

Real Time and Surveillance Weapon Detection using Deep learning

¹G. Hemalatha , ²K. Sudhakar

Department of CSE, Gokula Krishna College of Engineering, Sullurpet, India.

gollahemasai@gmail.com

Abstract: Security and safety is a big concern for today's modern world. For a country to be economically strong, it must ensure a safe and secure environment for investors and tourists. Having said that, Closed Circuit Television (CCTV) cameras are being used for surveillance and to monitor activities i.e. robberies but these cameras still require human supervision and intervention. We need a system that can automatically detect these illegal activities. Despite state-of-the-art deep learning algorithms, fast processing hardware, and advanced CCTV cameras, weapon detection in real-time is still a serious challenge. Observing angle differences, occlusions by the carrier of the firearm and persons around it further enhances the difficulty of the challenge. This work focuses on providing a secure place using CCTV footage as a source to detect harmful weapons by applying the state of the art open-source deep learning algorithms. We have implemented binary classification assuming pistol class as the reference class and relevant confusion objects inclusion concept is introduced to reduce false positives and false negatives. No standard dataset was available for real-time scenario so we made our own dataset by making weapon photos from our own camera, manually collected images from internet, extracted data from YouTube CCTV videos, through GitHub repositories, data by university of Granada and Internet Movies Firearms Database (IMFDB) imfdb.org. Two approaches are used i.e. sliding window/classification and region proposal/object detection. Some of the algorithms used are VGG16, Inception-V3, Inception-ResnetV2, SSDMobileNetV1, Faster-RCNN Inception-ResnetV2 (FRIRv2), YOLOv3, and YOLOv4. Precision and recall count the most rather than accuracy when object detection is performed so these entire algorithms were tested in terms of them. Yolov4 stands out best amongst all other algorithms and gave a F1-score of 91% along with a mean average precision of 91.73% higher than previously achieved.

Keywords: Real-Time Weapon Detection, CCTV Surveillance Analytics, Deep Learning-Based Object Detection, YOLOv4 and Region Proposal Networks, Confusion Object Inclusion Strategy.

1 INTRODUCTION

The crime rate across the globe has increased mainly because of the frequent use of handheld weapons during violent activity. For a country to progress, the law-and-order situation must be in control. Whether we want to attract investors for investment or to generate revenue with the tourism industry, all these needs is a peaceful and safe environment. The crime ratio because of guns is very critical in numerous parts of the world. It includes mainly those countries in which it is legal to keep a firearm. The world is a global village now and what we speak or write has an impact on the people. Even if the news they heard is crafted having no truth but as it gets viral in a few hours because of the media and especially social media, the damage will be done. People now have more depression and have less control over their anger, and hate speeches can get those people to lose their minds [1]. People can be brainwashed and psychological studies show that if a person has a weapon in this situation, he may lose his senses and commit a violent activity.

High incidents were recorded in past few years with the use of harmful weapons in public areas. Starting with the past year's attacks on a couple of Mosques in New Zealand, on March 15, 2019 at 1:40 pm, the attacker attacks the Christchurch AL-Noor Mosque during a Friday prayer killing almost 44 innocent and unarmed worshippers. On the same day just after 15 minutes at 1:55 PM, another attack happened killing seven more civilians. Active shooter incidents had also occurred in USA and then in Europe. The most significant cases were those at Columbine High School (USA, 37 victims), Andreas Broeivik's assault on Uotya Island (Norway, 179 victims) or the Charlie Hebdo newspaper attack killing 23. According to stats provided by the UNODC, among 0.1 Million people of a country, the crimes involving guns are very high i.e. 1.6 in Belgium, United States having 4.7 and Mexico with a number of 21.5.

CCTV cameras play an important role to overcome this problem and are considered to be one of the most important requirements for the security aspect. CCTVs are installed in every public place today and are mainly used for providing safety, crime investigation, and other security measures for detection [2]. CCTV footage is the most important evidence in courts. After a crime is committed, law enforcement agencies arrive at the scene and take the recording of footage with them. If we look at the surveillance system of different countries around the world, UK has about 4.5 million cameras, which are used for surveillance. Sweden has about 50000 cameras installed around 2010.

