Weapon Detection in Surveillance Videos Using Deep Neural Networks

Muhammad Ekmal Eiman Quyyum(✉) and Mohd Haris Lye Abdullah(✉)

Faculty of Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya,
Selangor, Malaysia
1171102157@student.mmu.edu.my, haris.lye@mmu.edu.my

Abstract. Object detection uses computer vision technique to identify and locate objects in an image or video. This feature can help to improve the security level as it can be deployed to detect a dangerous weapon with object detection methods. Driven by the success of deep learning methods, this study aims to develop and evaluate the use the deep neural network for weapon detection in surveillance videos. The YOLOv3 with Darknet-53 as feature extractor is used for detecting two types of weapons namely pistol and knife. The YOLOv3 Darknet-53 is further improved by optimizing the network backbone. This is achieved by adding a fourth prediction layer and customizing the anchor boxes in order to detect the smaller objects. The proposed model is evaluated with the Sohas weapon detection dataset. The performance of the model is evaluated in terms of precision, recall, mean average precision (mAP) and detection speed in frame per second (FPS).

Keywords: Deep neural network · Artificial Intelligence · Surveillance video · weapon detection

1 Introduction

Dangerous weapons are being used in criminal activities and terrorism. Therefore, implementation of weapon detection in a surveillance camera (CCTV) can improve the security level as it helps in automatically detecting weapon from the video feed. Current implementation for weapon detection using deep neural network provides sufficient high accuracy but suffers from poor detection speed and thus unable to work in real-time application. The trade-off between accuracy and speed are the main problem to implement the object detection in the real world especially for security purposes. Implementing weapon detection is an important requirement in surveillance video system. However, the weapon detection system faces difficulty to detect an object in low resolution surveillance footage. Furthermore, monitoring the CCTV for 24 hours requires a lot of manpower and due to human visual error, some security incidents may be missed.

The aim of this work is to develop and evaluate deep neural network model for a weapon detection system to achieve high accuracy and fast inference time. In addition, it aims to improve the deep neural model for a small weapon object. The system is developed based on TensorFlow framework and uses the YOLOv3 Darknet-53 model.

M. Y. bin Alias (Ed.): MECON 2022, AER 214, pp. 183–195, 2023.
https://doi.org/10.2991/978-94-6463-082-4_19

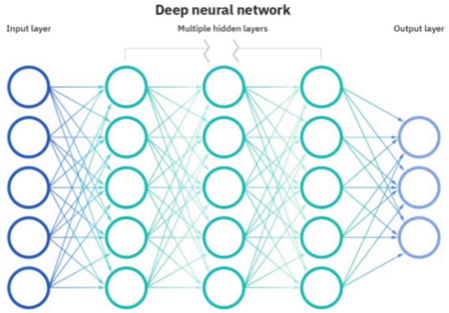


Fig. 1. A Deep Neural Network illustration

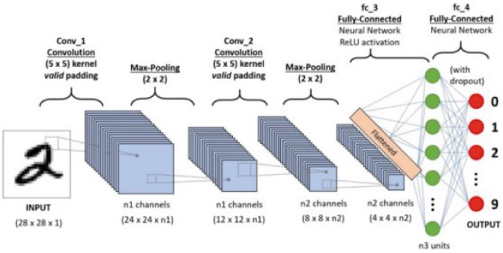


Fig. 2. A CNN Sequence to categorize the handwritten digits

1.1 Introduction of Deep Learning (DL)

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 [15]. Deep learning involves the use of Deep Neural Network (DNN) which involved multiple processing layers. The example of the deep neural network is shown in Fig. 1, it 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. CNN consists of filters as its basic computational component. As shown in Fig. 2, the filters are grouped into layer and multiple layers are stacked to form the model architecture. The filter applies convolution to extract meaningful information from the input image [14]. Figure 3 shows the output of convolution by using one specific predefined filter. In CNN, multiple filters are being used. The filter parameters are not predefined but are obtained through learning process.

The advantage of CNN is that it can automatically learns the filters from the training images [14]. The spatial pattern of an image learned by CNN model helps it to recognize and localize the object in the image. The example of CNN can be found in GoogLeNet, ResNet, AlexNet, MobileNets, GoogLeNet_DeepDream, VGGNet, ZFNet and LeNet [18].

