Yolov7 - human detection for Enemy Identification System

Zhen-Gyi Jiang, Lei Lin, Yu-Zhe Liu, Nai-Shen Hu, Cheng-Shun Lee

Department of Computer Information Science, Republic of China Military Academy, Taiwan

ABSTRACT

The main purpose of this paper is to identify adversaries who are difficult to discern with the naked eye on the battlefield. The more detailed information can be displayed, such as distance or the number of enemies. As gaining an edge in modern warfare is increasingly challenging, our goal is to develop recognition technology that can provide an early advantage, facilitating a quick understanding of battlefield dynamics. The content mainly focuses on non-contact recognition and integrates YOLOv7 and camera calibration technologies. The system utilizes computer vision technology to identify personnel attire, distinguishing friendly forces in green and enemy forces in red, which enhances the recognition process. Furthermore, the system also provides distance recognition for targets and cameras to enhance the observer's real-time understanding of the battlefield.

Key words：Enemy, Friendlies, YOLOv7, Detection, Fast R-CNN

1. Introduction

Reference to the ongoing war in Ukraine and tactical training in military academies, we can understand that distinguishing between friend and foe on the battlefield is not an easy task. The diverse conditions on the battlefield make identifying the enemy a significant challenge. Therefore, we believe that implementing a system to assist in identifying friend or foe would bring numerous benefits to the battlefield. This system may bring benefits in the future.

1. The system has a clear understanding of enemy information.
2. Integrating technology into glasses to reduce the likelihood of friendly fire incidents among soldiers.
3. Integration with drones can enhance reconnaissance accuracy, reduce friendly fire incidents, and improve the hit rate for attack drones.

Biometric recognition systems can be broadly categorized into two types based on the method of collecting biological feature data: contact-based and non-contact-based. Contact-based systems are less accepted due to concerns about privacy infringement and hygiene. Non-contact-based systems, on the other hand, are more widely accepted because they are perceived as less invasive. Generally, contact-based biometric recognition systems are considered more secure than non-contact-based systems. Contact-based systems can be further divided into three types: fingerprint recognition, which is the most widely used biometric technology, vein recognition, and signature recognition. Each has its own advantages and disadvantages.

Overall, the choice of a biometric system depends on factors such as security, acceptance, and practicality. In order to achieve the aforementioned results, we decided to develop our own system and chose to use contactless human body recognition as the foundation basis.

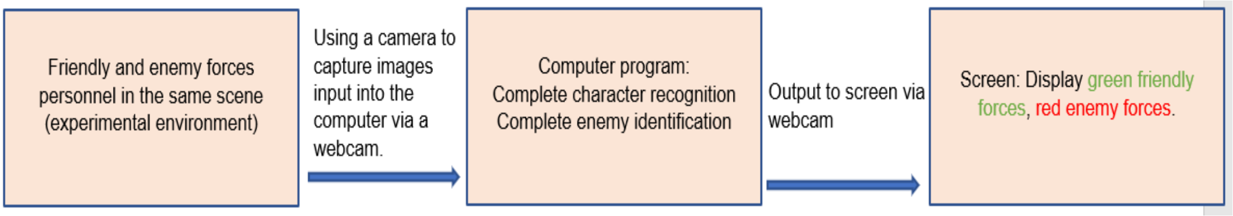

Friendly and enemy forces personnel in the same scene (experimental environment)
Using a camera to capture images input into the computer via a webcam.
Computer program: Complete character recognition Complete enemy identification
Output to screen via webcam
Screen: Display green friendly forces, red enemy forces.

Figure1: The framework captures images with a camera and outputs the results to the screen through a program, distinguishing between green friendly forces and red enemy forces

Then, we will train the model to identify individuals within a specific range to achieve the desired identification effect, with the aim of reaching our goal. By leveraging advanced computer vision technology, specifically deep learning, we will train a model to distinguish between friendly personnel and enemy personnel. The primary identification criterion will be clothing. Friendly units will be marked in green, while enemy units will be marked in red, making identification more intuitive.

We will not only identify friends and foes, but also provide distance and location information of the target through the camera. This will involve integrating ranging technology, image processing, and calculating the target's distance to enhance the observer's real-time understanding of the battlefield.

2.Related Work

2.1 Fast R-CNN [1]

R-CNN Region-based Convolutional Neural Network is a deep learning model used for object detection. It stands as a significant milestone in the field of object detection. It encompasses several advantages or characteristics, such as feature extraction and region classification. However, despite its good performance in object detection tasks, R-CNN [2] has a high computational cost. Subsequent improved models, such as Fast R-CNN, Faster R-CNN [3], and Mask R-CNN [4], have gradually evolved, optimizing for speed and efficiency. These subsequent models further enhance the effectiveness and speed of object detection.