The government of Poland was able to reduce drug cases by 60% and street fights by 40% by installing just 450 cameras in the city of Poznan. China has the world's biggest surveillance system and 170 million cameras around the nation, and these are expected to expand three times, through an additional 400 million to be connected by 2020. It took only seven minutes for Chinese officials to find and apprehend BBC reporter John Sudworth using their strong CCTV cameras network and facial recognition technology and put the criminal behind the bar.

In previous years, though having surveillance cameras installed, to use them for security purposes was not an easy and dependable method. A human has to be there all the time to monitor screens. CCTV operator has to monitor 20–25 screens for 10 hours. He has to look, observe, identify, and control the situation that can be harmful to the individuals and the property. As the number of screens increases, the concentration of the person decreases considerably to monitor each screen with time. It is impossible for the person monitoring the screens to keep the same level of attention all the time [3].

The solution to aforementioned problem is to install surveillance cameras with the ability to automatically detect weapons and raise alarm to alert the operators or security personals. However, there is not much work done on algorithms for weapon detection in surveillance cameras, and related studies are often considering concealed weapon detection (CWD), mostly using X-rays or millimeter waves images employing traditional machine learning techniques. In the past few years, deep learning in particular convolutional neural network (CNN) has given groundbreaking results in object categorizing and detection. It has achieved finest results thus far in classical problems of image processing such as grouping, detection and localization. Instead of selecting features manually, CNN automatically learns features from given data.

This article presents an automatic detection and classification method of weapons for real-time scenario using state of the art deep learning models. For real-time implementation relating the problem question of this work “detecting weapons in real-time for potential robbers/terrorist using deep learning”, detection and classification was done for pistol, revolver and other shot handheld weapons as in single class called pistol and related confusion objects such as cell phone, metal detector, wallet, selfie stick in not pistol class [3][4]. A major reason behind this was our research done on weapons used in robbery cases and it further motivated us to choose pistol and revolver as our target object. We go through several CCTV captured robbery videos on YouTube and found that almost 95% of cases have pistol or revolver as the weapon used. With the implementation of this system, many robbery crimes, and other incidents like what happened last year in New Zealand's Christchurch mosque could be controlled using early alarm system by alerting the operator and concerned authorities so action can be taken immediately.

Gun 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.
- For real-time scenarios, making a novel dataset manually was a very long and time-consuming process.
- Labeling the desired database is not an easy task, as all data needs to be labeled manually.
- Different detection algorithms were used, so a labeled dataset for one algorithm cannot be utilized for the other one.
- Every algorithm requires different labeling and pre-processing operations for the same-labeled 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, inter-object, or background blocking.

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 and 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 real time training data as well for better results [5].

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 preprocessed 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 among them Yolov4 performed best in terms of both speed and accuracy and our selected trained model predict images at almost every orientation, angle, and view, achieving the highest mean average precision of 91.73% along with a F1-score of 91% on Yolov3.

2 RELATED WORK

The problem of detection and classification of objects in real-time started after major developments in the CCTV field, processing hardware, and deep learning models. Very little work has been done in this field before and most of the previous effort was related to concealed weapon detection (CWD). Starting with concealed weapon detection (CWD), before its use in weapon detection, it was used for luggage control and other security purposes at airports and was based on imaging techniques like millimeter-wave and infrared imaging. Sheen *et al.* suggested CWD method based on a three-dimensional millimeter (mm) wave imaging method, for detecting hidden weapons at airports and other safe locations in the body. Z. Xue *et al.* suggested a CWD technique based on a fusion-based technique of multi-scale decomposition, which combines color visual picture with infrared (IR) picture integration [6]. R. Blum *et al.* suggested a CWD method based on the inclusion of visual picture and IR or mm wave picture using a multi-resolution mosaic technique to highlight the hidden weapon of the target picture.

