International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362
IJARST

LIVE TRACKING IN DETECTION OF WEAPONS IN SURVEILLANCE VIDEOS USING DEEP LEARNING

[1] S.Sushmitha , [2] Thakur Vaishnavi, , [3] Gunti Shresta, [4] Chagapuram Ashutosh Goud

[1] Assistant Professor, Department of Information Technology, Teegala Krishna Reddy Engineering College Hyderabad, Telangana, India.

[1] boggarapusushmitha@gmail.com

[2,3,4] UG Scholars Department of Information Technology, Teegala Krishna Reddy Engineering College , Hyderabad, Telangana, India.

[2] thakurvaishnavi1211@gmail.com , [3] guntishreshta26@gmail.com , [4] ashutosh.chagapuram@gmail.com

Abstract

Closed circuit television systems (CCTV) play a vital role in evidence collection against crimes and criminals. The existing systems does not classify normal and abnormal events leading the police to become more reluctant to attend the crime scenes unless there was a visual verification, either by manned patrols or by electronic images from the surveillance cameras. The Proposed work is being used for surveillance, monitoring and classifications of weapons, live tracking and many more purposes. Operations of proposed project has three processing modules, first processing module is for object detection using Convolutional Neural Networks(CNN) and second processing module will handle the classification of weapons, monitoring and alarm operations will be carried out by the third processing module.

1. INTRODUCTION

The use of weapons in public places has become a major problem in our society. These situations are more frequent in countries where weapons are legally purchased or their use is not controlled Crowded places are specially vulnerable. Unfortunately, mass shootings have become one of the most dramatic problems we face nowadays. Video surveillance systems, typically based on classic closed-circuit television (CCTV) are especially useful for intruder detection and remote alarm verification. However, these systems need to be continuously supervised by a human operator. In this respect, it is estimated that the concentration of a security guard watching a camera panel decreases catastrophically after 20 minutes. Security can be increased applying artificial vision algorithms on images obtained from video surveillance systems. Another advantage of these algorithms is the possibility of monitoring larger spaces using fewer devices thus requiring less dependence on the human factor. Machine learning techniques have been widely used in the field of video surveillance. The prevalent paradigm of deep learning has but increased the potential of machine learning in automatic video surveillance. The objective of this work is the development of two novel weapon detectors, for guns and knives, applying deep learning techniques and assess their performance.

1.1 PROBLEM STATEMENT

Design a system for recognising and detecting of weapons automatically in image or from the video and also real time through webcam or CCTV. Because it is real time and computer based human operator is not required and also human supervision is not required. This system can be public or private facility to restrict the weapons access.

1.2 OBJECTIVE

The terms “deep learning” and “machine learning” are frequently used interchangeably.

Deep learning is a subset of machine learning, both fields are a subset of artificial intelligence Deep learning involves the use of Deep Neural Network (DNN) which involved multiple processing layers. The example of the deep neural network consists of multiple layers between input and output layer. Each layer transforms the input signal to another feature space. Deep neural network model is inspired by the working mechanism observed in the human brain. Before it can be used, the model is trained using a large amount of data. Such training contributes towards excellent recognition accuracy in various applications. Convolutional neural network (CNN) is a type of deep neural network, mainly used for image and video.

