Research Article

Moahaimen Abdullah*, Ahmed H. Y. Al-Noori, Jameelah Suad, and Emad Tariq

A multi-weapon detection using ensembled learning

https://doi.org/10.1515/jisys-2023-0060
received May 10, 2023; accepted November 16, 2023

Abstract: Recently, the level of criminals and terrorists using light weapons (such as knives and firearms) has increased rapidly around the world. Unfortunately, most current surveillance systems are still based mainly on human monitoring and intervention. For that reason, the requirement for a smart system for detecting different weapons becomes crucial in the field of security and computer vision. In this article, a novel technique for detecting various types of weapons has been proposed. This system is based mainly on deep learning techniques, namely, You Only Look Once, version 8 (YOLOv8), to detect a different class of light weapons. Furthermore, this study focuses on detecting two armed human poses based on ensemble learning techniques, which involve combining the outputs of different Yolov8 models to produce an accurate and robust detection system. The proposed system is evaluated on the self-created weapons dataset comprising thousands of images of different classes of weapons. The experiment results of this work show the effectiveness of ensemble learning for detecting various weapons with high accuracy, achieving 97.2% of mean average precision.

Keywords: artificial intelligence, computer vision, object detection, convolutional neural networks, You Only Look Once

1 Introduction

Object detection, a subclass of computer vision fields, is the technique concerned with detecting instances of semantic items of a certain class from input images or video [1]. This detection technique plays a significant role in various real-world applications, such as robotics, security, autonomous driving [2,3], including plant disease detection [4], fire detection [5], and text detection [6], brain tumor detection [7], and human emotion detection [8], and movement detection [9]. Object detection, in general, involves categorization and location. Object detection models are commonly responsible for returning the recommended locations of the items in a particular class, the class labels, and the confidence scores.

Recently, the average of criminal and terrorist activities, especially those that depend on using light weapons, have rapidly increased worldwide [10]. For example, according to the Institute for Economics & Peace, the average of humans who fall globally from terrorism reach almost 28% of each 5,463 deaths in 2021. As a result, the requirements for alternative efficient detection systems capable of replacing human surveillance are crucial. These types of systems have the ability to detect different kinds of weapons simultaneously,

* **Corresponding author: Moahaimen Abdullah**, Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, 10011, Iraq, e-mail: moahaimen@gmail.com
Ahmed H. Y. Al-Noori: School of Science, Engineering and Environment, University of Salford, Salford M5 4NT, United Kingdom; Computer Engineering Department, College of Engineering, Al-Nahrain University, Baghdad, 64040, Iraq, e-mail: a.h.y.al-noori@edu.salford.ac, ahmed.hani.al-noori@nahrainuniv.edu.iq
Jameelah Suad: Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, 10011, Iraq, e-mail: dr.jameelahharbi@gmail.com
Emad Tariq: Business Management/Marketing Department, Liverpool Hope University, Liverpool L16 9JD, United Kingdom, e-mail: tariqe@hope.ac.uk

as well as the pose status of weapon carriers [11,12]. In this article, a novel deep learning-based system for detecting multiple weapons has been proposed. A new deep learning model, You Only Look Once version 8 (YOLO-v8), has been employed for this purpose. In addition, different types of YOLO-v8 have been used to detect the pose of armed humans. Furthermore, an ensemble learning approach has been proposed to increase the accuracy of weapon and pose detection together.

Deep learning approaches, specifically convolutional neural networks (CNNs), have been widely used in object detection tasks due to their ability to effectively learn and extract features from images [13].

A deep learning framework is constructed based on the acquisition of multiple levels of data features. The fundamental principle of deep learning lies in learning from the primary representation of data [14]. A vector of density values per pixel, traits (clusters of edges), and custom shapes can be used to describe images; some are better at communicating the data than others [15]. A CNN comprises several layers, including convolution, pooling, activation function, dropout, fully connected, and classification layers, which work together to perform deep learning, and these represent the fundamental architecture of the deep learning idea [16]. One of the most popular CNN-based object detection models is You Only Look Once (YOLO). YOLOv5 represented the fastest and most accurate detection among other versions [17]. However, when YOLOVv8 was issued, it became the most efficient and had the fastest speed detection compared with YOLOv5 and other YOLOv families. The proposed system utilizes three YOLOv8 network (yolov8s, yolov8x, and yolov8xl) models to detect nine classes of weapons.

The proposed system was trained and evaluated on a self-collected dataset. This dataset comprises a huge number of images containing various types of weapons, including handguns, rifles, and knives. The dataset includes 11 classes (two for armed human pose and nine for weapons). Each class contains 1,500 photos, which means the number of photos is equally distributed among the classes. This equality distribution increases the performance of weapon detection without bias or overfitting.

On the other hand, the proposed system has used one of the ensemble learning methods, known as the weighted boxes fusion method (WBF). The WBF method approach improves the classifiers of state-of-the-art object detection methods, which simultaneously increases the detection accuracy of various weapons. Moreover, WBF can be applied in real-time scenarios, making it a valuable tool for security fields and law enforcement.

In addition, estimating the person's pose is essential for different fields, such as action detection and sports [18]. For that reason, the proposed system has the ability to detect two different poses for armed persons: The aiming pose is considered highly dangerous as it indicates that the person is preparing to shoot. In contrast, the standard pose indicates that the person carries a weapon but may lead to firing later.

The experiment results of this work showed that the usage of the ensemble learning method achieved higher accuracy in detecting weapons than the other methods, with a mean average precision (mAP) of 97.2%. Therefore, the findings of this study have the potential to enhance security and defense measures in various settings, including airports, public areas, and military facilities. In general, the contribution of this work can be summarized in four main points: (1) A novel detection system for different types of light weapons based on the last version of the YOLO approach has been proposed. This system had the ability to detect various types of firearms simultaneously for any input video stream (live or recorded). (2) The detection system can recognize armed human posing, whether this posing is a standard or aiming pose. (3) The effectiveness of ensemble learning (including WBS) in improving the accuracy and robustness of the system detection is investigated. (4) A comprehensive dataset of thousands of images of different types of light weapons is provided.

The related works are presented in Section 2, the proposed detection system is shown in Section 3, followed by the experimental setup and results in Sections 4 and 5, respectively, and finally, the conclusions are presented in Section 6.

2 Related works

Olmos et al. [19] explored automatic handgun detection using the faster R-CNN framework, achieving an accuracy of 84.21%. However, their approach is computationally intensive due to the faster R-CNN framework and the separate region proposal mechanisms. In contrast, our ensemble learning model efficiently integrates multiple learning algorithms, reducing computational load while achieving a higher accuracy rate, thereby outperforming the methods presented by.

Fernandez-Carrobles et al. [20] also employed the faster R-CNN approach but faced challenges with slower inference speeds due to resource intensiveness. Unlike their model, our ensemble learning approach significantly enhances detection speed and overall performance without compromising on accuracy, making it a superior choice for real-time scenarios.