R. Pravesh *et al.* suggested a CWD technique using image fusion [7]. They used IR image and visual fusion to detect hidden weapons in a situation where the image of the scene was present over and under exposed area. Their methodology was to apply a homomorphic filter captured at distinct exposure conditions to visual and IR pictures. Current techniques attain high precision by using various combinations of extractors and detectors, either by using easy intensity descriptors, boundary detection, and pattern matching or by using more complicated techniques such as cascade classifiers with boosting. CWD though had worked for some sort of cases but it had many limitations. These systems were based on metal detection; non-metallic guns cannot be detected. They were costly to use in many locations because they need to be coupled with X-ray scanners and conveyor belts and responds to all metallic objects, so were not accurate. Economic cost and health risks limited the practical implementation of such methods. Furthermore, video-based firearm detection was a preventive measure for acoustic detection of gunshot and can be combined with it for implementation.

The idea of automated image processing for public security purposes in many fields has been well recognized and studied. CCTV was the ultimate need for this kind of work to progress. CCTV was first used back in 1946 in Germany and at that time, these cameras were installed to observe the launch of a rocket named V2. Although it had been used earlier, major improvements happened in the last two decades. With the advancement in CCTV technology, visual object recognition and detection for surveillance, control, and security were performed. In 1973, Charge-Coupled Device (CCD) was developed, which made the deployment of surveillance cameras possible by 1980. If we go a bit forward in time, a company named Axis Communication developed the first-ever network camera, which enabled the transformation of surveillance cameras from analog to digital. This transformation of analog to digital video made it possible for everyone to apply image processing, machine learning, and computer vision techniques on videos recorded from surveillance cameras. In 2003, Royal Palm Middle School in Phoenix used facial recognition for the first time for tracking missing children.

Several object detection algorithms were proposed in the field of computer vision to make surveillance system better. Object detection algorithms were used in several sectors like anomaly detection, deterrence, human detection, and traffic monitoring. R. Pravesh *et al.* discussed briefly object tracking and detection in surveillance cameras [6][7]. The authors had explained the tracking of an object using multiple surveillance cameras. Another author addressed techniques for detecting objects that come into contact with another object and are occluded. They also wrote regarding the segmentation of mean fluctuations. They outlined how mean segmentation of shifts can help detect objects. They used a Bayesian Kalman filter with a simplified Gaussian blend (BKF-SGM) algorithm to track the detected object. J.S Marques proposed distinct techniques for evaluating the efficiency of distinct algorithms for object recognition. B. Triggs *et al.* described histogram oriented gradient (HOG). HOG became a novel architecture for feature extraction. It was used mostly in applications involved in human detection. In 2005, the sliding window technique was proposed for the recognition of number plates. They had used a sliding window for the purpose of segmentation and a neural network for character recognition on the number plate.

As described above, objection detection for the computer vision tasks was used for some applications with big objects to identify like a person, transport or traffic monitoring, etc. Literature review on weapon detection left me with the opinion that regardless of many object detection algorithms, the algorithms proposed for weapon detection are very few. At last, the idea of firearm detection using the images and videos was proposed and false alarms were reduced by classifying neural networks with region-based descriptors and determining region of interest (ROI) using the sliding window technique and then trained the neural network classifier with image pixels.

With the development in CCTV's, object detection for different computer vision problems for real-time were performed and the idea to detect firearms were introduced first by A. S. V. Rao *et al.* in 2007 and a surveillance system was also implemented by them a year later in 2008 [8]. In the aforementioned work, writers created an accurate pistol detection model for RGB pictures. However, in the same scene, their method did not detect various pistols. The approach used comprises of first removing non-related items from the segmented picture using the K-mean clustering algorithm and then applying the SURF (Speed up Robust Features) method to detect points of interest. Darker gave the concept of SIFT based weapon detection algorithm and for ROI estimation, used the motion segmentation method. SIFT algorithm is prone to false alarms, so for estimating ROI, authors used motion segmentation rather than using SIFT on complete image. When ROI was determined, then SIFT was applied to detect firearms in their case.