2. LITERATURE SURVEY

The applications of the deep learning paradigm for weapon detection are still rather limited. Presented an automatic handgun detection system for video surveillance. This system was based on a Faster R- CNN with a VGG16 architecture trained using their own gun database. Results provided zero false positives, 100% recall and a precision (IoU=0.5) value of 84,21%. In Valledor a firearm detector for application to social media was presented. The detector employed a Faster R-CNN and an Inception v2 network for feature extraction. A public database of images containing several firearms was manually labelled and used for training. Benchmarking was performed on the COCO dataset obtaining a ROC curve that showed usable results. Internet Movie Firearm Database (IMFDB) to generate a handheld gun detector. For that purpose, a Faster R-CNN based on a VGG16 architecture was applied only for feature extraction. The best result achieved was 93.1% accuracy, using a Boosted Tree classifier. We have to note that the IMFDB dataset contains mostly profile images of pistols and revolvers at high resolution with homogeneous background, which is not a realistic situation. The presented a detection and classification system for X- ray baggage security imagery. The work explored the applicability of multiple detection approaches based on sliding window CNN, Faster R-CNN, Region based Fully Convolutional Networks. Their system was composed by images divided into six classes: camera, laptop, gun, gun component, knife and ceramic knife. For knife cases, the best results were obtained using a Faster R-CNN based on a ResNet-101 architecture with a 73.2% AP50. Detection results obtained a 95.73% of sensitivity, 97.30% of specificity, 96.26% of accuracy and 70% of mAP50. Regarding knife detection, the most relevant results have been obtained in the context of the COCO (Common Objects in Context) Challenges. COCO is a large-scale object detection dataset focused on detecting objects in context. Each year COCO launches a challenge based on any of the following artificial vision tasks: detection, segmentation, key points or scene recognition. The last object detection challenge using bounding boxes was released in 2017 where the best result for knife detection was obtained by the Intel Lab team. Employing a Faster R-CNN and a Hypernet architecture this team achieved 36.6% AP50. Knife detection was explored using a dataset of 8,527 infrared (IR) images.

A Google Net architecture was applied to classify IR images as person or person carrying hidden knife. The classification accuracy reported was 97.91%. In summary, the Faster R-CNN seems to be the prevalent deep architecture for gun and knife detection. This work also focuses on that architecture.

3. PROBLEM STATEMENT

Different approaches then used for weapon detection using sliding window and region proposal algorithms. HOG (Histogram of oriented Gradient) models were used to predict the objects in the frame. HOG significant work used low-level features, discriminative learning, and pictorial structure along with SVM. These algorithms were slow for real-time scenarios with 14s per image. Although these classifiers gave good accuracies, the slowness of the sliding window method was a big problem, especially for the real-time implementation purpose.

Disadvantages

Weapon detection in real-time is a very challenging task. As our desired object has a small size so, detecting it in an image is also very challenging in presence of other objects, especially those objects that can be confused with it. Deep learning models faced several below mentioned challenges for detection and classification task: The first and main problem is the data through which

CNN learn its features to be used later for classification and detection. No standard dataset was available for weapons. manually was a very long and time-consuming process. Labelling the desired database is not an easy task, as all data needs to be labelled manually. Different detection algorithms were used, so a labeled. dataset for one algorithm cannot be utilized for the other one. Every algorithm requires different labelling and pre-processing operations for the same-labelled database. As for real-time implementation, detection systems. require the exact location of the weapon so gun blocking or occlusion is also a problem that arises frequently and it could occur because of self, interobject, or background blocking.

4. PROPOSED SYSTEM

We propose an initial approach to systems designed for knife and firearm detection in images, respectively. In this work, we summarize this effort and present the current versions of the algorithm. Even if different methods are also used, the algorithms presented in this paper aim towards a similar goal; our motivation is to solve the problem of knife or firearm recognition in frames from camera video sequences. The aim of these approaches is to provide the capability of detecting dangerous situations in real life environments, e.g., if a person equipped with a knife or firearm starts to threaten other people. The algorithms are designed to alert the human operator when an individual carrying a dangerous object is visible in an image. Different approaches are used in this work for weapon classification and detection purpose but all have deep learning and CNN architecture behind them because of their state-of-the-art performance. Training from scratch took very much time so the Transfer learning approach was used and ImageNet and COCO (common objects in context) pre-trained models are used. Different datasets were made for classification and detection. For real-time purposes, we made our dataset by taking weapon photos from the camera, data was extracted manually from robbery CCTV videos, downloaded from imfdb (internet movie firearm database), data by university of Granada other online repositories. All the work has been done to achieve results in real-time. The main contributions of this work are: presentation of a first detailed and comprehensive work on weapon detection that can achieve detection in videos from real-time CCTV and works well even in low resolution and brightness because most of the work done earlier is on high definition training images but real-time scenario needs Realtime training data as well for better results, finding of the most suitable and appropriate CNN based object detector for the application of weapon detection in real- time .

