Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9975700, 9 pages
https://doi.org/10.1155/2021/9975700

Research Article

Weapon Detection Using YOLO V3 for Smart Surveillance System

Sanam Narejo,[1] **Bishwajeet Pandey**,[2] **Doris Esenarro vargas**,[3] **Ciro Rodriguez**,[4] **and M. Rizwan Anjum**[5]

[1]*Department of Computer Systems Engineering, Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan*
[2]*Gran Sasso Science Institute, L'Aquila, Italy*
[3]*Universidad Nacional Federico Villarreal, Lima, Peru*
[4]*Universidad Nacional Mayor de San Marcos, Lima, Peru*
[5]*Department of Electronic Engineering, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*

Correspondence should be addressed to Bishwajeet Pandey; dr.pandey@ieee.org

Received 4 March 2021; Revised 15 April 2021; Accepted 3 May 2021; Published 12 May 2021

Academic Editor: Zain Anwar Ali

Every year, a large amount of population reconciles gun-related violence all over the world. In this work, we develop a computer-based fully automated system to identify basic armaments, particularly handguns and rifles. Recent work in the field of deep learning and transfer learning has demonstrated significant progress in the areas of object detection and recognition. We have implemented YOLO V3 "You Only Look Once" object detection model by training it on our customized dataset. The training results confirm that YOLO V3 outperforms YOLO V2 and traditional convolutional neural network (CNN). Additionally, intensive GPUs or high computation resources were not required in our approach as we used transfer learning for training our model. Applying this model in our surveillance system, we can attempt to save human life and accomplish reduction in the rate of manslaughter or mass killing. Additionally, our proposed system can also be implemented in high-end surveillance and security robots to detect a weapon or unsafe assets to avoid any kind of assault or risk to human life.

1. Introduction

Violence committed with guns puts significant impact on public, health, psychological, and economic cost. Many people die each year from gun-related violence. Psychological trauma is frequent among children who are exposed to high levels of violence in their communities or through the media. Children exposed to gun-related violence, whether they are victims, perpetrators, or witnesses, can experience negative psychological effects over the short and long terms. Number of studies show that handheld gun is the primary weapon used for various crimes like break-in, robbery, shoplifting, and rape. These crimes can be reduced by identifying the disruptive behavior at early stage and monitoring the suspicious activities carefully so that law enforcement agencies can further take immediate action [1].

Levels of gun-related violence vary greatly among geographical locations and countries. The global death toll from use of guns may be as high as 1,000 dead each day [2]. According to statistics, 4.2 in 100000 people are killed in Pakistan every year in mass shootings. From street crimes to an individual institution attack, many precious lives suffered. This further indicates that manual surveillance system still needs human eye to detect the abnormal activities and it takes a sufficient amount of time reporting to security officials to tackle the situation.

Although the human visual framework is quick and precise and can likewise perform complex undertakings like distinguishing different items and recognizing snags with minimal cognizant idea, however, it is common truth that if an individual watches something very similar for quite a long time, there is an opportunity of sluggishness and lack of regard.

Nowadays, with the accessibility of huge datasets, quicker GPUs, advanced machine learning algorithms, and better calculations, we can now effectively prepare PCs and develop automated computer-based system to distinguish and identify numerous items on a site with high accuracy.

Recent developments indicate that machine learning [3–6] and advance image processing algorithms have played dominant role in smart surveillances and security systems [7, 8]. Apart from this, popularity of smart devices and networked cameras has also empowered this domain. However, human objects or weapon detection and tracking are still conducted at cloud centers, as real-time, online tracking is computationally costly. Significant efforts have been made in recent years to monitor robot manipulators that need high control performance in reliability and speed [9, 10]. The researchers have attempted to improve the response characteristics of the robotic system and to attenuate the uncertainties in [11]. The proposed developed robust model-free controller incorporates time delay control (TDC) and adaptive terminal sliding mode control (ATSMC) methods.

In this research work, we aim to develop a smart surveillance security system detecting weapons specifically guns. For this purpose, we have applied few compute vision methods and deep learning for identification of a weapon from captured image. Recent work in the field of machine learning and deep learning particularly convolutional neural networks has shown considerable progress in the areas of object detection and recognition, exclusively in images. As the first step for any video surveillance application, object detection and classification are essential for further object tracking tasks. For this purpose, we trained the classifier model of YOLO v3, i.e., "You Only Look Once" [12, 13]. This model is a state-of-the-art real-time object detection classifier. Furthermore, we are not just detecting the guns, rifles, and fire but also getting the location of the incident and storing the data for future use. We have connected three systems using socket programing as a demonstration for the real-life scenario as camera, CCTV operator, and security panels.