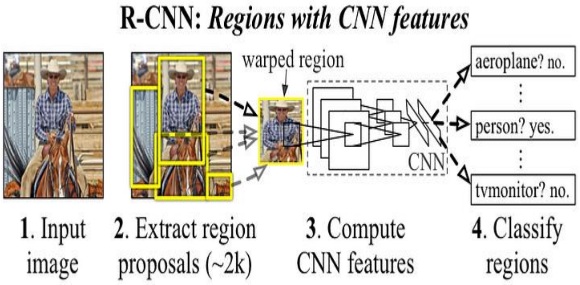


Figure2: It applies the region proposal method to send the regions of these proposals into a CNN to extract features, providing input for the subsequent CNN network

2.1.1 Region proposal [5]

Region Proposal is a crucial step in object detection, especially in systems utilizing Convolutional Neural Networks (CNNs) for this task. The primary objective of Region Proposal is to identify potential regions in an image that likely contain objects. By focusing computational efforts on these regions, subsequent steps like feature extraction and classification can be more efficient. The Region Proposal step helps narrow down the search space for object detection, making subsequent processing more efficient. After generating region proposals, the system can then focus on extracting features and classifying these regions to determine the presence of objects and their categories

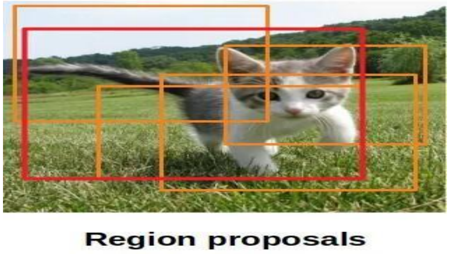


Figure3: First find the "areas that are more likely to be cats" on the image (such as the orange box in the figure), and then evaluate these areas

2.2 YOLOV7

YOLOv7 [7] has improved speed and accuracy by introducing several architectural reshaping, similar to Scaled YOLOv4, as it is written by the same author. It incorporates recent advancements in CNN neural networks, especially in reducing parameter quantity and enhancing computational efficiency, particularly in edge devices.

In addition to proposing new dynamic label assignment strategies, enhancing feature learning capabilities through hierarchical deep supervision and dynamic label assignment, it is also possible to effectively utilize parameter and memory usage expansion, as well as composite scaling methods. This is done without disrupting the original gradient paths, continuously strengthening the ability of continual learning. This approach can effectively reduce the parameters and computational load of state-of-the-art real-time object detection models by 40%, while achieving faster inference speed and higher detection accuracy by 50%.

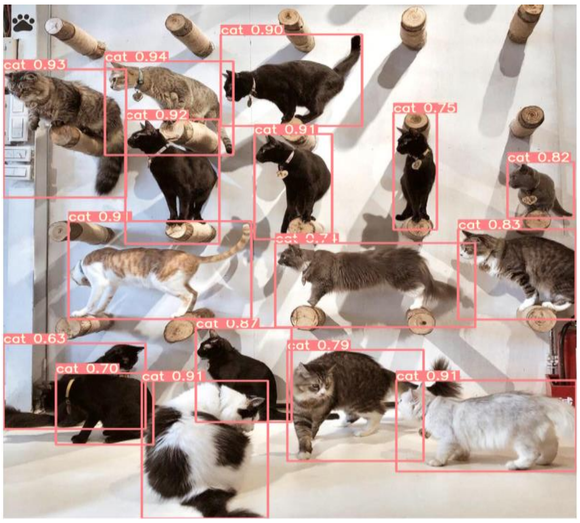


Figure 4: The performance of YOLOv7 in object detection models

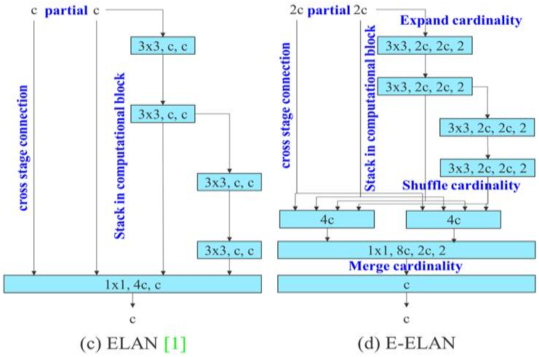


Figure 5: Picture (c), It allows deeper networks to learn and converge more effectively by controlling the shortest and longest gradient paths. Since stacking more modules may disrupt this stable state, YOLOv7 proposes extend ELAN (E-ELAN)
Picture (d), It achieves the effect of continuously enhancing network learning ability through techniques such as expand cardinality, shuffle cardinality, merge cardinality, etc., without disrupting the gradient path state

YOLOv7 can perform three tasks. One is Object Detection, which identifies objects in the scene and outlines them with bounding boxes. Additionally, it displays the predicted object names and probabilities on these boxes. One is Instance Segmentation, which involves coloring the recognized objects. This action is also referred to as Masking, and it provides a more detailed representation compared to object detection, where objects are outlined with bounding boxes only.

Another task is Key point Estimation. For humans, the model can identify and connect joint points in the scene, creating a representation similar to a stick figure. This can be used to predict actions and postures, such as fall detection.