CCTV video streams, making of a new dataset because real-time detection also needs real-time training data so we made a new database of 8327 images and pre-processed it using different OpenCV filters helped in detecting images in low brightness and resolution, introducing the concept of related confusion classes to reduce false positives and negatives, training and testing of our novel database on the latest state of the deep learning based classification and detection models. The main contributions of this work are: presentation of a first detailed and comprehensive work on weapon detection that can achieve detection in videos from real-time CCTV and works well even in low resolution and brightness because most of the work done earlier is on high definition training images but Realtime scenario needs real-time training data as well for better results, finding of the most suitable and appropriate CNN based object detector for the application of weapon detection in real-time CCTV video streams, making of a new dataset because real-time detection also needs real-time training data so we made a new database of 8327 images and pre-processed it using different OpenCV filters i.e. Equalized, Grayscale and clahe that helped in detecting images in low brightness and resolution, introducing the concept of related confusion classes to reduce false positives and negatives, training and testing of our novel database on the latest state of the deep learning based classification and detection models. To achieve high precision, increase number of frame per seconds and improve localization, we moved to the object detection and region proposal methods.

5.DESIGN

Next step is to bring down whole knowledge of requirements and analysis on the desk and design the software product. The inputs from users and information gathered in requirement gathering phase are the inputs of this step. The output of this step comes in the form of two designs; logical design and physical design. Engineers produce meta-data and data dictionaries, logical diagrams, data-flow diagrams and in some cases pseudo codes.

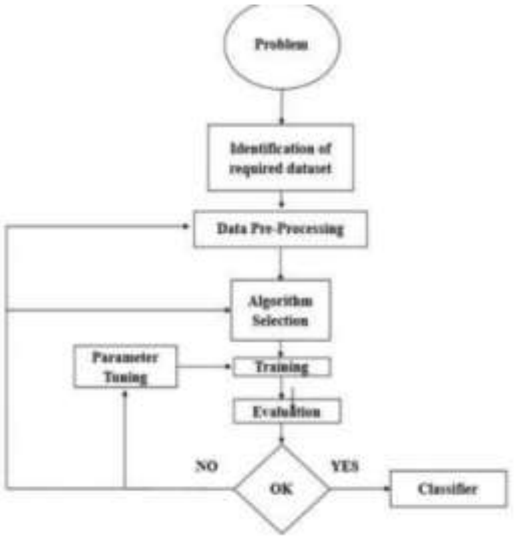


FIGURE 5.1 SYSTEM ARCHITECTURE

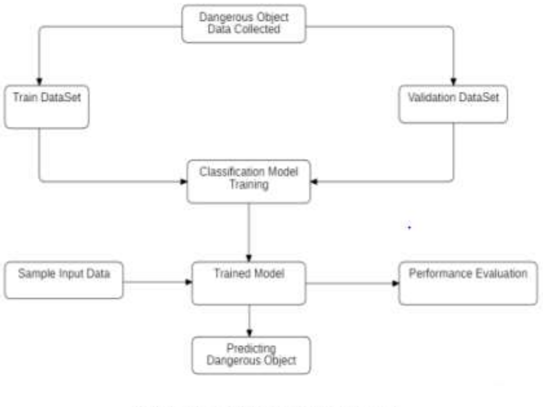

Dangerous Object Data Collected
Train DataSet
Validation DataSet
Classification Model Training
Sample Input Data
Trained Model
Performance Evaluation
Predicting Dangerous Object

FIGURE 5.2 DATA FLOW DIAGRAM

6. IMPLEMENTATION

6.1 CODING IMPLEMENTATION

The gun dataset has been extracted from [14]. The dataset is composed by 3,000 images of guns from different views and scenarios. In order to increase the accuracy of the detector, a data augmentation technique was applied to the dataset. The aim is to perform transformations that simulate realistic views of the object to be detected, see Figure 1: – Increasing brightness (10%) in order to simulate different illuminations – Image scaling to simulate different distances to the object – Mirroring and rotations (5o) to create different canonical views of the object With these transformations, the dataset was increased to a total of 15,000 images.

6.1.2 A OBJECT RECOGNITION

As the name suggests, it is the process of predicting the real class or category of an image to which it belongs by making probability high only for that particular class. CNN's are used to efficiently perform this process. Many state of the art Classification and Detection algorithms uses CNN as a backend to perform their tasks.