Verma and Dhillon [21] utilized transfer learning with the faster R-CNN model, achieving 93% accuracy. However, their model required immense computational resources, limiting real-time performance. Our ensemble learning model addresses these setbacks by efficiently processing data in a singular pass, promising swifter detection times and more versatile applications.

Gelana and Yadav [22] presented a method based on a CNN classifier via a sliding window approach, achieving a 93.84% accuracy. The reliance on high-quality CCTV footage and the sliding window technique pose significant limitations. In contrast, our ensemble learning model adapts to varying data quality and object dimensions, ensuring effective weapon detection in diverse practical applications.

Dwivedi et al. [23] utilized the VGG-16 network for weapon detection, achieving up to 99% accuracy for pistol detection. However, their focus on specific weapon types and the computational demands of the VGG-16 network limits its real-time applicability. On the other hand, our ensemble learning model offers broader weapon detection capabilities and efficient processing, proving its superiority in real-time scenarios and on devices with varying computational strengths.

The utilization of YOLOv8 models and the WBF technique in the presented study shows multiple improvements compared to the aforementioned approaches. Notably, the fusion of predictions from various models has the potential to enhance detection accuracy. In addition, this approach effectively addresses inaccuracies in predictions, inclusively handles multiple predictions, promotes consensus among models, and improves the handling of confidence scores. This facilitates a more exhaustive and reliable identification of objects, representing a noteworthy advancement compared to prior methodologies. Nonetheless, conducting a comprehensive evaluation by applying your approach to the same dataset with uniform conditions compared to alternative models would provide additional substantiation for these benefits.

3 Methods and materials

In this work, YOLOv8 in conjunction with weighted box fusion (WBF) had been presented over traditional methods due to the need for processing large and complex types of datasets. The time taken to detect and classify weapons in a scene can be crucial, especially in high-risk environments where timely response is paramount. YOLOv8 and WBF represent an optimal methodology as, with high-end computational infrastructure, these models can be trained in a reasonable amount of time.

YOLOv8 [24], part of the YOLO family of object detection models, excels at dealing with large-scale image data and offers real-time detection capabilities. Its unique architecture, which treats object detection as a single regression problem, allows it to capture temporal relationships between different objects in an image.

Complementing this, WBF is an ensemble method that improves the accuracy of bounding box predictions by considering the weights of boxes produced by different models. Its ability to effectively fuse multiple detections into a single, accurate bounding box prediction makes it an invaluable tool for handling complex scenes with overlapping detections.

Together, YOLOv8 and WBF provide a powerful combination for real-time object detection tasks, making them ideal for applications such as weapons detection in surveillance footage.

3.1 The YOLO network model for detecting objects

Joseph Redmon introduced YOLO, a real-time object identification system, in 2016. When Alexey Bochkovskiy started working on the YOLO algorithm, he wrote an article about it after making little progress. Then a sequence of YOLOs, YOLOV2, YOLOV3, and YOLOV4 took place.

On May 30, 2020, the Ultralytics LLC team released YOLOV5 [25] while YOLOV4 was still in development. YOLOV5, which utilized the Tesla P100, outperformed YOLOV4 by achieving 140 FPS; compared to its predecessor, the device can achieve a frame rate of 50 FPS. Although the architecture and benefits of YOLOV4 and YOLOV5 are similar, YOLOV5 is easier to train and use for object detection [26].

The YOLO algorithm aims to achieve enhanced accuracy and efficiency in object detection compared to traditional CNNs [26]. One can accomplish this by rapidly processing the image and treating the detection of objects as a unified operation. Yolov5 consists of several CNN networks; one of these important networks is the path aggregation network (PANET), for instance, segmentation (PANET) [27]. The pipeline applies a threshold to the output by utilizing the model's confidence level to reduce input, and subsequently processes it through a solitary CNN. The image is divided into multiple subregions using YOLO. Each entity is allocated a set of five anchor boxes to perform the detection task The selection of the zone with the highest probability is determined through the computation of the likelihood of a specific object. [26]. The seventh version of the YOLO family is known as YOLO-v7 [28]. The YOLO object recognition model initially connected bounding box estimates with object identification in a single end-to-end differentiable network. Compared to the preceding YOLO models, the YOLO-v7 model is lighter and easier to use because it was the first to be created using the PyTorch framework [29].

In this work, the most recent iteration of the YOLO family, YOLO-v8 has been adopted [24]. The object recognition model YOLO initially combined bounding box estimates with object identification in a single end-to-end differentiable network. For instance, according to the results obtained on the COCO dataset, YOLOv8 demonstrates a notable level of accuracy. As an illustration, the medium YOLOv8m model achieves an mAP of 50.2% when evaluated on the COCO dataset.

One example of an anchor-free model is YOLOv8. This suggests that it effectively predicts the center of an object rather than its displacement from a known anchor box. Anchor-free detection speeds up nonmaximum suppression (NMS) [30] by reducing the number of box predictions [31], a challenging postprocessing phase that, following inference, the system sorts through candidate detects. This technique divides the image into regions before determining each zone's bounding boxes and probabilities. The projected probabilities are used to weigh these boundary boxes [25].

The YOLOv8 model improves upon its predecessors by making changes to the convolutional layers in its stem and modifying the primary building block. Specifically, the first convolution in the stem, which was previously a 6 × 6 kernel size, has been replaced with a 3 × 3 kernel size, and Coarse-To-Fine (C2f) is now used in place of C3. The updated architecture of the bottleneck consists of concatenating two 3 × 3 convolutions connected by residual connections in C2f. Meanwhile, C3 only utilizes the output from the previous Bottleneck.

The YOLOv8 model also reverts to the ResNet block established in 2015, with the first convolution's kernel size changed from 1 × 1 to 3 × 3. Features are added directly to the neck without knowing the channel size, resulting in a decrease in the number of parameters and overall tensor size. Figure 1 shows the efficient layer aggregation networks.

While model architecture has received a lot of attention in deep learning research, YOLOv5 and the training procedure of YOLOv8 play a pivotal role in determining its effectiveness. The YOLOv8 algorithm demonstrates improved photo enhancement capabilities through online training. At each epoch, the model sees a slightly different selection of the photographs sent to it. One of these augmentations is mosaic enhancement. The model combines four photos and must discover objects in novel locations, partially occluded, and against various background pixels [24].