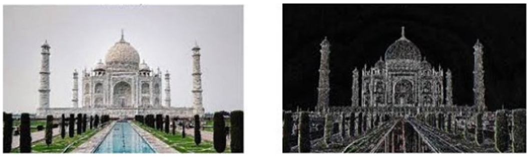


Fig. 3. Output of Convolution

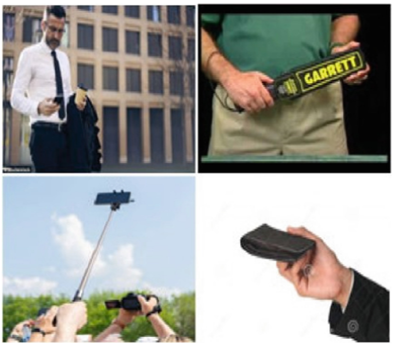


Fig. 4. Handled object similar to weapon [1]

2 Literature Review

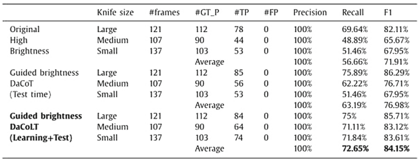
Detecting dangerous weapon from surveillance video is difficult. According to [2], the attention of security personnel monitoring the CCTV will deteriorate after 20 minutes. Paper [3] shows that after 12 minutes of continuous video monitoring, a security guard is likely to miss up to 45% of screen activity, and after 22 minutes, up to 95% of activity is missed. By implementing deep learning, the monitoring of the surveillance video feed can be automated to detect important events.

Among the proposed object detection solution, YOLO architecture is one of the most used method in real-time weapon detection [1, 4, 5]. YOLO architectures give a balanced performance in accuracy and faster inference time compared to other CNN models [1]. Besides, [6] uses Faster R-CNN with different feature extractors such as Inception-ResNetV2, ResNer50, VGG16 and MobileNetV2 and the performances are compared with YOLOv2 model. Only Faster R-CNN with Inception-ResNetV2 produce a better mAP than the YOLOv2. Faster R-CNN uses Region Proposal Network (RPN) to make the prediction more accurate, however the downside of this architecture is it reduces the inference time.

To improve weapon detection performance, relevant confusion object is added to the training set [1]. The weapon such as pistol and revolver are small, and it is likely to be confused with other small objects. Figure 4 shows sample of small hand-held objects that are likely to be confused as weapon. The YOLO model cannot achieve good result with small objects as it has been pre- trained with high quality training images [1][7].

In [8], YOLOv3 algorithm has shown good performance to detect small and large objects due to its network with three prediction scales. The performance of YOLOv3

Table 1. Comparison of different method pre- processing and their performance [9]



	Knife size	#frames	#GT_P	#TP	#FP	Precision	Recall	F1
Original	Large	121	112	78	0	100%	69.64%	82.11%
High	Medium	107	90	44	0	100%	48.89%	65.67%
Brightness	Small	137	103	53	0	100%	51.46%	67.95%
			Average			100%	56.66%	71.91%
Guided brightness	Large	121	112	85	0	100%	75.89%	86.29%
DaCoT	Medium	107	90	56	0	100%	62.22%	76.71%
(Test time)	Small	137	103	53	0	100%	51.46%	67.95%
			Average			100%	63.19%	76.98%
Guided brightness	Large	121	112	84	0	100%	75%	85.71%
DaCoLT	Medium	107	90	64	0	100%	71.11%	83.12%
(Learning+Test)	Small	137	103	74	0	100%	71.84%	83.61%
			Average			100%	**72.65%**	**84.15%**

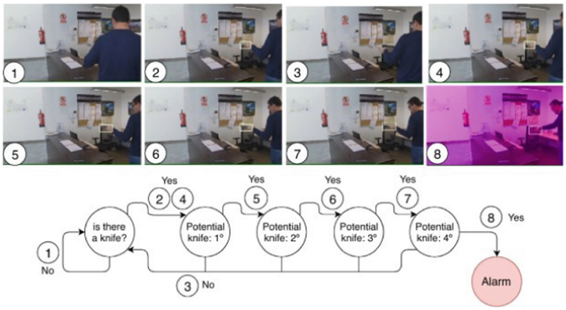


Fig. 5. The AATpI illustration, where the white box represents a true positive [9]

can be further improved by choosing the suitable number of candidate anchor boxes and their aspect ratio dimensions for each scale.