IMAGE CLASSIFICATION : The classification model takes an image and slide the kernel/filter over the whole image to get the feature maps. From the feature extracted, it then predicts the label based on the probability.

OBJECT LOCALIZATION: This method outputs the actual location of an object in an image by giving the associated height and width along with its coordinates. OBJECT DETECTION This task uses the properties of the aforementioned algorithms. The detection algorithm tells us the bounding box having x and y coordinates with associated width and height along with the class label. Non-max suppression is used to output the box with our desired threshold . This process gives the following results altogether: Bounding Box, Probability. In past object detection was very limited because of less data and low processing power of computers but with the passage of time the computing power of computers increased and world moved from CPU's to Graphic Processing Units (GPU). GPU's were firstly made for increasing the graphic quality of the systems and for gaming but later GPUs were used extensively for deep learning. In ImageNet, competitions started and contained about 1000 classes. This was the evolution of machine learning and deep learning.

6.1.3 CLASSIFICATION APPROACH

There are many ways to generate region proposals, but the simplest way of generating them is by using the sliding window approach. The sliding window method is slow because filter slides over the entire frame and has limitations, which were tackled by the region proposal approach, so we have the following two approaches used in our work for both classification and detection models are:

SLIDING WINDOW/CLASSIFICATION MODELS: In the method to the sliding window, a box or window is moved over a picture to select an area and use the object recognition model to identify each frame patch covered by the window. It is an exhaustive search over the whole picture for objects. Not only do we need to search in the picture for all feasible places, we also need to search on distinct scales. This is because models are usually trained on a particular range. The outcomes are in tens of thousands (104) of picture spots being classified. The sliding window method is computationally very costly because of the search with various aspect ratios and especially for each pixel of an image if the stride or step value is less

REGION PROPOSAL/OBJECT DETECTION MODELS

This technique takes an image as the bounding boxes of input and output proposals related to all areas in a picture most probable to be the object. These regional proposals may be noisy; coinciding not containing the object flawlessly , but there is a proposal among these region proposals related to the original target object. As this method takes a picture as the bounding boxes of input and output related to all patches in a picture most probable to be a category, so it proposes a region with the maximum score as the location of an object. Instead of considering all possible regions of the input frame as possibilities, this method uses detection proposal techniques to select regions . Region-based CNNs (RCNN) was the first detection model to introduce CNNs under this approach. The selective search method of this approach produces 2000 boxes having maximum likelihood. Selective search is a widely used proposal generation method because it is very fast having a good recall value. It is dependent on the hierarchical calculation of desired areas established on the compatibility of color, texture, size, and shape. Unlike the other region proposal-based methods it divides the input image into an SxS grid and then simultaneously predicts the probability and bounding boxes for an object with a center falling into a grid cell The general methodology used in training and optimization. It starts with defining a problem, finding the required dataset, applying pre-processing methods, and then finally training and evaluating the dataset. If the evaluation is correct then we save those weights as a classifier but if it's incorrect then comes the process of back propagation algorithm along with the gradient descent algorithm. In backpropagation, weights are optimized by subtracting the partial derivative of cost function J(Ɵ) with a multiplier of the learning rate alpha from the old or previous weight value. Gradient descent is the main weight optimization algorithm. It is used as a base in all optimizers used for the modeling and it helps in converging the model and reaching the minima where we get the best and desired weights values.

7. ALGORITHM

Evolution of R-CNN objects detectors:

The Regions-CNN method was developed in 2014. The processing of a R-CNN can be divided into three steps [8]. Firstly, an algorithm called selective search generates approximately 2,000 region proposals (or regions of interest). These region proposals are independent divisions of the image where an object could be located. Secondly, a CNN extracts features individually from each region proposal. Finally, the object is classified using a Support Vector Machine (SVM) methodology. Region proposals are considered as positive when their Intersection over Union (IoU) measure against the ground truth exceeds an arbitrary value. Later, the object bounding box localization is calculated by overlapping the selected region proposals. One of the main disadvantages of the R-CNN was its slow execution time. Fast R-CNN was proposed in 2015 as an improvement of R-CNN. Fast R-CNN is twenty-five times faster than its predecessor mainly due to two modifications. Feature extraction is performed using a CNN on the whole input image. Region proposals are selected as in the R-CNN approach by an external selective search method and included in the last layers of In 2016 the Faster R-CNN method introduced a new region proposal extraction method called Regional Proposal Network (RPN) . The idea of a RPN is to take advantage of the convolutional layers to obtain region proposals directly. Consequently, a sliding window is applied on the CNN feature map in order to extract region proposals of different sizes. The RPN is not responsible for classifying localized objects, this task is subsequently carried out by a Fast R- CNN.

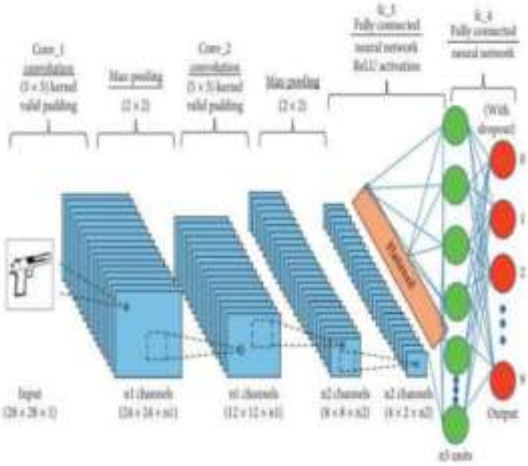
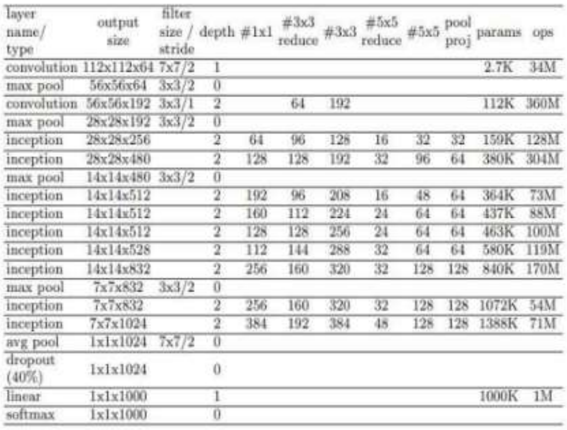

Conv_1 convolution (5 × 5) kernel valid padding
Max pooling (2 × 2)
Conv_2 convolution (5 × 5) kernel valid padding
Max pooling (2 × 2)
fc_3 Fully connected neural network ReLU activation
fc_4 Fully connected neural network
(With dropout)
Flattened
0
1
2
9
Input (28 × 28 × 1)
n1 channels (24 × 24 × n1)
n1 channels (12 × 12 × n1)
n2 channels (8 × 8 × n2)
n2 channels (4 × 4 × n2)
n3 units
Output

FIGURE 6.2.1 CNN ARCHITECTURE

Faster-CNN base architectures:

The CNN selected as Faster R-CNN base architecture should depend on its final purpose. Many CNNs employ a very deep architecture with the aim of obtaining a higher accuracy at a high computational cost. On the other hand, other architectures can be used that sacrifice precision in order to obtain models that can be integrated into embedded systems. In this work, GoogleNet and SqueezeNet CNN architectures have been tested and compared with the purpose of exposing their advantages and disadvantages for our task of weapon detection in video. GoogleNet The GoogleNet network is a CNN developed in 2014. This network demonstrated high accuracy for object detection in the ImageNet contest Large-Scale Visual Recognition Challenge 2014, being the winning architecture with a 6.66% error rate. The architecture of this CNN is mainly composed by Inception layers which are based on covering large image areas while keeping a high resolution for small areas with high feature density. For that, the network applies convolutions in parallel with different filter sizes. The GoogleNet architecture is composed by a total of 22 layers, Training using GoogleNet for the Faster R-CNN was carried out applying a stochastic gradient descent optimization algorithm with a momentum of 0.9 to accelerate gradient vectors, a L2 regularization method and an initial learning rate of 1e−3. The optimization was run for 30 epochs.