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. This work focuses on the state of the art deep learning network rather SIFT and HOG features which use handcrafted rules for feature extraction, selection, and detection in real-time visual scenario using CCTV cameras. X. Zhang *et al.* concluded an important finding that helped my work [6][9]. They concluded that the automatic feature representation gave improved results rather than manual features. Not only the learned features were better in performance, they also had learned the deep representation of the data and reduced a lot of manual work, and saved time and energy.

H. Martinez *et al.* proposed the technique of knife and gun detection in surveillance systems [10]. They had used HOG as a feature extractor along with backpropagation of artificial neural networks for classification purposes. The detection was performed using different scenarios, first weapon only and then using HOG and background subtraction methods for human before the desired object and claimed to have an accuracy of 83%. The aforementioned work uses the CNN along with non-linearity of ReLu, convolutional neural layer, fully connected layer, and dropout layer of CNN to reach a result for detection with multiple classes and implemented their work using the Tensor flow open-source platform. Their system achieved a test accuracy of 90.2 % for their dataset. MichałGrega *et al.* proposed knives and firearm detection in CCTV images. They had applied MPEG-7 and principle component analysis along with the sliding window approach, which made their work slower for real-time scenarios, although they claimed to achieve good accuracy on their test dataset.

S. Solak *et al.* had also used the deep learning technique to detect weapons and used the Faster RCNN model [11][12]. The work was performed on imfdb, which in my opinion is not suitable to train a model for real-time case. They claimed to have an accuracy of 93.1% on that dataset but in the case of weapon detection, only achieving higher accuracy is not enough, and precision and recall must be considered. Siham Tabik *et al.* work was very much related to the real-time scenario. They used Faster RCNN to detect weapons in real-time using sliding window and region proposal methods. Best results were obtained by using the region proposal technique. The sliding window was also very time-consuming and took 14 s/image, on the other hand, the region proposal method processed the image in 140ms with 7 fps. They trained the network on Faster RCNN using only one class focusing on reducing the false positive.

Recent past objection detection work with the application to firearms was proposed in 2019, where a group of researchers, Javed Iqbal *et al.* proposed orientation aware detection of the object. This system is more suitable for long and thin objects like rifles etc. The predicted bounding box in their case was aligned with the object and had the less unnecessary area to deal with. Images of very high quality were used for training and testing purposes, which may make it less suitable for real-time scenarios. Jose Luis Salazar Gonz'alez *et al.* work was very much related to achieve real-time results. They did immense experimentation using different datasets and trained Faster-RCNN using Feature Pyramid Network with Resnet50 and improves the previous state of the art by 3.91 %.

3 PROPOSED METHODOLOGY

Deep learning is a branch of machine learning inspired by the functionality and structure of the human brain also called an artificial neural network. The methodology adopted in this work features the state of art deep learning, especially the convolutional neural networks due to their exceptional performance in this field. The aforementioned techniques are used for both the classification as well as localizing the specific object in a frame so both the object classification and detection algorithms were used and because our object is small with other object in background so after experimentation it found the best algorithm for our case. Sliding window/classification and region proposal/object detection algorithms were used, and these techniques will be discussed later in this section.

It had started by doing the classification using different deep learning models and achieved good precision but for the real-time scenarios, the low frame per seconds of classification models were the real issue in implementation. Oxford VGG, Google Inceptionv3 and InceptionResnetv2 were trained using the aforementioned approach. To achieve high precision, increase number of frame per seconds and improve localization, it moved to the object detection and region proposal methods. The different state of the art deep learning models for object detection were used and the results were compared in terms of precision, speed, and standard metric of F1 score. State of the art deep learning based SSDMobileNetv1, YOLOv3, FasterRCNN-InceptionResnetv2, and YOLOv4 were trained and tested.

Different datasets were made keeping in mind the classification and detection problem as both have a separate requirement for performing the tasks to achieve high accuracy, mean average precision as well as frame per second for the real-time implementation. To understand object classification and detection let us first briefly understand object recognition as both the aforementioned types come under the umbrella of this and combined classification and localization make detection possible for any kind of detection problem giving class name as well as the region where our desired object is in the frame.