In the paper titled “Brightness guided pre-processing for automatic cold steel weapon detection in surveillance videos with deep learning” [9], a method called DaCoLT (Darkening and Contrast at Learning and Test stages) is proposed to improve the robustness of the weapon detection model with regards to variation of brightness. There are two data-augmentation stages used. The method applies training with brightness and visual quality adjustment. Table 1 illustrates different method of pre-processing result with different performances.

In order to evaluate weapon detection signal flagged by the model, the proposed AATpI (Alarm Activation time per Interval) [10] is used. AATpI is a metric that indicates how long it takes for an automated detection alarm system to identify at least *k* successful frames of true positives. Figure 5 shows the alarm detection system diagram. The crime alert can then be generated automatically and send to the authority through email or Short Message Service (SMS).

In summary, YOLOv3 model can be used for object detection improvement as it achieves good result in accuracy and fast detection speed. In comparison to CNN architectures such as Faster R-CNN and SSD, the complexity of its network increases the training time.

3 Approach

3.1 Design Requirements

YOLOv3 is an object detection algorithm based on convolutional neural network and uses the concept of bounding box regression [16]. Instead of processing region of interest (ROI) to detect the target object, it directly predicts bounding boxes and its classes for the image with a single stage architecture. YOLOv3 components consists of feature extractor, prediction layer and grid cells with anchor boxes. Given an image, the feature extractor compute the salient feature and it is passed to the prediction layer to produce the predicted bounding boxes.

3.1.1 Feature Extractor

The Darknet-53 is a feature extractor designed specifically for YOLOv3 and it consists of 53 convolutional layers as shown in Fig. 7. The Darknet-53 feature extractor or backbone is chosen mainly due to its high accuracy when compared to other models. Darknet-19 is a smaller model with 19 convolutional layers. Figure 6 [11], shows the performance of Darknet- 53 is higher when compared to Darknet-19 and ResNet-101. However, when compared with Darknet-19 as backbone, the detection speed of YOLOv3 with Darknet-53 is slower in terms of detection speed as measured in FPS (Frame Rate per Second).

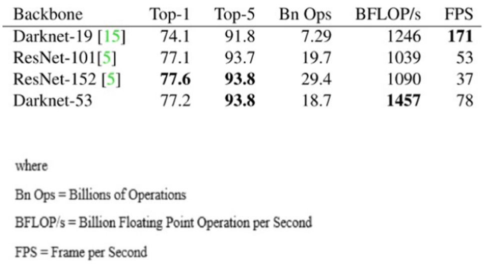
3.1.2 Bounding Box Regression

The bounding box is defined as the rectangular spatial box that defines the predicted object location in the image. Every bounding box has width (*b_w*), height (*b_h*), class (knife or pistol) represented by *c* and bounding box centre (*b_x, b_y*). Figure 8 shows the bounding box definition.

During image classification and localisation, the model will predict bounding box, class of the object, center of bounding box and probability of object in bounding box.

3.1.3 Grid Cells and Anchor Boxes

The YOLOv3 algorithm works by dividing the image into N grid cells with equal size ($S \times S$). Each grid cell has a set of predefined anchor boxes with certain height and



Backbone	Top-1	Top-5	Bn Ops	BFLOP/s	FPS
Darknet-19 [15]	74.1	91.8	7.29	1246	**171**
ResNet-101[5]	77.1	93.7	19.7	1039	53
ResNet-152 [5]	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78

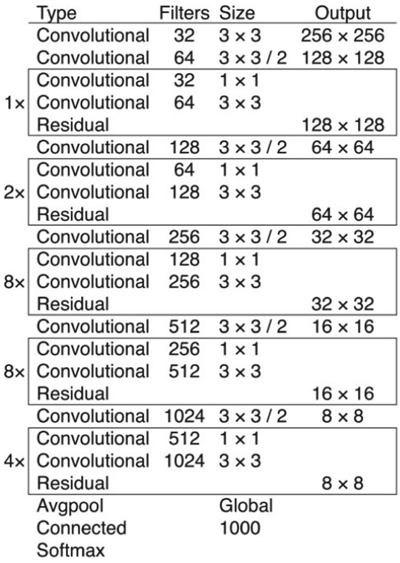
where

Bn Ops = Billions of Operations

BFLOP/s = Billion Floating Point Operation per Second

FPS = Frame per Second

Fig. 6. Comparison of Backbone [11]