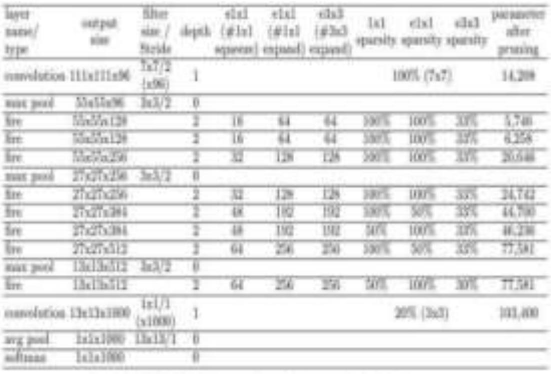


layer name/ type	output size	filter size / stride	depth	#1x1	#3x3 reduce	#3x3	#5x5 reduce	#5x5	pool proj	params	ops
convolution	112x112x64	7x7/2	1							2.7K	34M
max pool	56x56x64	3x3/2	0								
convolution	56x56x192	3x3/1	2		64	192				112K	360M
max pool	28x28x192	3x3/2	0								
inception	28x28x256		2	64	96	128	16	32	32	159K	128M
inception	28x28x480		2	128	128	192	32	96	64	380K	304M
max pool	14x14x480	3x3/2	0								
inception	14x14x512		2	192	96	208	16	48	64	364K	73M
inception	14x14x512		2	160	112	224	24	64	64	437K	88M
inception	14x14x512		2	128	128	256	24	64	64	463K	100M
inception	14x14x528		2	112	144	288	32	64	64	580K	119M
inception	14x14x832		2	256	160	320	32	128	128	840K	170M
max pool	7x7x832	3x3/2	0								
inception	7x7x832		2	256	160	320	32	128	128	1072K	54M
inception	7x7x1024		2	384	192	384	48	128	128	1388K	71M
avg pool	1x1x1024	7x7/2	0								
dropout (40%)	1x1x1024		0								
linear	1x1x1000		1							1000K	1M
softmax	1x1x1000		0								

Figure 6.2.2 Google Net Architecture

SqueezeNet:

The SqueezeNet network [9] is a CNN developed in 2016. The main goal of this network was the deployment of a small CNN architecture with fewer parameters instead of improving the accuracy. SqueezeNet achieved the same accuracy than AlexNet on the ImageNet dataset obtaining a model size with 50x fewer parameters. Therefore, it is a valuable alternative for embedded systems, field- programmable gate arrays (FPGAs) and other constrained systems. The SqueezeNet architecture follows three strategies to reduce the number of parameters while maintaining the accuracy level: Most convolutions replace 3x3 filters for 1x1 filters. As a consequence, the number of parameters is reduced 9x by convolution SqueezeNet applies fire modules to achieve the previous strategies. A fire module is composed by a SqueezNet convolution layer (1x1 filters) and an expand layer (mixture of 1x1 and 3x3 convolution filters). Three parameters are included in a fire module: s1x1 (from squeeznet layer), e1x1, and e3x3 (from expanded layer). All are related to the number of filters used in these layers. Fire module sets that s1x1 must be less than the sum of e1x1 and e3x3, so the squeeze layer helps to limit the number of input channels to the 3x3 filters. The SqueezeNet architecture is composed by a total of 13 layers In this approach, training using a SqueezeNet for the Faster R-CNN was carried out applying a stochastic gradient descent optimization algorithm with a momentum of 0.9 to accelerate gradient vectors, a L2 regularization method and an initial learning rate of 1e−4. As with the GoogleNet approach, 30 epochs were used to train the classifier