One of the advantages of the YOLOv8 network is its high detection accuracy, quick detection speed, and lightweight properties. The benchmark model, YOLOv8l, an extended model, YOLOv8x, and two simplified preset variants, YOLOv8s and YOLOv5m, are four main models in YOLOv8 [32]. In this work, the model

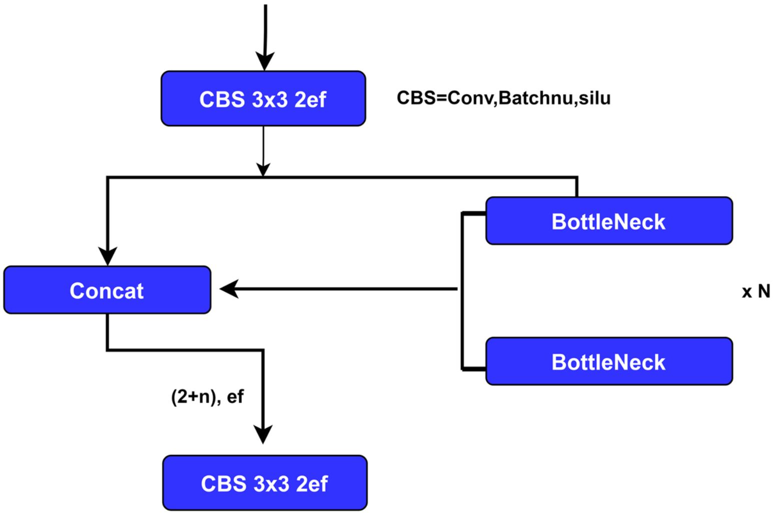

Figure 1: Extended efficient layer aggregation networks.

Yolov8x6 was utilized. The main differentiations among them primarily reside in the varying quantity of feature extraction modules and convolution kernels found in different segments of the network, as well as in the magnitude and quantity of model parameters.

3.2 WBF

To enhance the detection accuracy of the YOLOv8 models, the WBF technique has been employed, which is known as a method that fuses bounding box predictions for the same image from *N* different models. The program searches a list of fused boxes for a matching box by iteratively reviewing the predicted boxes from each model in a cycle. The term “match” refers to a box that significantly overlaps the box in question intersection over union (IoU) > threshold (THR). It adds the box to a list of fused boxes and box clusters and then updates the box coordinates and confidence score in the fused box using all T boxes gathered in the cluster. It uses fusion equation (1) that considers the corresponding boxes’ confidence scores. The fused box’s coordinates are weighted sums of the boxes that make it up, and its confidence score is set to be the average confidence of all the boxes that make up the box. Confidence scores in the fused list are rescaled once all boxes have been processed and the final boxes are acquired.

WBF becomes helpful, especially when all models predict the boxes incorrectly. NMS (or soft-NMS), then, will only leave one inaccurate box in this situation. However, WBF will fuse all anticipated boxes. Like WBF, the nonmaximum weighted (NMW) technique, as shown in earlier studies [33,34], does not alter confidence scores. Instead, NMW weighs the boxes based mainly on the IoU value. NMW does not produce the best outcomes for model ensembles since it also requires knowing how many models forecast a particular box in a cluster.

Each box in the fused box has its coordinates and confidence ratings updated by the WBF algorithm. The average confidence of all the boxes that comprise the fused box is its confidence score. The algorithm calculates the fused box’s coordinates by taking the weighted sum of the coordinates of the individual boxes that make it up. The weights are determined based on the confidence scores of the corresponding boxes. As a result, higher confidence boxes contribute more to the coordinates of the fused box than the lower confidence boxes, as shown in equations (1)–(3).

$$C = \frac{\sum_{i=1}^{T} C_i}{T}, \tag{1}$$

$$X1{,}2 = \frac{\sum_{i=1}^{T} C_i \times X1{,}2_i}{\sum_{i=1}^{T} C_i}, \tag{2}$$

$$Y1{,}2 = \frac{\sum_{i=1}^{T} C_i \times Y1{,}2_i}{\sum_{i=1}^{T} C_i}, \tag{3}$$

where C is the confidence, C_i is the confidence score of the ith bounding box. $X1{,}2_i$ and $Y1{,}2_i$: These represent the coordinates of the ith bounding box. In object detection, a bounding box is often represented by two sets of coordinates: ($X1$, $Y1$) for the top-left corner and ($X2$,$Y2$) for the bottom-right corner. So, $X1{,}2_i$ would be the X-coordinates for the ith bounding box, and similarly, $Y1{,}2i$ would be the Y-coordinates

After processing each box, the WBF algorithm rescales the confidence scores in the Fused list. To accomplish this, the number of models N is divided by the number of boxes in a cluster and multiplied by the confidence score by that number. A few boxes in a cluster could mean that only some models could forecast it. Therefore, it is necessary to lower the confidence scores in certain situations. This can be done by applying one of three equations, equations (4)–(6).

$$C = C \times \frac{\min(T, N)}{N}, \tag{4}$$

or

$$R = C \times \frac{T}{N}, \tag{5}$$

$$\text{AP} = \frac{1}{C} \sum_{K=1}^{N} P(K) \Delta R(K). \tag{6}$$

4 The proposed system

In the proposed system, the initial phase involves obtaining a live video stream from a strategically located camera. Following this, the captured data undergoes a preprocessing stage to prepare these data for the subsequent stage, where the WBF algorithm is employed. This fusion stage plays a critical role in boosting detection accuracy and enhancing the assessment's confidence. A unique feature of the proposed system is the capability to trigger an alarm when it identifies threatening poses, thereby alerting security personnel. The presented multi-weapon detection system (MWDS) not only enhances video analytics but also ensures prompt responses to potential threats. An implementation and the involved algorithms are discussed later in this article. Figure 2 presents the visual illustration of the system architecture.

4.1 Dataset construction and preprocessing

In this work, a unique dataset has been utilized to develop the system to address the absence of uniform weapon datasets. The construction of this dataset occurred through a bifurcated process. The study's initial phase centered on acquiring a heterogeneous array of nine distinct weapons that were obtained in different situations and recorded in a video format. In addition, the subsequent phase of the study focused on two distinct categories of human postures associated with handling and employing weapons, namely, STANDARD_POSE and AIMING_POSE. To guarantee the accurate identification of objects, each constituent within the dataset was annotated by utilizing bounding boxes through the RoboFlow platform. The ultimate dataset, tailored for YOLOV8, comprised 13,500 tagged and annotated images, representing each weapon class more than 1,500 images. Numerous visuals comprised various elements, necessitating distinct annotations for each item. Implementing equal distribution of weapons and balanced class representation aimed to enhance