This work is an attempt to design and develop a system which can detect the guns, rifles, and fire in no time with less computational resources. It is evident from technological advancements that most of the human assisted applications are now automated and computer-based. Eventually, in future these computer-based systems will be replaced by more smart machines, robots, or humanoid robots. In order to provide visionary sense to robots, object detection plays fundamental part for understanding the objects and its interpretation. Thus, our proposed system can also be implemented in surveillance and security robots to detect any weapon or unsafe assets.

2. Literature Review

Reducing the life-threatening acts and providing high security are challenging at every place. Therefore, a number of researchers have contributed to monitoring various activities and behaviors using object detection. In general, a framework of smart surveillance system is developed on three levels: firstly, to extract low-level information like features engineering and object tracking; secondly, to identify unusual human activities, behavior, or detection of any weapon; and finally, the high level is about decision making like abnormal event detection or any anomaly. The latest anomaly detection techniques can be divided into two groups, which are object-centered techniques and integrated methods. The convolutional neural network (CNN) spatial-temporal system is only applied to spatial-temporal volumes of interest (SVOI), reducing the cost of processing. In surveillance videos of complex scenes, researchers in [14] proposed a tool for detecting and finding anomalous activities. By conducting spatial-temporal convolution layer, this architecture helps one to capture objects from both time domain and frequency domain, thereby extracting both the presence and motion data encoded in continuous frames. To do traditional functions to local noise and improve detection precision, spatial-temporal convolution layers are only implemented within spatial-temporal quantities of changing pixels. Researchers proposed anomaly-introduced learning method for detecting anomalous activities by developing multi-instance learning graph-based model with abnormal and normal bimodal data, highlighting the positive instances by training coarse filter using kernel-SVM classifier and generating improved dictionary learning known as anchor dictionary learning. Thus, abnormality is measure by selecting the sparse reconstruction cost which yields the comparison with other techniques including utilizing abnormal information and reducing time and cost for SRC.

Hu et al. [15] have contributed in detecting various objects in traffic scenes by presenting a method which detects the objects in three steps. Initially, it detects the objects, recognizes the objects, and finally tracks the objects in motion by mainly targeting three classes of different objects including cars, cyclists, and traffic signs. Therefore, all the objects are detected using single learning-based detection framework consisting of dense feature extractor and trimodal class detection. Additionally, dense features are extracted and shared with the rest of detectors which heads to be faster in speed that further needs to be evaluated in testing phase. Therefore, intraclass variation of objects is proposed for object subcategorization with competitive performance on several datasets.

Grega et al. presented an algorithm which automatically detects knives and firearms in CCTV image and alerts the security guard or operator [16]. Majorly, focusing on limiting false alarms and providing a real-time application where specificity of the algorithm is 94.93% and sensitivity is 81.18% for knife detection. Moreover, specificity for fire alarm system is 96.69% and sensitivity is 35.98% for different objects in the video. Mousavi et al. in [17] carried out video classifier also referred to as the Histogram of Directed Tracklets which identifies irregular conditions in complex scenes. In comparison to traditional approaches using optical flow which only measure edge features from two subsequent frames, descriptors have been developing over long-range motion projections called tracklets. Spatiotemporal cuboid footage sequences are statistically gathered on the tracklets that move through them.

Ji et al. developed a system for security footage which automatically identifies the human behavior using convolutional neural nets (CNNs) by forming deep learning model which operates directly on the raw inputs [18].

Therefore, 3D CNN model for classification requires the regularization of outputs with high-level characteristics to increase efficiency and integrating the observations of a variety of various models.

Pang et al. presented real-time concealed various object detection under human dress in [19]. Metallic guns on human skeleton were used for passive millimeter wave imagery which relies on YOLO algorithm on dataset of small scale. Subsequently, comparison is undertaken between Single MultiBox Detector algorithm, YOLOv3-13, SSD-VGG16, and YOLOv3-53 on PMMW dataset. Moreover, the weapon detection accuracy computed 36 frames per second of detection speed and 95% mean average precision. Warsi A et al. have contributed to automatically detecting the handgun in visual surveillance by implementing YOLO V3 algorithm with Faster Region-Based CNN (RCNN) by differentiating the number of false negatives and false positives [20], thus, taking real-time images and incorporating with ImageNet dataset then training it using YOLO V3 algorithm. They have compared Faster RCNN to YOLO V3 using four different videos and as a result YOLO V3 imparted faster speed in real-time environment.