	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			

Fig. 7. Architecture of Darknet-53 [11]

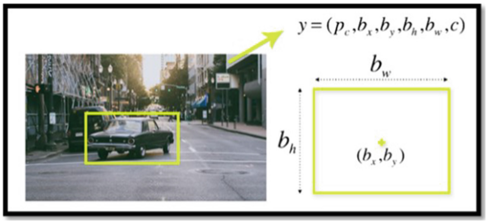


Fig. 8. Bounding box of detected object

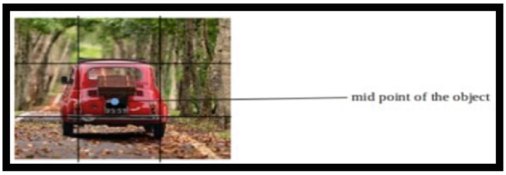


Fig. 9. Example of grid cells of size *(3 × 3)*

width. Each anchor box predicts the object label, probability of the object presents in the cell within the anchor box and the bounding box coordinate (*w, h, x, y*). For instance, in Fig. 9, the image is divided into 9 (3 × 3) grid cells. Each grid cell with anchor boxes, predicts the presence of the object in the cell. The size of the grid cell will influence the ability of the model to make prediction for small object. Therefore, smaller grid cell is used for detecting smaller object as shown in Fig. 10.

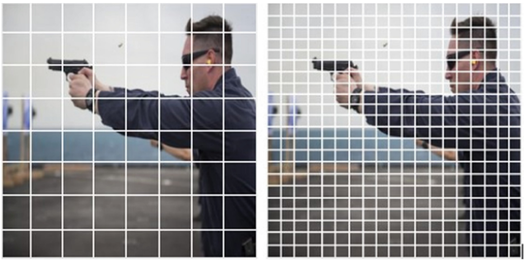


Fig. 10. Comparison between large grid cells and small grid cells

3.2 Optimization of YOLOv3 Darknet-53

Modification on YOLOv3 model has been made to improve the performance of the model for detecting gun and pistol. The modification is made by using custom anchor boxes, dataset expansion and adding extra prediction layer.

3.2.1 Custom Anchor Boxes

By default, YOLOv3 is designed for object detection in COCO (Common Objects in Context) dataset for 80 object categories. Hence, the model needs to be modified to accommodate the detection of object from two classes namely knife and pistol. In addition, anchor boxes need to be customized to detect the small knife and pistol object in the Sohas weapon detection dataset [12].

The custom anchor boxes are generated by using K- Means clustering on the Sohas training dataset. K-means clustering is an unsupervised algorithm to group similar data points together in order to discover the underlying pattern. This can be achieved by grouping the data into a selected number of clusters. A total of 4014 data points has been used to obtain anchor boxes dimension that cover the pistol and knife objects in the training image set. Next, the K-means clustering is used to group them into predetermined number of clusters. Experiment on the use of different number of clusters and its impact on the coverage of the ground truth object is shown in Fig. 11. The higher the number of clusters, the performance as measured by Average IoU (Intersection Over Union) becomes higher. The number of cluster N is chosen to be 9 since the curve is flattened at N=9. Therefore, 9 anchor boxes is defined for each grid cell and the anchor box parameters are obtained from the cluster's centroid.

3.2.2 Dataset

In order to evaluate the performance of the detection model, the Sohas weapon detection dataset [12] is used. The dataset contains 4014 images with weapons. The images are categorized based on the type of handheld weapon objects used, namely pistol and knife. A total of 3250 images is used for training and 764 images are used for testing the trained model. The number of images for the two classes is approximately balanced. A total of 1425 images from the pistol category and 1825 images from the knife category is used for training. The test set consist of 374 pistol images and 390 knife images.