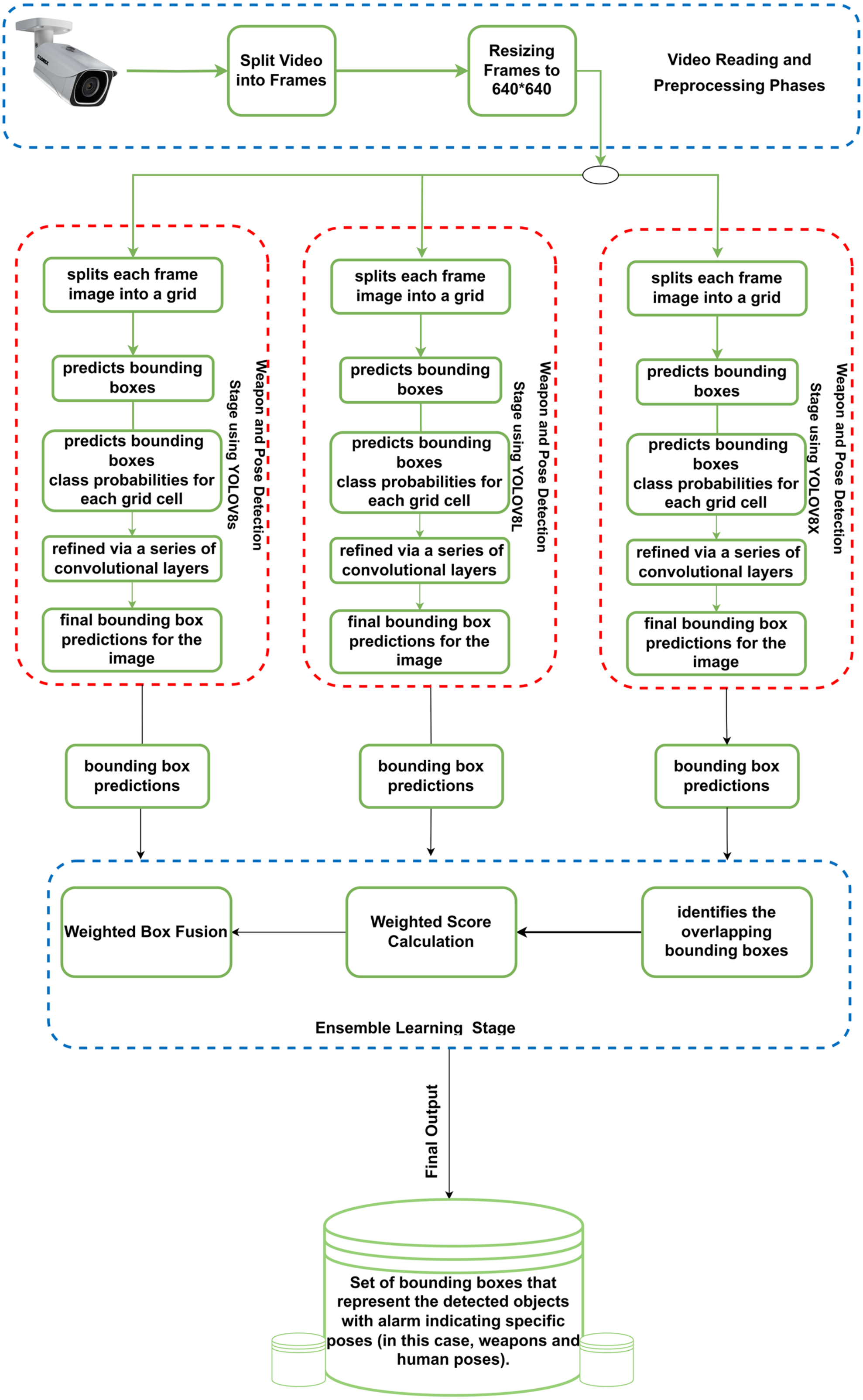
Split Video into Frames
Resizing Frames to 640*640
Video Reading and Preprocessing Phases
splits each frame image into a grid
predicts bounding boxes
predicts bounding boxes class probabilities for each grid cell
refined via a series of convolutional layers
final bounding box predictions for the image
Weapon and Pose Detection Stage using YOLOV8s
Weapon and Pose Detection Stage using YOLOV8L
Weapon and Pose Detection Stage using YOLOV8X
bounding box predictions
Weighted Box Fusion
Weighted Score Calculation
identifies the overlapping bounding boxes
Ensemble Learning Stage
Final Output
Set of bounding boxes that represent the detected objects with alarm indicating specific poses (in this case, weapons and human poses).

Figure 2: Workflow of the proposed system.

training outcomes, as balanced datasets have resulted in superior performance. For optimal clarity, the image resolution was established at 1,024 × 1,024 pixels; Figure 3 shows the dataset-building steps.

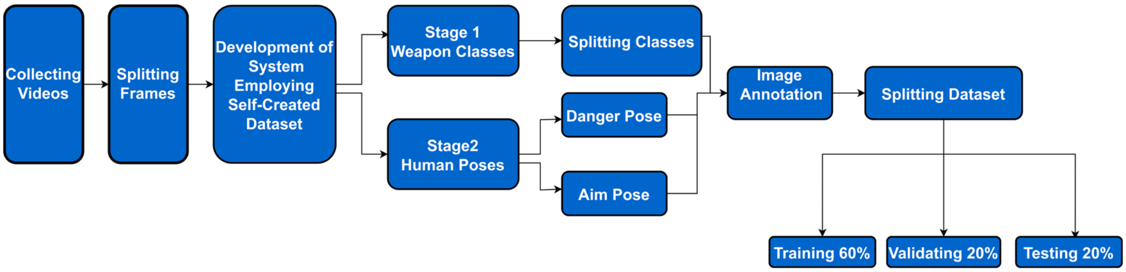

Figure 3: The dataset-building step.

The system under consideration initiates by processing video data, which is partitioned into frames to facilitate efficient manipulation. The preprocessing stage involves resizing images to a standard size of 640 × 640 pixels, utilizing the advanced object detection capabilities of YOLOv8. The advanced feature extraction capabilities of YOLOv8 obviate the necessity for manual feature engineering or intricate data transformations. In addition, it incorporates pre-built augmentation techniques, such as random scaling, rotations, and translations. The aforementioned characteristics improve the model's capacity to generalize, mitigate the likelihood of overfitting, and streamline preprocessing procedures. The utilization of YOLOv8 by the system enhances the performance of object detection while simultaneously preserving computational efficiency and ensuring superior outcomes.

4.2 The Yolov8 network model for object detection

The YOLOv8.0s model has three main components: (1) the backbone, (2) the head, and (3) the detection layer. The backbone is responsible for extracting features from the input image. On the other hand, the head is responsible for predicting the bounding boxes and class probabilities for each object in the image. Finally, the detection layer is responsible for combining the predictions from the head and generating the final output.

In phase 1, the YOLOv8s backbone is first fed with an input image measuring 640 × 640 pixels. The structure mentioned earlier comprises 53 convolutional layers, encompassing 22.8 million parameters. The initial step involves downsampling the image by a factor of 2 through a max pooling layer, thereby decreasing the image dimensions to 320 × 320 pixels. Following this, the reduced resolution image undergoes a sequence of convolutional layers, which facilitate extracting its distinctive features. Following the procedure mentioned earlier, the resultant image transforms into a tensor with dimensions of 10 × 10 × 1,024, commonly referred to as the "Bottleneck input." This tensor is then directed toward the initial concatenation layer.