3. Methodology

In this work, we have attempted to develop an integrated framework for reconnaissance security that distinguishes the weapons progressively, if identification is positively true it will caution/brief the security personals to handle the circumstance by arriving at the place of the incident through IP cameras. We propose a model that provides a visionary sense to a machine to identify the unsafe weapon and can also alert the human administrator when a gun or firearm is obvious in the edge. Moreover, we have programmed entryways locking framework when the shooter seems to carry appalling weapon. On the off chance conceivable, through IP webcams we can likewise share the live photo to approach security personals to make the move in meantime. Also, we have constructed the information system for recording all the exercises to convey impact activities in the metropolitan territories for a future crisis. This further ends up in designing the database for recording all the activities in order to take prompt actions for future emergency. Figure 1 presents the overall generalized approach of our research work divided into three parts.

The most important and crucial part of any application is to have a desired and suitable dataset in order to train the machine learning models. Therefore, we manually collected huge amount of images from Google. A few of the image samples are shown in Figure 2. For each weapon class, we collected at least 50 images. Using google-images-download is one of the best ways to collect images for constructing one's own dataset. We further saved those images to a folder called "images." One must save images in ".jpg" form; if the images are in different extensions, it will be a little troublesome and will generate errors when provided for training. Alternatively, since the images are processed in terms of batches, therefore prior to training, the sizes of all the images are transformed into the same width and height 416×416 pixels.

Object detection is primarily related to computer vision that includes distinguishing objects in computerized images. Object detection is a domain that has benefited immensely from the recent advancements in the realm of deep learning. YOLO is basically a pretrained object detector. It is a CNN model. A CNN is a deep learning algorithm which can take in a raw input image and assign learnable weights and biases to various aspects/objects in the image. A convolutional layer in CNN model is responsible of extracting the high-level features such as edges, from the input image. This works by applying kxk filter known as kernel repeatedly over raw image. This further results in activation maps or feature maps. These feature maps are the presence of detected features from the given input. Thus, the preprocessing required is much lower as compared to other classification algorithms, whereas in standard approach, filters are hand-engineered and in CNN these are learned through a number of iterations and training. Figure 3 indicates a basic CNN architecture as classification model for 10 different weapons. Subsequently, the next layer is Max-Pooling or Subsampling layer, which is responsible for reducing the spatial size of the convolved features. This is to decrease the computational power required to process the data through dimensionality reduction. ReLU is a rectified linear unit activation expressed in (1), which is related to the feature of non-saturating activation. It eliminates undesirable values from an activation map effectively by setting them to nil. Finally, the last layers are fully connected layers transforming the data into a 1-dimensional array. To create a particular long feature vector, the flattened output is fed to a feedforward neural network and backpropagation is applied to every iteration of training. These layers are liable to learn nonlinear combinations of the high-level features as represented by the output of the convolutional layer.

$$\text{ReLU:}\quad f(x) = \max(0, x). \tag{1}$$

As mentioned earlier that YOLO is a pretrained object detector, a pretrained model simply means that another dataset has been trained on it. It is extremely time consuming to train a model from scratch; it can take weeks or a month to complete the training step. A pretrained model has already seen tons of objects and knows how each of them must be classified. The weights in the abovementioned pretrained model have been obtained by training the network on COCO and Imagenet dataset. Thus, it can only detect objects belonging to the classes present in the dataset used to train the network. It uses Darknet-53 as the backbone network for feature extraction and uses three scale predictions. The DarkNet-53 is again convolutional neural network that has 53 layers as elucidated in Figure 4. DarkNet-53 is a fully convolutional neural network. Pooling layer is replaced with a convolution operation with stride 2. Furthermore, residual units are applied to avoid the gradient dispersion.

Initially, CNN architectures were quite linear. Recently, numerous variations are introduced, for example, middle