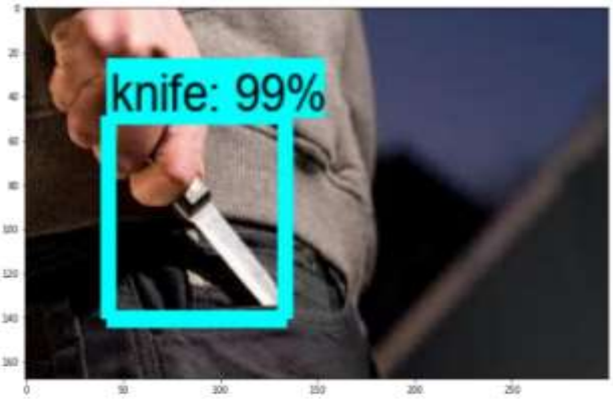


layer name/ type	output size	filter size / Stride	depth	s1x1 (#1x1 squeeze)	e1x1 (#1x1 expand)	e3x3 (#3x3 expand)	1x1 sparsity	e1x1 sparsity	e3x3 sparsity	parameter after pruning
convolution	111x111x96	7x7/2 (x96)	1					100% (7x7)		14,208
max pool	55x55x96	3x3/2	0							
fire	55x55x128		2	16	64	64	100%	100%	33%	5,746
fire	55x55x128		2	16	64	64	100%	100%	33%	6,258
fire	55x55x256		2	32	128	128	100%	100%	33%	20,646
max pool	27x27x256	3x3/2	0							
fire	27x27x256		2	32	128	128	100%	100%	33%	24,742
fire	27x27x384		2	48	192	192	100%	50%	33%	44,700
fire	27x27x384		2	48	192	192	50%	100%	33%	46,236
fire	27x27x512		2	64	256	256	100%	50%	33%	77,581
max pool	13x13x512	3x3/2	0							
fire	13x13x512		2	64	256	256	50%	100%	30%	77,581
convolution	13x13x1000	1x1/1 (x1000)	1					20% (3x3)		103,400
avg pool	1x1x1000	13x13/1	0							
softmax	1x1x1000		0							