In phase 2, during this stage, the input at the Bottleneck layer is subjected to convolution with 1,024 filters of dimensions 3 × 3 and a stride of 1. The aforementioned procedure results in a decrease in the number of parameters of the model, thereby producing a tensor of dimensions 1 × 1 × 1,024, commonly referred to as the "Bottleneck output." In conjunction with the ultimate convolutional layer derived from the underlying network, the result is transmitted to the subsequent concatenation layer. Upon conducting a sequence of convolutional operations, the concatenated resultant feature map, also known as scale 1, is acquired.

In the third phase, the Head network applies a sequence of convolutional layers to the feature map obtained from phase 2. These layers facilitate the extraction of more intricate features from the input. Concurrently, the fully connected layers utilize the aforementioned input to make predictions regarding both object-bounding boxes and their respective classes. The prognostications derived from the topmost neural network are subsequently transmitted to the third stratum of concatenation. Upon conducting supplementary convolutional operations, the amalgamated feature map, specifically of scale 2, is derived.

The fourth phase refers to the fourth stage of a particular process or project, where the detection layer amalgamates the concatenated feature map obtained from phase 3 to produce a roster of bounding boxes and class probabilities for every object present in the image. Subsequently, an NMS technique is employed on the aforementioned list to eliminate intersecting bounding boxes. The feature maps of varying dimensions (e.g., 320 × 320, 160 × 160, 80 × 80) generated during the process stages are accountable for detecting objects of diverse sizes employing a bounding box. As a result, a set of three bounding boxes are predicted for each feature map at every location. The ultimate model provides a forecast of the whereabouts of the targeted entities through a bounding box and an abjectness score. The following summarizes the YOLOV8 algorithm:

Input: Video data, classes (weapon, poses)

Output: List of bounding boxes with recognized
class (weapon, poses)

Frame Preparation:

Video is split into frames, resized to 1,024 × 1,024 pixels, and preprocessed for quality enhancement.

Region of Interest (ROI) Detection:

YOLOv8X's backbone, with 52 layers and 10.2 M parameters, grids each frame to identify object origin points.

Bounding Box Prediction:

YOLOv8X's head, having 10 layers and 3.7 M parameters, predicts bounding boxes and class probabilities for each grid cell.

Bounding Box Refinement:

IOU and non-max suppression refine bounding boxes, selecting those with the highest probability.

Iteration:

Steps 2–4 are repeated, with the head's 512 output channels enabling up to 512 bounding box and class probability predictions per video frame, until all bounding boxes are processed.

4.3 WBF

As mentioned earlier, WBF ensemble learning is a technique that combines the outputs of multiple YOLO models to improve the accuracy of object detection. The process works by first training three YOLO models, each with a different input size: Yolov8s (128 × 128), Yolov8l (256 × 256), and Yolov8x (512 × 512). Once the models are trained, they are used to generate predictions for a test set of images. The predictions from the three models are then combined using a weighted voting scheme to produce a final set of predictions.

In a setting where three YOLOv8 models are operating concurrently, each model will independently analyze an input image and subsequently produce a unique set of bounding boxes that signify detected objects. The resulting output will also provide the class labels for each object and the confidence scores for each detection. Once these bounding boxes are received, the WBF algorithm begins its fusion process.

The WBF algorithm plays a role, in combining detection results from models. It begins by gathering all the bounding boxes produced by the three models. This creates a collection of bounding boxes that represent overlaps and differences for the objects identified in the image. Since there can be variations, in the detection outputs, the primary goal of WBF is to address these inconsistencies and obtain a set of bounding boxes.

The WBF algorithm operates through two essential steps: box grouping and box fusion. Box grouping forms the first phase of the process. Here, bounding boxes that have a substantial level of overlap are gathered into groups. This overlap is quantified by a metric known as IoU, which calculates the area of overlap between two bounding boxes as a proportion of their total area. By using the IoU metric, boxes that are highly likely to denote the same object are identified and clubbed together.

Each bounding box within this group is then assigned a weight. Typically, these weights correspond to the confidence scores of the individual detections, but they can also be modified based on the reliability and performance of each YOLOv8 model. Higher weight values can be assigned to models that have exhibited superior performance in previous detection tasks, thereby allowing their outputs to carry more influence in the fusion process.

Following the box grouping and weight assignment, the second phase of WBF, box fusion, takes place. In this stage, a new bounding box is created for each group of boxes, representing a “consensus detection” that summarizes the collective insight from the grouped boxes. The spatial coordinates of this new box are calculated as the weighted average of the coordinates of the boxes in the group, with the weights serving to reflect the relative confidence in each detection.

The final output of the WBF process is a refined set of bounding boxes. These boxes are an optimized and harmonized representation of the detection outcomes from the three models. By reducing redundancy and emphasizing agreement among the models, WBF enhances the overall precision and reliability of object detection tasks. The use of WBF is thus a strategic approach in multi-model settings, driving toward the goal of achieving more accurate and robust object detection.

5 Results and discussion

In this section, the results of this work are presented and, at the same time, given a comprehensive discussion. As mentioned earlier, the experiments in the study involve nine distinct types of weapons and two different kinds of armed human positions. Three portions of the dataset were utilized for training the network: 70% for training, 10% for testing, and 20% for validation, as illustrated in Table 1, which represents the number of images used in each training, validating, and testing, respectively.

Table 1: Dataset for weapons

Training	Validating	Testing
9,400	2,700	1,400

The input image utilized to train the Yolov8 model was 640 × 640 pixels in size and was formatted in RGB. It received training during 150 epochs.

5.1 Experiments evaluation metrics

Compared to comparable works, the total mAP is 96%, which is relatively high. The assessment of item detection’s memory and accuracy is another primary focus of this work. Equations (7)–(9) are definitions for mAP, recall, and precision, respectively, where P is referred to precision and R is referred to recall. K is an index used to sum over all the classes or instances from 1 to N, where N is the total number of classes or instances, C is the number of object categories, N is the number of IoU thresholds, k is the IoU threshold, and so on.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{7}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{8}$$

$$\mathrm{mAP} = \frac{1}{C}\sum_{K=1}^{N} P(K)\Delta R(K). \tag{9}$$

Various classification models. These matrices offer a range of performance evaluation metrics and parameters, including the confusion matrix, that can be utilized to evaluate the effectiveness of classification models.