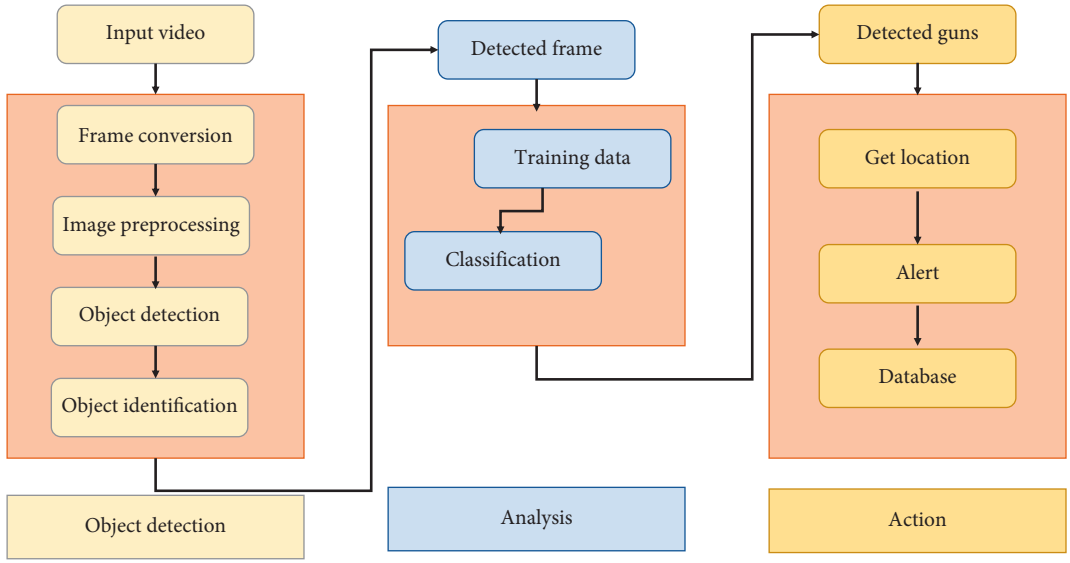

Figure 1: The flow of research methodology.

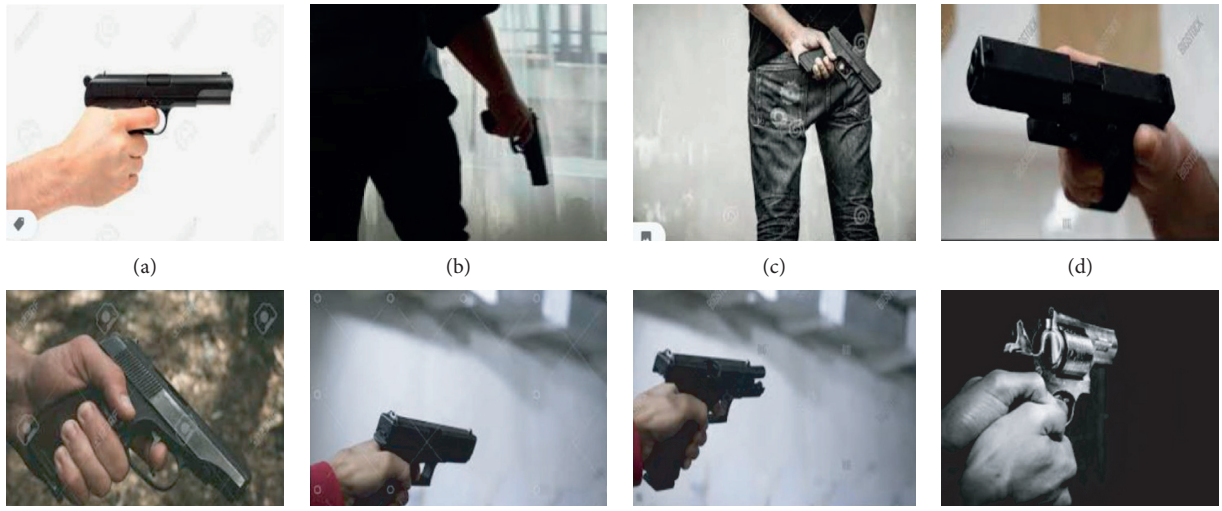

Figure 2: Sample images from collected dataset.

blocks, skip connections, and aggregations of data between layers. These network models have already acquired rich feature representations by getting trained over a wide range of images. Thus, selecting a pretrained network and using it as a starting point to learn a new task is a concept behind transfer learning. In order to recognize the weapons, we took the weights of a pretrained model and trained another YOLO V3 model.

YOLO V3 is designed to be a multiscaled detector rather than image classifier. Therefore, for object detection, classification head is replaced by appending a detection head to this architecture. Henceforth, the output is vector with the bounding box coordinates and probability classes. YOLO V3 inherits Darknet-53 as its backbone, a framework to train neural networks with 53 layers as indicated in Figure 4. Moreover, for object detection task additional 53 layers are stacked over it, accumulating to a total of a 106-layer fully convolutional architecture. Due to its multiscale feature fusion layers, YOLO V3 uses 3 feature maps of different scales for target detection as shown in Figure 5.