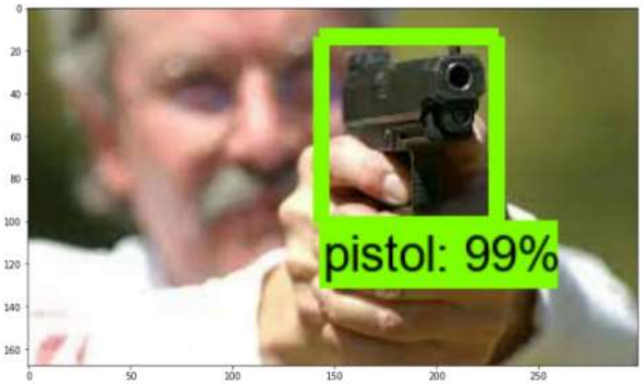
FIGURE6.2.3: SQUEEZENET ARCHITECTURE

8. RESULTS

DETECTION OF WEAPONS FROM IMAGES

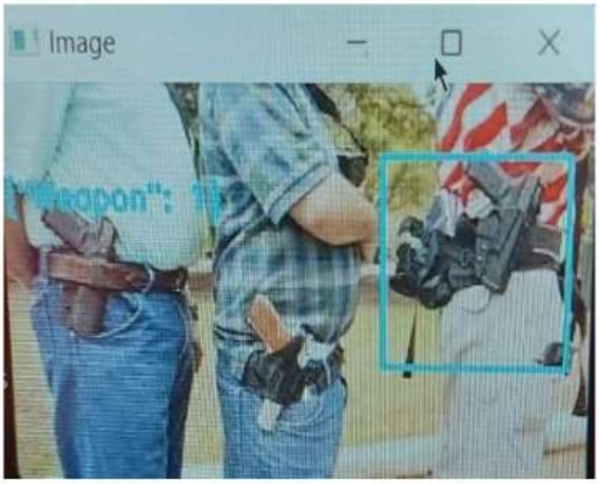


KNIFE DEDECTION

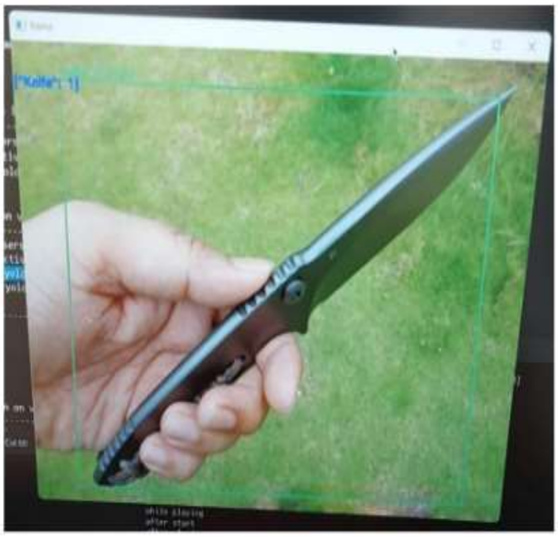


PISTOL DEDCTION

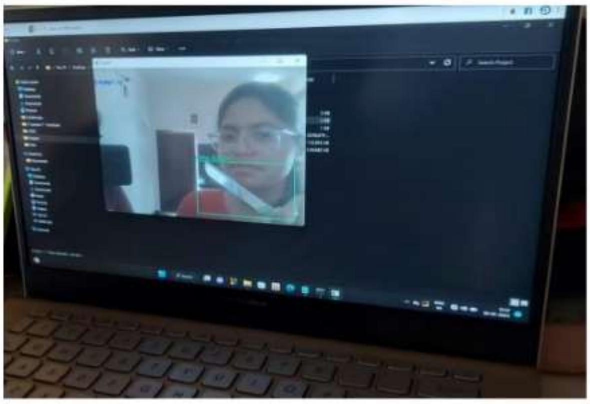
IN REAL TIME WEAPONS DETECTIONS



GUN DETECTION

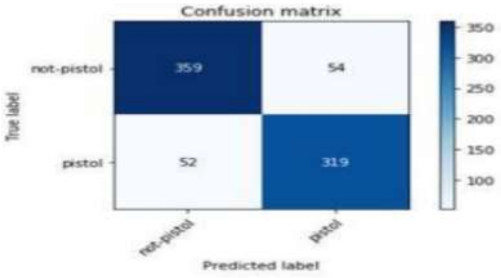


KNIFE DETECTION



REAL TIME IDENTIFIYING KNIFE AT HOME

NO DECTION OF WEAPONS



CONFUSION MATRIX

9. CONCLUSION

Public and crowded areas are still the target of many violent acts. Video surveillance can be helped by automatic image analysis using artificial vision. This paper describes the implementation of several weapon detectors for video surveillance based on Faster R-CNN methodologies. For training, gun and knife images COCO dataset have been used. Several transformations such as rotations, scaling or brightness were applied in order to augment the datasets. Detectors were developed using the GoogleNet and SqueezNet architectures as CNN base on a Faster R-CNN. The best result for gun detection was obtained using a SqueezeNet architecture achieving a 85.45% AP50. For knife detection, GoogleNet approach accomplished 98% accuracy. Both detector results improve upon previous literature studies evidencing the effectiveness of our detectors.

10. FUTURE SCOPE

The future work includes reducing the false positives and negatives even more as there is still a need for improvement. We might also try to increase the number of classes or objects in the future but the priority is to further improve precision and recall.

11. REFERENCES

[1]. R. Xu, S. Y. Nikouei, Y. Chen et al., “Real-time human objects tracking for smart surveillance at the edge,” in Proceedings of the 2018 IEEE International Conference on Communications (ICC), pp. 1–6, Kansas City, MO, USA, May 2018.

[2]. G. K. Verma and A. Dhillon, “A handheld gun detection using faster R-CNN deep learning,” in Proceedings of the 7th International Conference on Computer and Communication Technology, pp. 84–88, Kurukshetra, Haryana, November 2017.

[3]. S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. Faughnan, “Smart surveillance as an edge network service: from harr-cascade, SVM to a lightweight CNN,” in Proceedings of the 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp. 256–265, Philadelphia, PA, USA, April 2018.

[4]. Akcay, S., Kundegorski, M.E., Willcocks, C.G., Breckon, T.P.: Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. IEEE Transactions on Information Forensics and Security 13(9),

2203–2215 (Sep 2018). https://doi.org/10.1109/TIFS.2018.2812196

[5]. Dastidar, J.G., Biswas, R.: Tracking human intrusion through a CCTV. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN). pp. 461–465 (Dec 2015). https://doi.org/10.1109/CICN.2015.95