3.1. 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. Fig. 1 depicts that classification and localization come under the category of recognition and combined classification and localization is performed to do object detection. Let us have a brief overview of the object classification, localization, and detection.

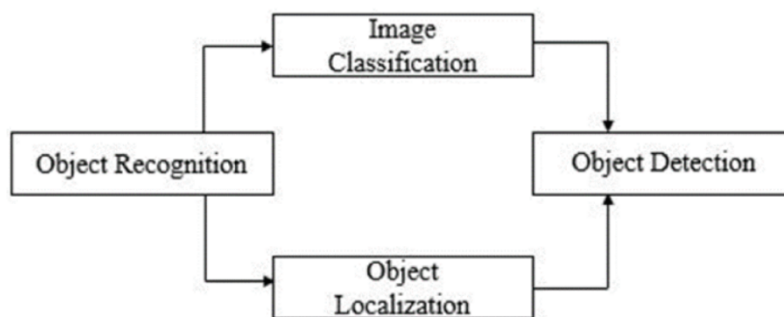


Fig. 1. Object Recognition to detection Hierarchy.

3.1.1. 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.

3.1.2. 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.

3.1.3 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. In the beginning, the models were not very deep, means there were not many layers as they are now in an algorithm. Because of the aforementioned developments, in 2012 A.Krizhevsky presented a model called Alex Net trained on ImageNet and got the first position in that competition. This was the beginning of object detection in deep learning. It gave a way to researchers and then every year the algorithms and models keep on coming. All these algorithms contain layers that work on the principle of the convolutional neural network (CNN).

3.2. Classification and Detection 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 they have the following two approaches used in our work for both classification and detection models are:

- Sliding window/Classification Models
- Region proposal/Object Detection Model

3.2.1. 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 need to search in the picture for all feasible places, 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 (10^4) 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.

3.2.2. 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 (R-CNN) 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. Yolo series is among the state of the art object detection models. 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.

3.3. Training Mechanism

Fig. 2 describes 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 save those weights as a classifier but if it's incorrect then comes the process of backpropagation algorithm along with the gradient descent algorithm. In backpropagation, weights are optimized by subtracting the partial derivative of cost function $J(O)$ with a multiplier of the learning rate α 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 it get the best and desired weights values.