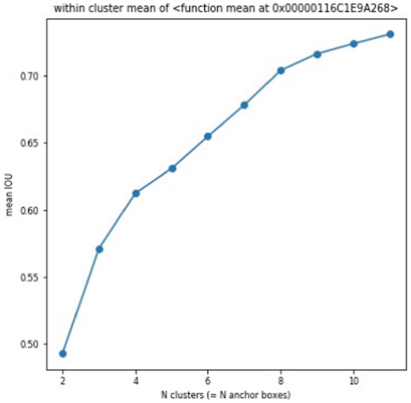


Fig. 11. Architecture of Darknet-53 [11]

The Sohas dataset lacks blurry image sample typically found in surveillance video frame. Therefore, the training and test samples is expanded with blurry images from surveillance video to obtain a larger and more diverse dataset. This can help the model to generalize better with images obtained from surveillance videos.

The additional images are obtained by extracting the images from the YouTube video. Five surveillance videos obtained from Closed-Circuit Television (CCTV) camera are downloaded from YouTube. One frame is extracted every ten seconds of the video and the image that contained weapons are used. This dataset is named as Dataset 1 with a total of 479 images. A total of 339 images is allocated for model training and 140 images for testing.

3.2.3 Addition of Prediction Layer

As the knife and pistol objects are seen smaller in the video, sometimes it can be undetected by the YOLOv3 model. We propose to add another prediction layer with smaller scale to the YOLOv3 backbone.

The default YOLOv3 model consists only of three prediction layers. The prediction layer Predict 1 with grid cells size 13 × 13 is used for detecting large object. The Predict 2 layer is used for detecting the medium size object and it has the grid cells of size 26 × 26. Predict 3 layer has the resolution 52 × 52 and the scale is still not suitable for detecting small object like knife and gun. We add another prediction layer (Predict 4) at grid cells size 104 × 104 to cater for the detection of smaller object. The Fig. 12 shows the Predict 4 layer added to the original YOLOv3 network.

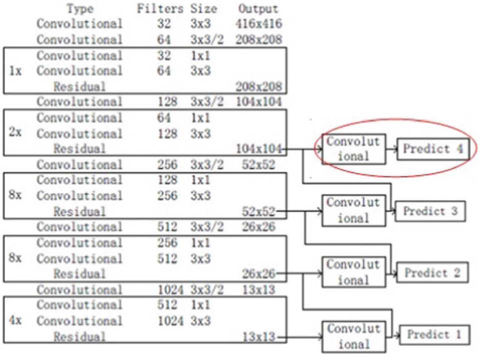
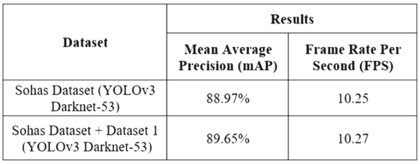


Fig. 12. Prediction layer Predict 4 added

Table 2. Comparison of performance for YOLOv3 with different dataset



Dataset	Results	
	Mean Average Precision (mAP)	Frame Rate Per Second (FPS)
Sohas Dataset (YOLOv3 Darknet-53)	88.97%	10.25
Sohas Dataset + Dataset 1 (YOLOv3 Darknet-53)	89.65%	10.27

4 Discussion of Findings

The YOLOv3 model is trained and evaluated with the Sohas dataset. In the second experiment trial, the surveillance video images from YouTube (Dataset 1) is added to the Sohas dataset and is used for model training and evaluation. Object detection performance in Table 2 shows improved result when the training images is augmented with additional surveillance video images.

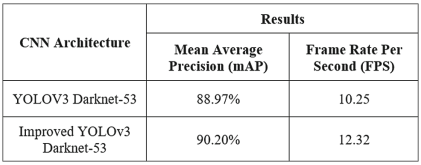
The next experiment evaluates the performance of the customized YOLOv3 model with one additional prediction layer (Improved YOLOv3 Darknet-53). Comparison of the performance of the original YOLOv3 model and the improved model is shown in Table 3. It is observed that the Mean Average Precision (mAP) improved from 88.97% to 90.20% with a slight reduction in detection speed from 10.25 to 12.32 Frame Rate Per Second.