The model has undergone training. The term "true positive" (TP) denotes the number of positive predictions that are correct. The metric used to quantify the number of pessimistic projections that are inaccurate is referred to as the false negative (FN) rate. The term "false positive" (FP) denotes the number of positive predictions that are incorrect. The true negative (TN) metric represents the cumulative count of accurate negative predictions. The following equations present the specifics of the evaluation metrics derived from the confusion matrix.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

IoU [35] is a standard for object detection that determines how much the expected and actual bounding boxes resemble one another. *k* is the IoU threshold *T*. The normalized index IoU, whose value range is [0, 1], can be explained using equation (12).

$$\text{IoU} = \frac{\text{area}(\text{box}(\text{Pred}) \cap \text{box}(\text{Truth}))}{\text{area}(\text{box}(\text{Pred}) \cup \text{box}(\text{Truth}))}. \tag{12}$$

The IoU is no longer accurate when the two bounding boxes do not overlap, or there are more ways for them to overlap when the overlapping regions are the same but the overlapping direction differs. As a result, generalized IoU (GIoU) [36] is used to evaluate the anticipated bounding box; equation (13) shows that GIoU has the same basic features as IoU and makes up for IoU's flaws.

$$\text{GIoU} = \frac{|A \cap B|}{|A \cup B|} - \frac{|C \backslash (A \cup B)|}{|C|} = \text{IoU} - \frac{|C \backslash (A \cup B)|}{|C|}, \tag{13}$$

where *A* and *B* are two bounding boxes with arbitrary forms, and *C* is the smallest rectangular box that can completely enclose them. The GIoU value range is [1, 1].

5.2 Approach results

The essential aim of this investigation is to augment security protocols and thwart acts of terrorism through the creation of a sophisticated multiweapon detection apparatus.

The presented methodology is unique compared to prior research endeavors. In addition, in this work, an encompassing dataset covering the diverse array of armaments frequently employed by terrorists in their assaults on noncombatants has been collected. Furthermore, this system utilizes an ensembled model to facilitate rapid and accurate identification and detection of various weapons, including AKM, M4, RPG, PKS, 14.5, Dragunov Sniper, Pistol, and Knife.

As indicated in

Table 2 shows that the quantity of images per class is uniform, resulting in a balanced dataset expected to yield improved outcomes.

To thoroughly analyze the performance of the MWDS object detection system, the mAP results have been categorized, as seen in Table 3, for easier comparison.

According to Table 3, the detection of different weapons classes has achieved high accuracy. For instance, AKM, Javlin, and RPG have achieved higher accuracy detection with 99.36, 99.33, and 99.22, respectively. On the other hand, the armed pose (standard and aim pose) has been achieved at 95.49 and 96.36, respectively. Figure 4 illustrates some of the detection of both weapons and poses for different video streams. In addition, the MWDS showed improvement among other state of the art, as depicted in Table 4, which represents mAP scores of different versions of YOLO and the proposed system. As noticed, the MWD showed higher mAP among other

Table 2: The total number of samples and precision, recall, and mAP

Weapon class	Number of samples	*P*	*R*	mAP
Pistol	1,500	0.941	0.967	0.967
AKM	1,500	0.992	0.949	0.958
M4	1,500	0.925	0.982	0.979
Javlin	1,500	0.99	0.989	0.995
14.5	1,500	0.983	0.997	0.987
RPG	1,500	0.995	0.973	0.985
PKS	1,500	0.988	0.913	0.951
Sniper	1,500	0.931	0.941	0.97
War-Knife	1,500	0.905	0.934	0.933
Danger_Pose	1,500	0.90	0.915	0.929
Aim_Pose	1,500	0.914	0.95	0.958

Table 3: Shows the total accuracy, *F*1 score, and sensitivity

Class	Accuracy (%)	*F*1 score	Sensitivity
Pistol	97.44	0.953	0.968
AKM	99.36	0.969	0.950
M4	95.82	0.953	0.981
Javlin	99.33	0.990	0.989
14.5	99.07	0.990	0.997
RPG	99.22	0.984	0.973
PKS	99.09	0.950	0.914
Sniper_	96.45	0.936	0.941
War-Knife	95.55	0.920	0.932
Danger _Pose	95.49	0.907	0.915
Aim_Pose	96.36	0.932	0.950

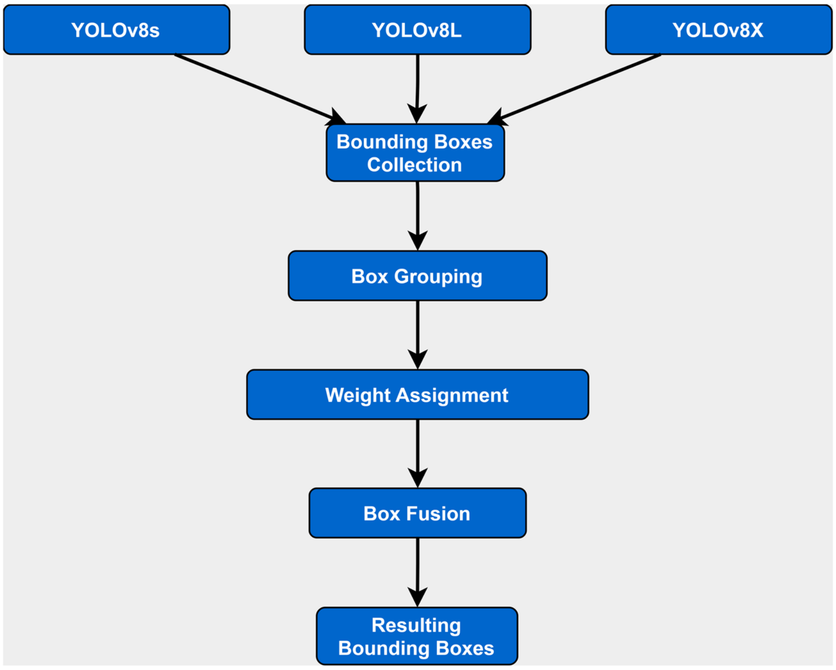

Figure 4: Weapons detected by the ensemble WBF learning.

Table 4: The total number of samples and precision, recall, and mAP

Weapon class	Yolov5	Yolov7	Yolov8	MWDS
Pistol	0.948	0.948	0.941	0.967
AKM	0.962	0.958	0.958	0.958
M4	0.958	0.956	0.976	0.979
Javlin	0.957	0.949	0.991	0.995
14.5	0.955	0.952	0.983	0.987
RPG	0.960	0.952	0.983	0.985
PKS	0.945	0.951	0.95	0.951
Sniper	0.948	0.952	0.962	0.97
War-Knife	0.949	0.945	0.927	0.933
Standard_Pose	0.905	0.90	0.921	0.929
Aim_Pose	0.953	0.941	0.955	0.958
Total	0.949	0.946	0.958	0.97

states of arts for the most class of weapons. For example, for Pistol, the Pistol obtained 0.967 compared with 0.948, 0.948, and 0.941 for Yolov5, Yolov7, and Yolov8, respectively.

This work employed a composite of training, validation, and test datasets derived from the collected data to facilitate model training and evaluation. The model was subjected to training and fine-tuning processes using the training and validation datasets during the initial phase. The model's effectiveness was assessed using an independent test dataset that the model had not previously encountered. The outcomes of this evaluation exhibit a significant degree of precision in categorization. Table 3 demonstrates the total accuracy of the detection for the proposed system.