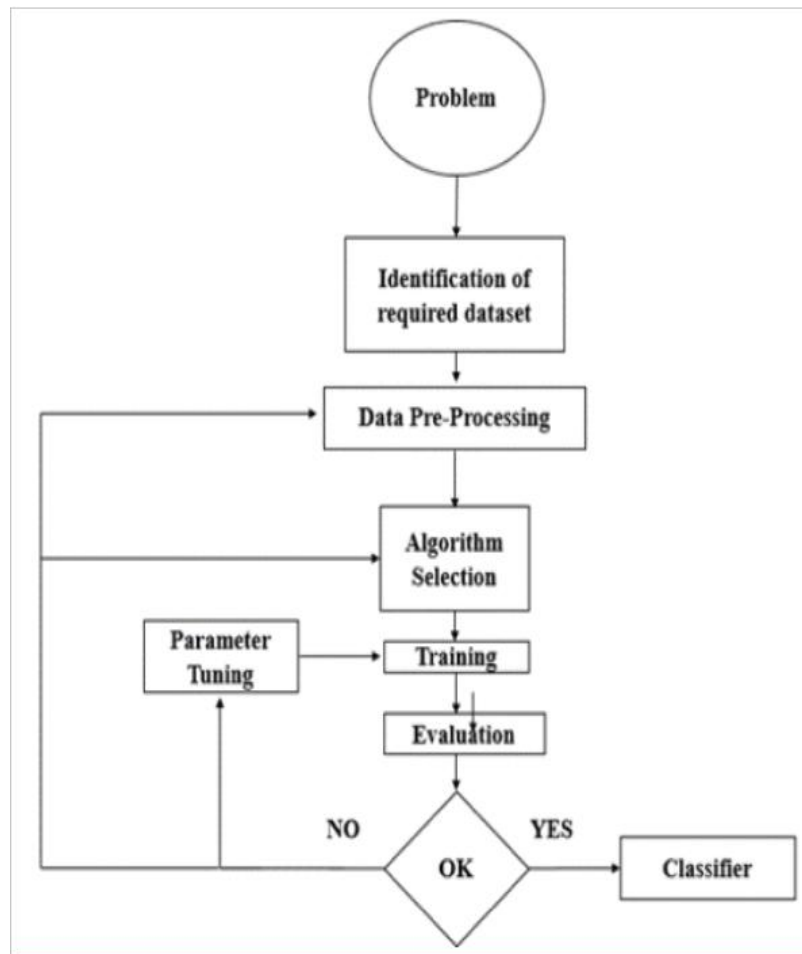


Fig. 2. Training and Optimization Flow Diagram.

3.4. Confusion Object Inclusion

They have formulated the problem to reduce the number of false positives and negatives by adding relevant confusion object. The weapon category includes all the handheld weapons such as, pistol, revolver, shotgun and other than weapon includes the objects that can most be confused with pistol classes e.g. mobile, metal detector, selfie stick, purse, etc. By understanding the differences between classification and detection algorithms, sliding window, and region proposal methods, let's now look at the algorithms used for both approaches.

3.5. Classifiers and Object Detectors

The classifiers used under the sliding window approach:

- VGG16
- InceptionV3
- Inception ResnetV2

The object detectors used for real-time detection are:

- SSD MobilNetV1
- YoloV3
- Faster RCNN-Inception ResNetV2

Three databases named database1, database2, and database3 were created one by one after experimentation on different algorithms with diverse images, first for classification and then for object detection. Although the results obtained from the classification algorithms were not bad but the frames per second were very slow for real-time implementation. Detail for each database will be discussed in the next section.

4 EXPERIMENTAL RESULTS AND ANALYSIS

We have detected weapons in real-time CCTV streams in low resolution, dark light with real-time frame per second. Most of the work done before was on detecting images and videos of high quality and because those models were trained on high-quality datasets, it is not possible to then detect an object of low resolution in real-time. The results are analyzed after training and testing models on datasets mentioned in Table 1. As described in the methodology section the results for different approaches are evaluated. The main problem statement is of real-time detection because 97% of weapon used in robbery cases were pistol or revolver, so different dataset results have been evaluated here for sliding window and region proposal approach. The performance of these models was analyzed by comparing them in terms of the standard metrics of F1-score and frame per seconds along with mean average precision (MAP) for the best performed model and these terms are calculated by using the below equations. F1 score is ratio of the precision and recall functions.

4.1. Anchor Boxes

YOLO uses anchor boxes to help predict bounding boxes for detecting objects of various sizes and aspects ratios, improving detection accuracy for things like weapons. Masks (which anchors go to which layer). Each [yolo] layer uses 3 anchors via the mask line:

- First YOLO layer (13×13, small objects)
mask = 6,7,8 → anchors: (116,90), (156,198), (373,326)
- Second YOLO layer (26×26, medium objects)
mask = 3,4,5 → anchors: (30,61), (62,45), (59,119)
- Third YOLO layer (52×52, large objects)
mask = 0,1,2 → anchors: (10,13), (16,30), (33,23)

Table 1. Shows the configuration is using these anchor boxes:

(10,13),	(16,30)	(33,23)
(30,61),	(62,45)	(59,119)
(116,90),	(156,198)	(373,326)

And they're divided among the 3 YOLO layers like this:

- Small objects (52×52 grid) : (10,13), (16,30), (33,23)
- Medium objects (26×26 grid) : (30,61), (62,45), (59,119)
- Large objects (13×13 grid) : (116,90), (156,198), (373,326)