4.1 Precision vs Recall Graph

In order to measure the performance of weapon object detection by the proposed model, the precision recall curve is used. The IoU threshold for positive class identification is set to 0.5. The Precision P and Recall R metric are calculated by using formula (1) and (2).

$$P = \frac{TP}{TP + FP} \quad (1)$$

Table 3. Comparison of performance for YOLOv3 with Improved YOLOv3 by using Sohas Dataset



CNN Architecture	Results	
	Mean Average Precision (mAP)	Frame Rate Per Second (FPS)
YOLOV3 Darknet-53	88.97%	10.25
Improved YOLOv3 Darknet-53	90.20%	12.32

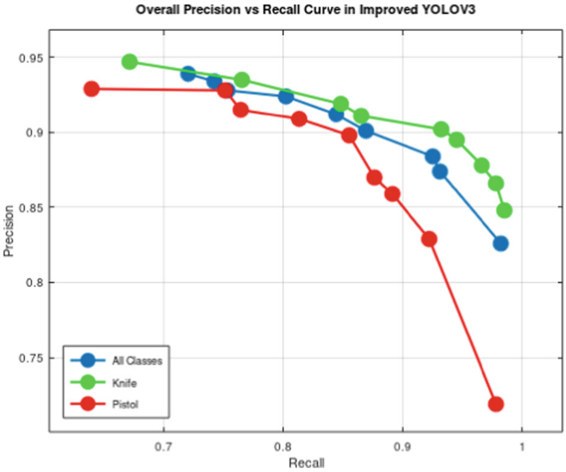


Fig. 13. Precision vs Recall graph for Improved YOLOv3

$$R = \frac{TP}{TP + FN} \tag{2}$$

The precision-recall (PR) curve graph for the knife and pistol detection is plotted based on the precision and recall with various confidence level of the model predictions. Figure 13 shows the PR curve for the detection of knife and pistol. The detection of combined knife and pistol classes is shown as well. The performance of the trained model on the knife object detection is better than pistol. This is due to the appearance of the pistol that is more similar to commonly handheld objects.

4.2 Result on Image and Video

Figure 14 and 15 shows sample object detection results using improved YOLOv3 Darknet-53. The results demonstrate the ability of the improved YOLOv3 model in detecting the small pistol and knife object. The model manages to detect the object successfully even with cluttered background and in low brightness scene.

When the model is tested on an actual surveillance scene, it manages to detect the weapon successfully. This shows the capability of the trained model for detecting small weapon object in a blurry scene as well.

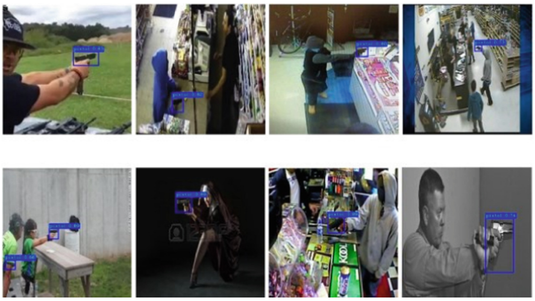


Fig. 14. Object detection for Pistol in different background using Improved YOLOv3 Darknet-53

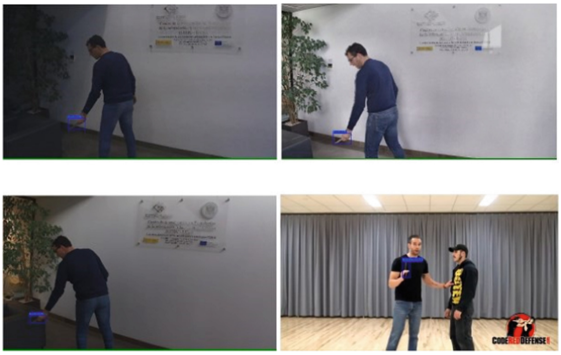


Fig. 15. Object detection for Knife with different brightness using Improved YOLOv3 Darknet-53

5 Conclusion and Future Work

The weapon detection using YOLOv3 architecture is implemented and trained with the Sohas weapon dataset. The original YOLOv3 model is improved and optimized to increase its performance for detecting small weapon object that include knife and pistol. The YOLOv3 anchor box parameters are optimized with the use of K-Mean clustering method. Both the original and improved YOLOv3 shows good accuracy on the Sohas evaluation dataset. The experiment results show that addition of extra prediction layer for detecting small object in YOLOv3 lead to better performances in mAP with slightly slower detection speed. Training data augmentation from actual surveillance video images helps the model gain slight performance improvement.