The 150-epoch Yolov8s model began picking up new information during training in the second epoch. Due to the excellent learning via epochs offered, this learning is ideal for the study's job. Accuracy, precision, and mAP, the measures used in this study, increased rapidly with each epoch; Figure 5 shows some results from detected objects. Because of its small size, the War-Knife has the lowest accuracy (0.933%) among other weapons, which is still very high, while the Javelin, 14.5, and RPG all have the best accuracy (99.2%). All classes' combined mAP is 96%.

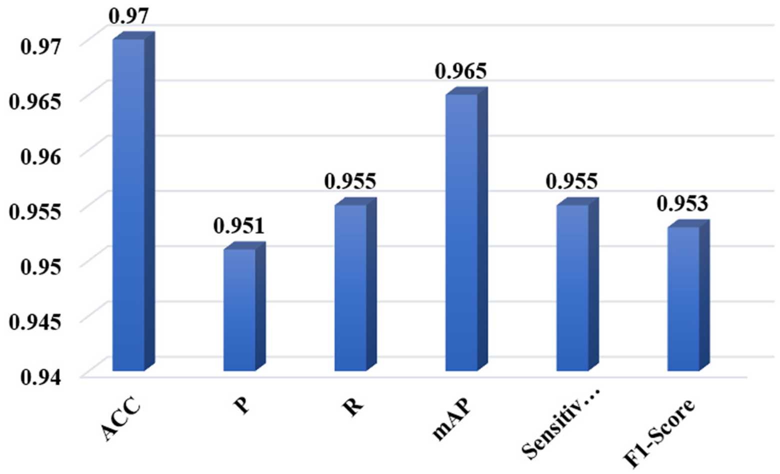

Figure 5: Accuracy, precision, recall, mAP, specificity, precision, and *F*-score for MWDS.

The weapon classes exhibited varying levels of accuracy, with the AKM demonstrating the highest degree of accuracy. Figure 6 shows the detection confidence of the classes.

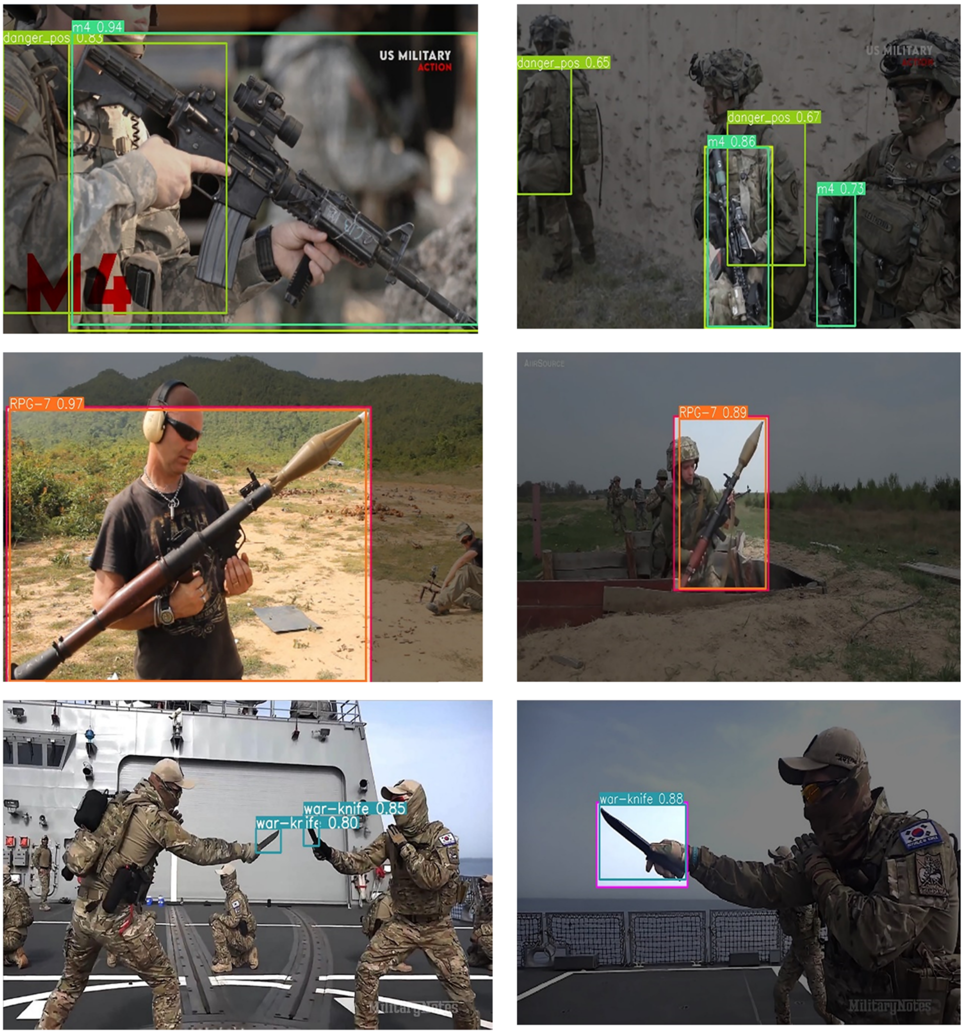

Figure 6: The detected classes with the confidence result.

6 Conclusion

In this work, a novel system that is capable of detecting different types of weapons has been proposed. Furthermore, the proposed system (namely, the MWDS) can detect whether the armed human pose is standard or dangerous with high accuracy. To achieve this, a novel deep learning algorithm, "YOLOv8," has been employed for this purpose. This was done by implementing three Yolov8 models, namely, Yolov8s, Yolov8l, and Yolov8x. To improve the robustness of detection for the proposed system, the WBF Ensemble Learning technique has been added as a subsequent phase in the proposed system. The WBF phase is responsible for amalgamating the results obtained from three Yolov8 models. As a result, this addition improved the detection accuracy to 30% when compared with the same system without the WBF assemble learning. The findings of this study show that the MWDS approach exhibits superior performance compared to other approaches across all scenarios, with an average detection accuracy rate of 97%.

7 Future work

Therefore, there are many directions that can be suggested in the future work. The following are some of these directions.

1. Expansion to Other Object Classes: Expanding the system's detection capabilities to include other object classes beyond weapons is valuable. Incorporating additional classes, such as suspicious objects or specific threats, can enhance system's versatility and applicability in diverse security scenarios.
2. Integration of contextual Information: Considering contextual information, such as scene analysis and behavior modeling, can further enhance the system's performance. Incorporating contextual cues and prior knowledge can improve the accuracy of weapon detection and provide more comprehensive situational awareness.

Acknowledgements: The authors would like to thank the Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq. All images included in this paper are originally produced by the authors.