4.2. Analysis and Discussion

- Some classification models showed good results but they were not suitable for a real-time scenario, were slow, not much accurate, and fast as compared to the object detection models as they performs very well and achieved high precision and recall.
- The reason why some classification models have a good F1-score is the training and evaluation on initial datasets we made when starting this work, but after experimentation, we come to know that these models are not suitable for real-time scenarios having the background objects.
- Object detection models performed well for the real-time scenario and performance comparison in terms of speed and F1-score between the detection models can be seen. Inference results are obtained using the NVIDIA RTX 2080ti for each model.
- The standard metrics of mean average precision (MAP), recall and F1-score are calculate and all the models have been compared at a benchmark IoU threshold of 0.50 or 50%.
- Comparison in terms of test accuracy vs F1-score for the best-performed models of both classification and detection approaches is shown in the Fig. Accuracy and F1-score for VGG, Inceptionv3, InceptionResNetv2, SSD Mobile Net, FasterRCNN-InceptionResNetv2, Yolov3 and yolov4 are 78.20%, 85.20%, 92.20%, 79%, 96%, 94%, 99% and 81.69%, 84.36%, 85.74%, 59%, 87%, 86% and 91% respectively.
- It shows the inference or detection results of our model for pistol and not pistol class on images, videos, and real-time CCTV streams.



Fig. 3. Detection Results

It shows detection results Only weapon in the whole frame without any background at different angles, brightness, sharpness, and quality.



Fig. 4. Detection Results-Top left to bottom right

It shows that detection Results-Top left to bottom right: (a) Image with front and side view, (b) Image vertical view (c) Image with Dark background and Low Resolution fully tilted side view, (d) Low brightness image side view slightly tilted (e) Image with the back view (f) Full front view (g) Small CCTV object (h) Very small object with side view (i) Image with full side view.

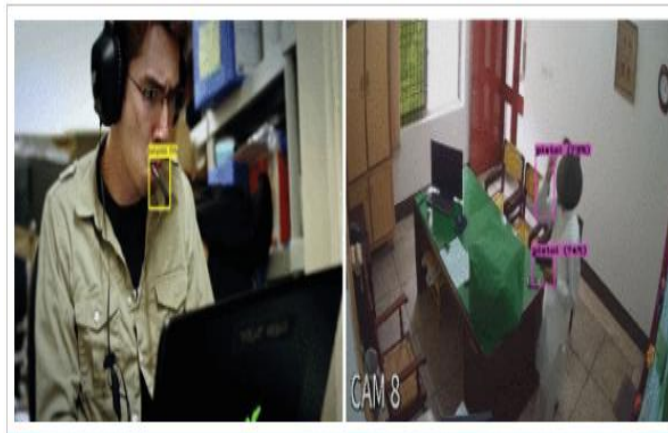


Fig. 5. Misdetections: False positives and Negatives.

It is very hard to do a comparison with studies conducted previously on this subject because each study has its own dataset, models and metrics used to evaluate performance. It should also be noticed achieve real time detection, we also need to have a real time dataset for training because with high quality training images we cannot achieve results in real time.

Each study also has different testing conditions, either just on images, videos or on images with high quality but our approach from start was to achieve real time results. In some studies, the performance metric used is accuracy, others have precision or mean average precision (MAP) but mostly MAP is used as standard so we have given comparison results in terms of MAP and precision at a standard iou threshold of 50%, which ever was available.

5 CONCLUSIONS

For both monitoring and control purposes, this work has presented a novel automatic weapon detection system in real-time. This work will indeed help in improving the security, law and order situation for the betterment and safety of humanity, especially for the countries who had suffered a lot with these kind of violent activities. This will bring a positive impact on the economy by attracting investors and tourists, as security and safety are their primary needs. We have focused on detecting the weapon in live CCTV streams and at the same time reduced the false negatives and positives. To achieve high precision and recall we constructed a new training database for the real-time scenario, then trained, and evaluated it on the latest state-of-the-art deep learning models using two approaches, i.e. sliding window/classification and region proposal/object detection. Different algorithms were investigated to get good precision and recall.