4. Experimental Results

Image classification includes, for example, the class of one object in a picture. However, object localization is to recognize the area of at least one article in a picture and drawing a proliferating box around their degree as shown in Figure 6. Moreover, Figure 7 illustrates the detection of rifle from an animated video. The shape of the detection kernel is computed by $1 \times 1 \times (\text{bb x } (4 + 1 + \text{nc}))$. Hence, bb is the number of bounding boxes, "4" is for the 4 bounding box coordinate positions and 1 is object confidence, and nc is the number of

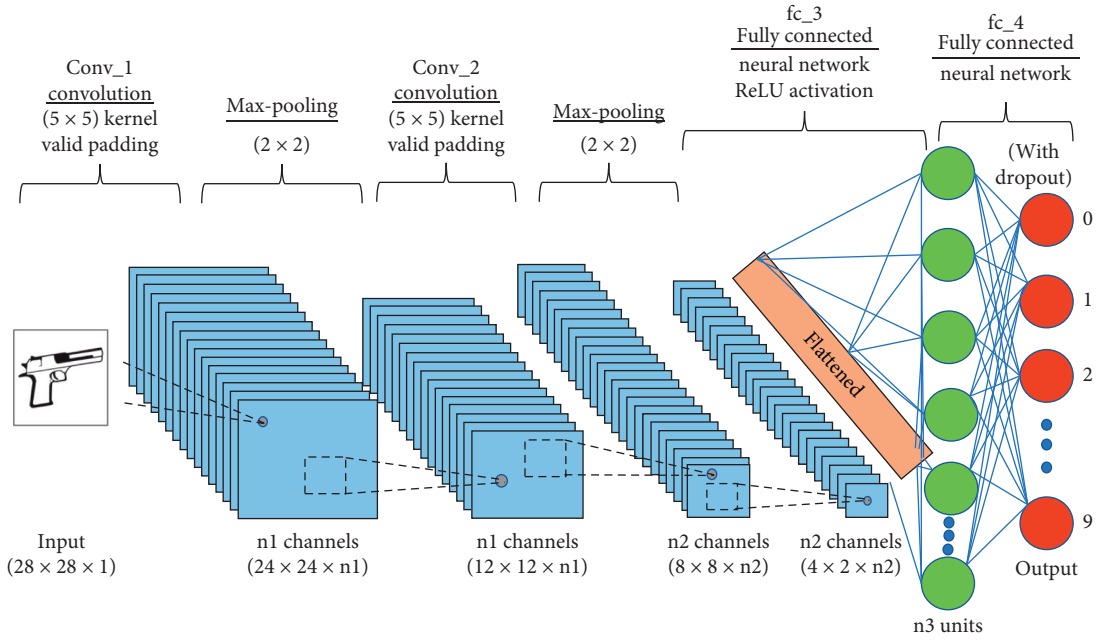

Figure 3: Feedforward convolutional neural network (CNN).

classes. The downsampling of the input image is for three scale predictions and is computed by strides 32, 16, and 8. The loss function over here is comprised on three sections, location error (L_{box}), confidence error (L_{cls}), and classification error (L_{obj}), as presented in (2).

$$\text{Loss} = L_{box} + L_{cls} + L_{obj}. \tag{2}$$

Literature suggests that YOLO v2 often struggled with small object detections. This happened due to loss of fine-grained features as the layers downsampled the input. In conclusion, YOLO v2 applies an identity mapping, concatenating feature maps from a previous layer to capture low-level features. However, YOLO v2's architecture was lacking some of the influential essentials that are encapsulated in most of state-of-the-art algorithms. The early models were lacking in the residual blocks, skip connections, and upsampling. On the other hand, YOLO v3 incorporates all of these. The detection of smaller objects can be seen from cumulative results demonstrated in Figure 8. We retrained both YOLO V2 and YOLO V3. Alternatively, we also conducted comparative analysis of the models with traditional CNN which was trained from the very scratch with null weights. The obtained results are summarized in Table 1.

The subsequent part of our research is based on the recording of location where the weapon was detected so that the alarm is generated. For this purpose, at backend we have also created a Database. A desktop application is also developed in order to provide connectivity with the database system. There are four attributes that are collected from the site where an object like weapon was detected. The collected information needs to be translated into a geographical format of longitude and latitude. For this purpose, geocoding was performed. It is the method of translating addresses to geographical details, longitude, and latitude, to map their positions. As it can be seen from the relational table provided in Figure 9, the attributes are latitude, longitude, time, and location where weapons were seen or identified. At backend DAO (Data Access Object) layer is also available to show the user the data from the database. It is component of Java Foundation Classes (JFC), which is a

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3/2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3/2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3/2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3/2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3/2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 4: Architectural details of DARKNET-53 layers [10].

