Recognizing Human Actions: A Local SVM Approach ".

[2] Zhe Cao ; Thomas Simon ; Shih-En Wei ; Yaser Sheikh (2017); "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields"

[3] Toshev Alexander ; Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[4] Moon Gyeongsik ; Ju Yong Chang ; Kyoung Mu Lee. "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[5] Yang ; Xu, et al ; "Deep label distribution learning for apparent age estimation." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015.

[6] Amarjot Singh ; Devendra Patil ; SN Omkar ; "Eye in the Sky: Real-time Drone Surveillance System (DSS) for Violent Individuals Identification using ScatterNet Hybrid Deep Learning Network". IEEE Computer Vision and Pattern Recognition (CVPR) Workshops 2018

[7] S. Penmetsa ; F. Minhuj ; A. Singh ; S. Omkar ; "Autonomous uav for suspicious action detection using pictorial human pose estimation and classification. " ELCVIA: electronic letters on computer vision and image analysis, 13(1):18–32, 2014.

[8] T. Pfister ; K. Simonyan ; J. Charles ; A. Zisserman ; "Deep convolutional neural networks for efficient pose estimation in gesture videos." In Asian Conference on Computer Vision, pages 538–552, 2014.

[9] K. Goya ; X. Zhang ; K. Kitayama ; ; I. Nagayama. ; "A method for automatic detection of crimes for public security by using motion analysis " International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.

[10] A. Geiger ; F. Moosmann ; O. Car ; B. Schuster ; " Automatic camera and range sensor calibration using a single shot. In Robotics and Automation (ICRA) " 2012 IEEE International Conference on, pages 3936–3943, 2012.

[11] Hu ; Qiyang et al ; "Video Synthesis from a Single Image and Motion Stroke." arXiv preprint arXiv:1812.01874 (2018).