Through a series of experiments, we concluded that object detection algorithms with ROI (Region of Interest) perform better than algorithms without ROI. We have tested many models but among all of them, the state-of-the-art Yolov4, trained on our new database, gave very few false positive and negative values, hence achieved the most successful results. It gave 91.73% mean average precision (MAP) and a F1-score of 91% with almost 99% confidence score on all types of images and videos. We can say that it satisfactorily qualifies as an automatic real-time weapon detector. Looking at the results, we got the highest mean average precision (MAP) F1-score as compared to the research done before for real-time scenarios. 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.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare that they have no conflicts of interest related to this study.

LICENSING

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

REFERENCES

- [1] T. Santos, H. Oliveira, and A. Cunha, "Systematic review on weapon detection in surveillance footage through deep learning," *Computer Science Review*, vol. 51, p. 100612, Dec. 2023, doi: 10.1016/j.cosrev.2023.100612.
- [2] M. H. Faheem, Q. E. U. Haq, T. M. Khan, and J. H. Kim, "Deep-Learning-Based weapon detection using drones equipped with infrared and color cameras," *International Journal of Digital Crime and Forensics*, vol. 17, no. 1, pp. 1–16, Nov. 2025, doi: 10.4018/ijdcf.393088.
- [3] D. Berardini, L. Migliorelli, A. Galdelli, and M. J. Marín-Jiménez, "Edge artificial intelligence and super-resolution for enhanced weapon detection in video surveillance," *Engineering Applications of Artificial Intelligence*, vol. 140, p. 109684, Nov. 2024, doi: 10.1016/j.engappai.2024.109684.
- [4] P. Yadav, N. Gupta, and P. K. Sharma, "A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods," *Expert Systems With Applications*, vol. 212, p. 118698, Aug. 2022, doi: 10.1016/j.eswa.2022.118698.
- [5] A. Castillo, S. Tabik, F. Pérez, R. Olmos, and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," *Neurocomputing*, vol. 330, pp. 151–161, Nov. 2018, doi: 10.1016/j.neucom.2018.10.076.
- [6] G. Wang, H. Ding, M. Duan, Y. Pu, Z. Yang, and H. Li, "Fighting against terrorism: A real-time CCTV autonomous weapons detection based on improved YOLO v4," *Digital Signal Processing*, vol. 132, p. 103790, Oct. 2022, doi: 10.1016/j.dsp.2022.103790.
- [7] R. Pravesh and B. C. Sahana, "A dual-stage deep learning framework for simultaneous fire and firearm detection in smart surveillance systems," *Results in Engineering*, vol. 27, p. 106330, Jul. 2025, doi: 10.1016/j.rineng.2025.106330.

- [8] A. S. V. Rao, S. Kainth, A. Bhattacharya, and T. Amgoth, "An efficient weapon detection system using NSGCU-DCNN classifier in surveillance," *Expert Systems With Applications*, vol. 255, p. 124800, Jul. 2024, doi: 10.1016/j.eswa.2024.124800.
- [9] P. Yadav, N. Gupta, and P. K. Sharma, "Robust weapon detection in dark environments using Yolov7-DarkVision," *Digital Signal Processing*, vol. 145, p. 104342, Dec. 2023, doi: 10.1016/j.dsp.2023.104342.
- [10] H. Martinez, F. J. Rodriguez-Lozano, F. León-García, J. M. Palomares, and J. Olivares, "Distributed Fog computing system for weapon detection and face recognition," *Journal of Network and Computer Applications*, vol. 232, p. 104026, Sep. 2024, doi: 10.1016/j.jnca.2024.104026.
- [11] K. G. Dhuri and S. Patil, "Crime prediction against women in surveillance videos using deep learning models: A review," *Computers & Electrical Engineering*, vol. 128, p. 110651, Sep. 2025, doi: 10.1016/j.compeleceng.2025.110651.
- [12] S. Solak, B. Cetin, and M. H. B. Ucar, "Real-time theft detection in urban surveillance: A comparative analysis of YOLO-based approach," *Journal of Engineering Research*, Nov. 2025, doi: 10.1016/j.jer.2025.11.009.