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3/2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3/2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3/2	32 × 32
3×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3/2	16 × 16
4×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3/2	8 × 8
	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

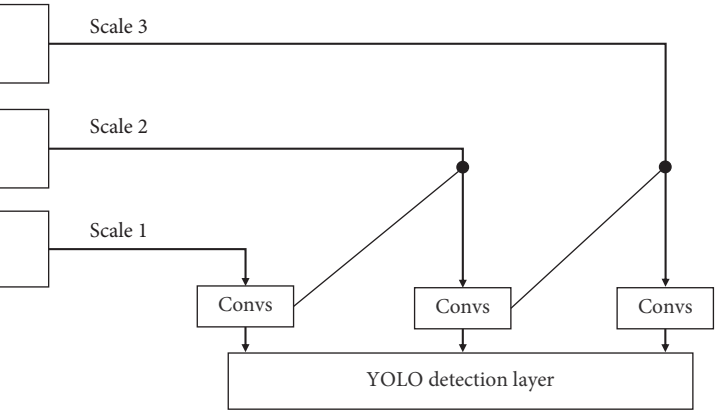

Figure 5: Architectural description of YOLO V3.

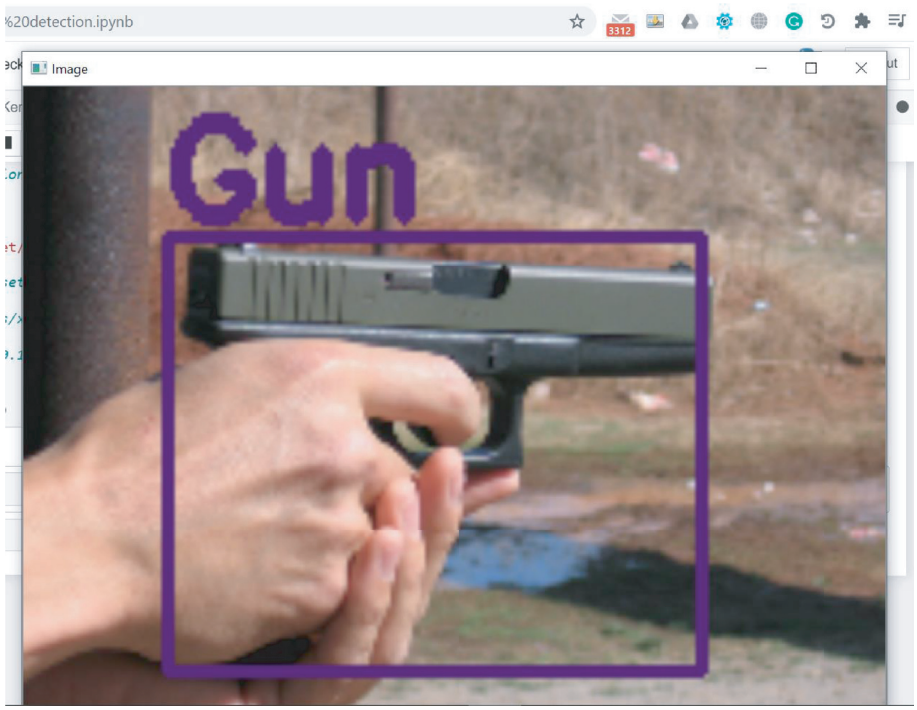

Figure 6: Bounding box around detected object; weapon category GUN.

GUI-providing API for Java programs. Swing provides packages that let us render our Java programs a complex collection of GUI components and it really is platform independent. Figure 10 presents the class diagram and implementation of DAO layer.

Our proposed system is further compared with the existing literature in Table 2. In [21], the proposed system includes CNN-based VGG-16 architecture as feature extractor, followed by state-of-the-art classifiers which are implemented on a standard gun database. The researchers

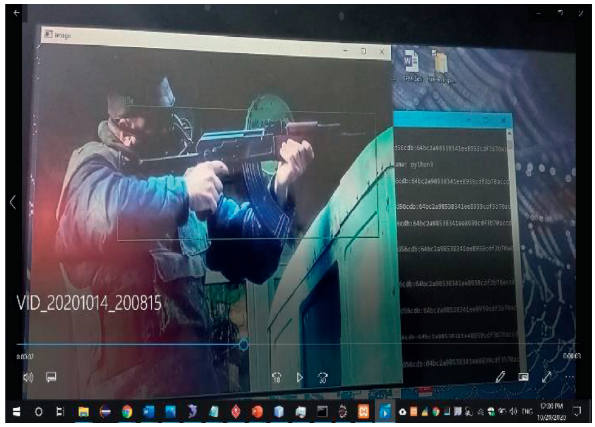

Figure 7: Real-time weapon detected from a video surrounded by bounding box. Weapon category rifle.