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

Author contributions: Conceptualization, Moahaimen Talib; methodology, Moahaimen Talib; software, Moahaimen Talib; validation, Moahaimen Talib Abdullah, Ahmed Al-Noori; formal analysis, Jameelah Suad; investigation, Moahaimen Talib Abdullah, Ahmed Al-Noori; resources, Moahaimen Talib; data curation, Jameelah Suad; writing—original draft preparation, Jameelah Suad; writing—review and editing, Ahmed Al-Noori, Jameelah Suad; visualization, Moahaimen Talib Abdullah; supervision, Jameelah Suad; project administration, Jameelah Suad; funding acquisition, Moahaimen Talib Abdullah, Ahmed Al-Noori.

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this article.

Data Availability statement: The datasets generated and analyzed during the current study are available in the Roboflow Universe [https://app.roboflow.com/weapondetection-pzoyk/weapon-detection_merged/deploy] and publicly allowed.

References

[1] Szeliski R. Computer vision: Algorithms and applications; 2010. [Online]. http://szeliski.org/Book/.

[2] Dollár P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell. 2012;34(4):743–61. doi: 10.1109/TPAMI.2011.155.

[3] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. Int J Robot Res. 2013 Sep;32(11):1231–7. doi: 10.1177/0278364913491297.

[4] Hussein M, Abbas AH. Plant leaf disease detection using support vector machine. Al-Mustansiriyah J Sci. 2019 Aug;30(1):105–10. doi: 10.23851/mjs.v30i1.487.

[5] Fakhir Mutar A. Study fire detection based on color spaces. Al-Mustansiriyah J Sci. 2019 May;29(4):93–9. doi: 10.23851/mjs.v29i4.414.

[6] Abood Ramadhan Z, Alzubaydi D. Text detection in natural image by connected component labeling. Al-Mustansiriyah J Sci. 2019 Aug;30(1):111–8. doi: 10.23851/mjs.v30i1.531.

[7] Archana KV, Komarasamy G. A novel deep learning-based brain tumor detection using the Bagging ensemble with K-nearest neighbor. J Intell Syst. 2023 Jan;32(1):20220206. doi: 10.1515/jisys-2022-0206.

[8] Guo J. Deep learning approach to text analysis for human emotion detection from big data. J Intell Syst. 2022 Jan;31(1):113–26. doi: 10.1515/jisys-2022-0001.

[9] Chaturvedi RP, Ghose U. A review of small object and movement detection based loss function and optimized technique. J Intell Syst. 2023 Jan;32(1):20220324. doi: 10.1515/jisys-2022-0324.

[10] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2016, [Online]. http://pjreddie.com/yolo/.

[11] Abbas ZK, Al-Ani AA. Anomaly detection in surveillance videos based on H265 and deep learning. Int J Adv Technol Eng Explor. 2022;9(92):910–22. doi: 10.19101/IJATEE.2021.875907.

[12] C. Asia-Pacific Signal and Information Processing Association. Annual Summit and Conference; 2019. Lanzhou Shi, Asia-Pacific Signal and Information Processing Association, IEEE Signal Processing Society, and Institute of Electrical and Electronics Engineers, 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

[13] Kaya V, Tuncer S, Baran A. Detection and classification of different weapon types using deep learning. Appl Sci (Switz). 2021 Aug;11(16):7535. doi: 10.3390/app11167535.

[14] Bengio Y. Learning deep architectures for AI. Found Trends Mach Learn. 2009;2(1):1–27. doi: 10.1561/2200000006.

[15] Song HA, Kim BK, Xuan TL, Lee SY. Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task. Neurocomputing. 2015 Oct;165:63–74. doi: 10.1016/j.neucom.2014.08.095.

[16] Masood S, Ahsan U, Munawwar F, Rizvi DR, Ahmed M. Scene recognition from image using convolutional neural network. Procedia Comput Sci. 2020;167:1005–12. Elsevier B.V. doi: 10.1016/j.procs.2020.03.400.

[17] Jiang L, Liu H, Zhu H, Zhang G. Improved YOLO v5 with balanced feature pyramid and attention module for traffic sign detection. MATEC Web Conf. 2022;355:03023. doi: 10.1051/matecconf/202235503023.

[18] Osokin D. Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose. 2018 Nov, [Online]. http://arxiv.org/abs/1811.12004.

[19] Olmos R, Tabik S, Herrera F. Automatic handgun detection alarm in videos using deep learning. Neurocomputing. 2018 Jan;275:66–72. doi: 10.1016/j.neucom.2017.05.012.

[20] Fernandez-Carrobles MM, Deniz O, Maroto F. Gun and knife detection based on Faster R-CNN for video surveillance. 2019, [Online]. http://visilab.etsii.uclm.es.

[21] Verma GK, Dhillon A. A handheld gun detection using faster R-CNN deep learning. In ACM International Conference Proceeding Series. Association for Computing Machinery; 2017 Nov. p. 84–8. doi: 10.1145/3154979.3154988.

[22] Gelana F, Yadav A. Firearm detection from surveillance cameras using image processing and machine learning techniques. In Advances in intelligent systems and computing. Singapore: Springer Verlag; 2019. p. 25–34. doi: 10.1007/978-981-13-2414-7_3.

[23] Dwivedi N, Singh DK, Kushwaha DS. Employing data generation for visual weapon identification using convolutional neural networks. Multimed Syst. 2022;28(1):347–60.

[24] Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics (Version 8.0.0) [Computer software]; 2023. https://github.com/ultralytics/ultralytics.

[25] ACG, Krishnan K, Angel Viji KS. Multiple Object Tracking using Deep Learning with YOLO V5. 2021, [Online]. www.ijert.org.

[26] de Azevedo Kanehisa RF, de Almeida Neto A. Firearm detection using convolutional neural networks. ICAART. 2019;2(2):707–14.

[27] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. 2018 Mar, [Online]. http://arxiv.org/abs/1803.01534.

[28] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022 Jul, [Online]. http://arxiv.org/abs/2207.02696.

[29] Couturier R, Noura HN, Salman O, Sider A. A deep learning object detection method for an efficient clusters initialization. arXiv preprint arXiv:2104.13634; 2021.

[30] Neubeck A, ETH Zurich, Van Gool L. ETH Zurich. Efficient Non-Maximum Suppression; 2006.

[31] Yan B, Fan P, Lei X, Liu Z, Yang F. A real-time apple targets detection method for picking robot based on improved YOLOv5. Remote Sens (Basel). 2021 May;13(9):1619. doi: 10.3390/rs13091619.

[32] Song Q, Li S, Bai Q, Yang J, Zhang X, Li Z, et al. Object detection method for grasping robot based on improved YOLOv5. Micromachines (Basel). 2021;12(11):1273.

[33] Institute of Electrical and Electronics Engineers. 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW): Hong Kong, 10–14 July 2017.

[34] Zhou H, Li Z, Ning C, Tang J. CAD: Scale invariant framework for real-time object detection. Venice, Italy: IEEE; 2017.

[35] Yu J, Jiang Y, Wang Z, Cao Z, Huang T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia; 2016. p. 516–20.

[36] Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society; 2019 Jun. p. 658–66. doi: 10.1109/CVPR.2019.00075.