References

1. M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, “Weapon Detection in Real-Time CCTV Videos Using Deep Learning,” IEEE Access, vol. 9, pp. 34366–34382, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3059170.
2. M. M. Fernandez-Carrobles, O. Deniz, and F. Maroto, Gun and Knife Detection Based on Faster R-CNN for Video Surveillance, vol. 11868 LNCS. Springer International Publishing,2019.

3. J. L. Salazar González, C. Zaccaro, J. A. Álvarez- García, L. M. Soria Morillo, and F. Sancho Caparrini, "Real-time gun detection in CCTV: An open problem," Neural Networks, vol. 132, pp. 297–308, 2020, doi: https://doi.org/10.1016/j.neunet.2020.09.013.
4. S. Narejo, B. Pandey, D. Esenarro Vargas, C. Rodriguez, and M. R. Anjum, "Weapon Detection Using YOLO V3 for Smart Surveillance System," Math. Probl. Eng., vol. 2021, 2021, doi: https://doi.org/10.1155/2021/9975700.
5. T. S. S. Hashmi, N. U. Haq, M. M. Fraz, and M. Shahzad, "Application of Deep Learning for Weapons Detection in Surveillance Videos," 2021 Int. Conf. Digit. Futur. Transform. Technol. ICoDT22021, 2021, doi: https://doi.org/10.1109/ICoDT252288.2021.9441523.
6. R. M. Alaqil, J. A. Alsuhaibani, B. A. Alhumaidi, R. A. Alnasser, R. D. Alotaibi, and H. Benhidour, "Automatic Gun Detection from Images Using Faster R-CNN," Proc. - 2020 1st Int. Conf. Smart Syst. Emerg. Technol. SMART-TECH 2020, pp. 149–154, 2020, doi: https://doi.org/10.1109/SMART-TECH49988.2020.00045.
7. H. Jain, A. Vikram, Mohana, A. Kashyap, and A. Jain, "Weapon Detection using Artificial Intelligence and Deep Learning for Security Applications," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 193–198, 2020, doi: https://doi.org/10.1109/ICESC48915.2020.9155832.
8. M. Ju, H. Luo, Z. Wang, B. Hui, and Z. Chang, "The application of improved YOLO V3 in multi-scale target detection," Appl. Sci., vol. 9, no. 18, pp. 1– 14, 2019, doi: https://doi.org/10.3390/app9183775.
9. 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, doi: https://doi.org/10.1016/j.neucom.2018.10.076.
10. U. V. Navalgund and P. K. Priyadharshini, "Crime Intention Detection System Using Deep Learning," 2018 Int. Conf. Circuits Syst. Digit. Enterp. Technol. ICCSDET 2018, 2018, doi: https://doi.org/10.1109/ICCSDET.2018.8821168.
11. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, [Online]. Available: http://arxiv.org/abs/1804.02767.
12. F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," Knowledge- Based Syst., vol. 194, p. 105590, 2020, doi: https://doi.org/10.1016/j.knosys.2020.105590.
13. Anyoha, R. (2017, August 28). The History of Artificial Intelligence. Retrieved from sitn.hms.harvard.edu: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
14. aravindpai. (2020, February 17). CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning. Retrieved from analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/
15. Education, I. C. (2020, June 3). Artificial Intelligence (AI). Retrieved from ibm.com: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence
16. Karimi, G. (2021, April 15). Introduction to YOLO Algorithm for Object Detection. Retrieved from section.io: https://www.section.io/engineering-education/introduction-to-yolo-algorithm-for-object-detection/
17. Khandelwal, R. (2020, Jan 6). Evaluating performance of an object detection model. Retrieved from towardsdatascience.com: https://towardsdatascience.com/evaluating-performance-of-an-object-detection-model-137a349c517b
18. Kumar, A. (2021, November 7). Different Types of CNNArchitectures Explained: Examples. Retrieved from vitalflux.com: https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

19. Reynoso, R. (2021, May 25). A Complete History of Artificial Intelligence. Retrieved from g2.com: https://www.g2.com/articles/history-of-artificial-intelligence
20. Rosebrock, A. (2016, July 1). Intersection over Union (IoU) for object detection. Retrieved from pyimagesearch.com:
21. West, D. M., & Allen, J. R. (2018, April 24). How artificial intelligence is transforming the world. Retrieved from brookings.edu: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/#:~:text=Summary,transforming%20every%20walk%20of%20life.