Figure 8: Cumulative result of detecting weapon with precision value.

Table 1: Experimental results for trained deep learning models.

S. no	Models	Accuracy
1	Traditional CNN	95
2	YOLO V2	96.76
3	YOLO V3	98.89

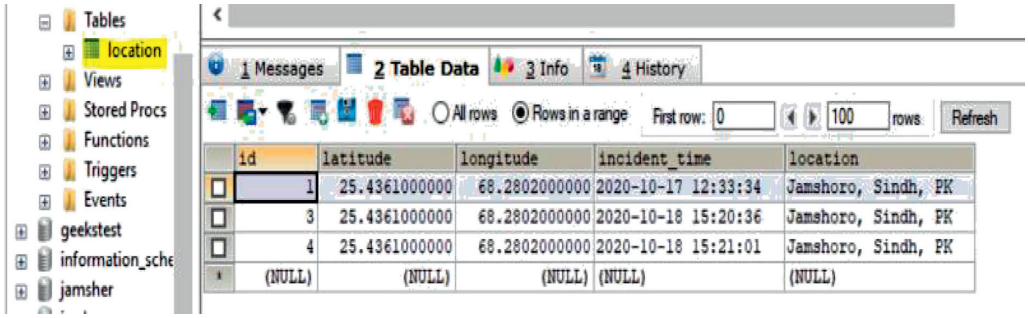

Figure 9: Image presenting the recorded database.

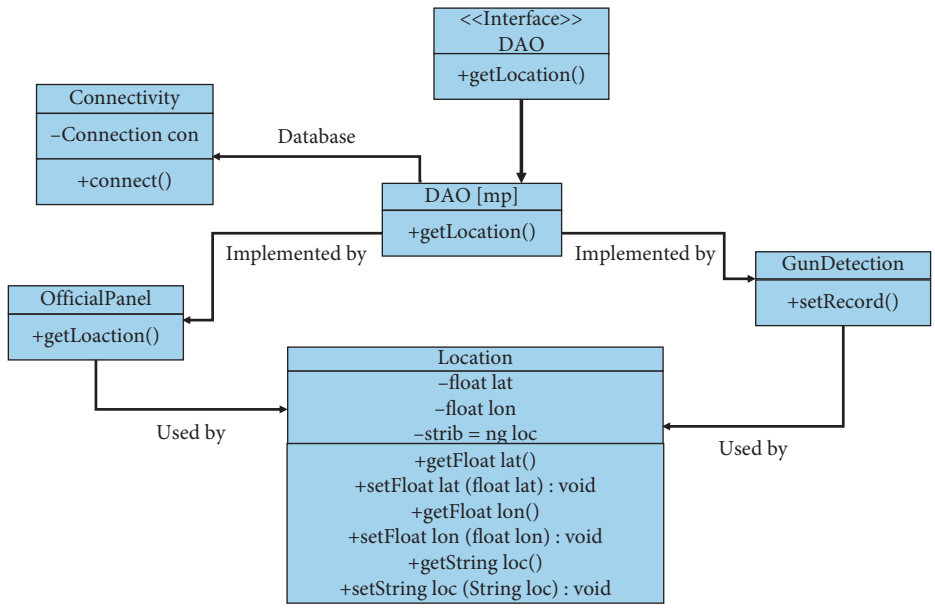

Figure 10: Class diagram for DOA layer.

Table 2: Comparison with the existing work.

S. no	Models	Dataset	Accuracy (%)
1	Our trained model YOLO V3	Image dataset collected for current research	98.89
2	Alexnet + SVM [22]	Gun video database [24]	95
4	Faster RCNN [23]	Streaming video	95.4
5	CNN VGG-16 [21]	IMDB	93.1

investigated four machine learning models, namely, BoW, HOG + SVM, CNN, and Alexnet + SVM, to recognize the firearms and knifes from a dataset of images [22]. Their work suggests that pretrained Alexnet + SVM performed the best. As it is evident from the previous studies, researchers have widely applied CNN and its variant for weapon or knife identification from CCTV videos [23]. It is obvious from Table 2 that the implemented YOLO v3 outperforms the rest of the other models.

5. Conclusion and Future Work

In this study, the state-of-the-art YOLO V3 object detection model was implemented and trained over our collected dataset for weapon detection. We propose a model that provides a visionary sense to a machine or robot to identify the unsafe weapon and can also alert the human administrator when a gun or a firearm is obvious in the edge. The experimental results show that the trained YOLO V3 has better performance compared to the YOLO V2 model and is less expensive computationally. There is an immediate need to update the current surveillance capabilities with improved resources to support monitoring the effectiveness of human operators. Smart surveillance systems would fully replace current infrastructure with the growing availability of low-cost storage, video infrastructure, and better video processing technologies. Eventually, the digital monitoring systems in terms of robots would fully replace current surveillance systems with the growing availability of cheap computing, video infrastructure, high-end technology, and better video processing.

Data Availability

The data are available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. A. Velastin, B. A. Boghossian, and M. A. Vicencio-Silva, "A motion-based image processing system for detecting potentially dangerous situations in underground railway stations," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 2, pp. 96–113, 2006.

[2] United Nations, *Office on Drugs and Crime, Report on "Global Study of Homicide"*, https://www.unodc.org/documents/data-and-analysis/gsh/Booklet1.pdf.

[3] P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, R. Jidhesh, and T. Vadivel, "Intelligent face recognition and navigation system using neural learning for smart security in internet of things," *Cluster Computing*, vol. 22, no. S4, pp. 7733–7744, 2019.

[4] V. Babanne, N. S. Mahajan, R. L. Sharma, and P. P. Gargate, "Machine learning based smart surveillance system," in *Proceedings of the 2019 Third International Conference on*

I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 84–86, IEEE, Palladam, India, December 2019.

[5] A. Joshi, N. Jagdale, R. Gandhi, and S. Chaudhari, "Smart surveillance system for detection of suspicious behaviour using machine learning," in *Intelligent Computing, Information and Control Systems. ICICCS 2019. Advances in Intelligent Systems and Computing*, A. Pandian, K. Ntalianis, and R. Palanisamy, Eds., vol. 1039, Berlin, Germany, Springer, Cham, 2020.

[6] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Engineering Applications of Artificial Intelligence*, vol. 67, pp. 226–234, 2018.

[7] 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.

[8] 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.

[9] S. Ahmed, A. Ahmed, I. Mansoor, F. Junejo, and A. Saeed, "Output feedback adaptive fractional-order super-twisting sliding mode control of robotic manipulator," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 45, no. 1, pp. 335–347, 2021.

[10] S. Ahmed, H. Wang, and Y. Tian, "Adaptive fractional high-order terminal sliding mode control for nonlinear robotic manipulator under alternating loads," *Asian Journal of Control*, 2020.

[11] S. Ahmed, H. Wang, and Y. Tian, "Adaptive high-order terminal sliding mode control based on time delay estimation for the robotic manipulators with backlash hysteresis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 2, pp. 1128–1137, 2021.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[13] A. Farhadi and R. Joseph, "Yolov3: an incremental improvement," *Computer Vision and Pattern Recognition*, 2018.

[14] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29573–29588, 2018.

[15] Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel, and F. Porikli, "Fast detection of multiple objects in traffic scenes with a common detection framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1002–1014, 2015.

[16] M. Grega, A. Matiolański, P. Guzik, and M. Leszczuk, "Automated detection of firearms and knives in a CCTV image," *Sensors*, vol. 16, no. 1, p. 47, 2016.

[17] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 148–155, IEEE, Waikoloa, HI, USA, January 2015.

[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3 D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[19] L. Pang, H. Liu, Y. Chen, and J. Miao, "Real-time concealed object detection from passive millimeter wave images based on the YOLOv3 algorithm," *Sensors*, vol. 20, no. 6, p. 1678, 2020.

[20] A. Warsi, M. Abdullah, M. N. Husen, M. Yahya, S. Khan, and N. Jawaid, "Gun detection system using YOLOv3," in *Proceedings of the 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, pp. 1–4, IEEE, Kuala Lumpur, Malaysia, August 2019.

[21] 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.

[22] S. B. Kibria and M. S. Hasan, "An analysis of feature extraction and classification algorithms for dangerous object detection," in *Proceedings of the 2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE)*, pp. 1–4, IEEE, Rajshahi, Bangladesh, December 2017.

[23] 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, 2019.

[24] V. Gun, "Database," http://kt.agh.edu.pl/grega/guns/.