3. Experimental

3.1 Data Collection

Initially, we collected approximately 1000 images as samples from the internet, including photos of individuals wearing camouflage uniforms from the Republic of China, the United States, and the People's Republic of China. Some of these photos also featured individuals in casual attire. Subsequently, we used bounding boxes to initiate the training and learning process for our equipment. This enables us to ultimately achieve the capability to distinguish between individuals wearing camouflage attire (marked with a green box) and those in casual clothing (marked with a red box) when captured by a camera as shown in figures 6 to 8.

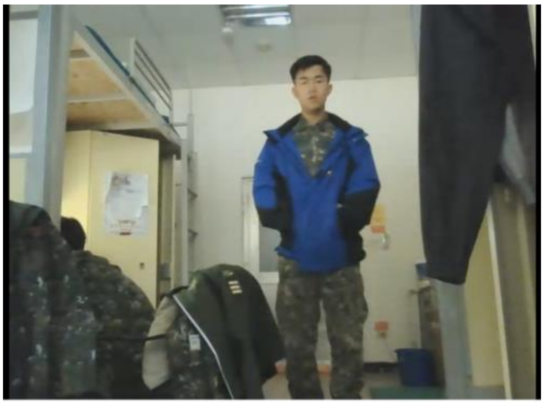


Figure 6: Testing individuals in casual clothing (not displayed due to limited training samples for casual attire)

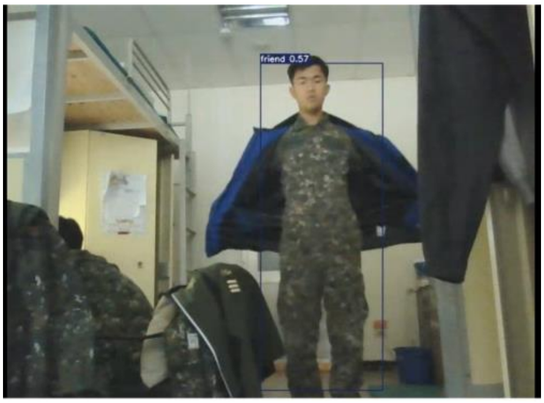


Figure9: Testing individuals wearing casual outerwear but revealing camouflage attire

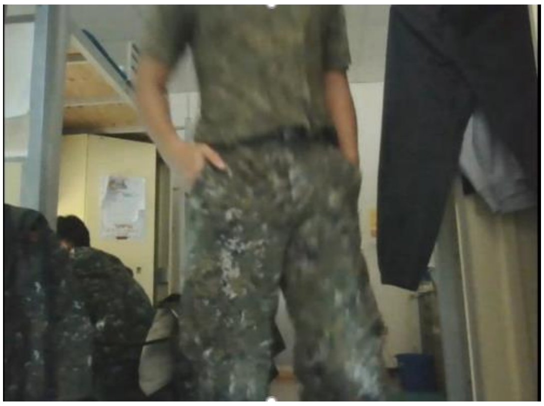


Figure 7: In close-range testing of individuals wearing camouflage attire, issues have been identified with the inability to capture certain elements

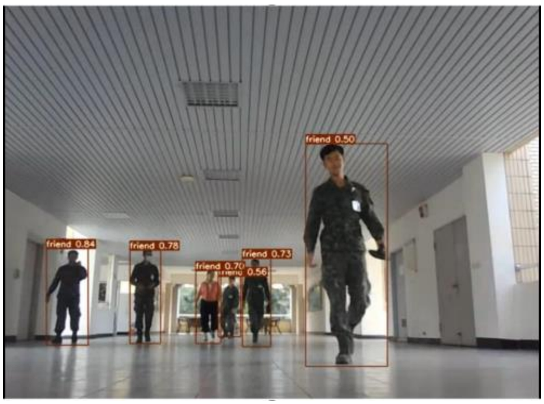


Figure 8: During long-range testing of individuals clad in camouflage attire, recognition remains viable

When we train, batch is set to 5 (the memory will be overloaded at higher levels) epochs are set to 5000 times to do the operation with yolov7 and GPU RTX4050, which takes a total of 3 hours Compared to the training data on the network, our database needs to be expanded and filtered, and then the image data will be removed from the individual cases. The purpose is to improve the accuracy of our machine. Additionally, we will strive to avoid blurry photos during machine testing to prevent training issues, as well as situations where there are too many personnel in the photos, leading to accuracy problems for the machine.

The box loss (bounding box regression loss) in yolov7 is computed using the Generalized Intersection over Union (GIOU) loss function for bounding boxes. The objective is to minimize this loss, ensuring that predicted bounding boxes are as accurate as possible. The GIOU loss considers the mean of box predictions, where smaller values indicate more accurate bounding boxes.

In our experiments, we have observed some differences in the box loss curve compared to data found online. While typical curves show a rapid decrease followed by a gradual stabilization, our loss function exhibits significant fluctuations in the middle section, indicating instability.

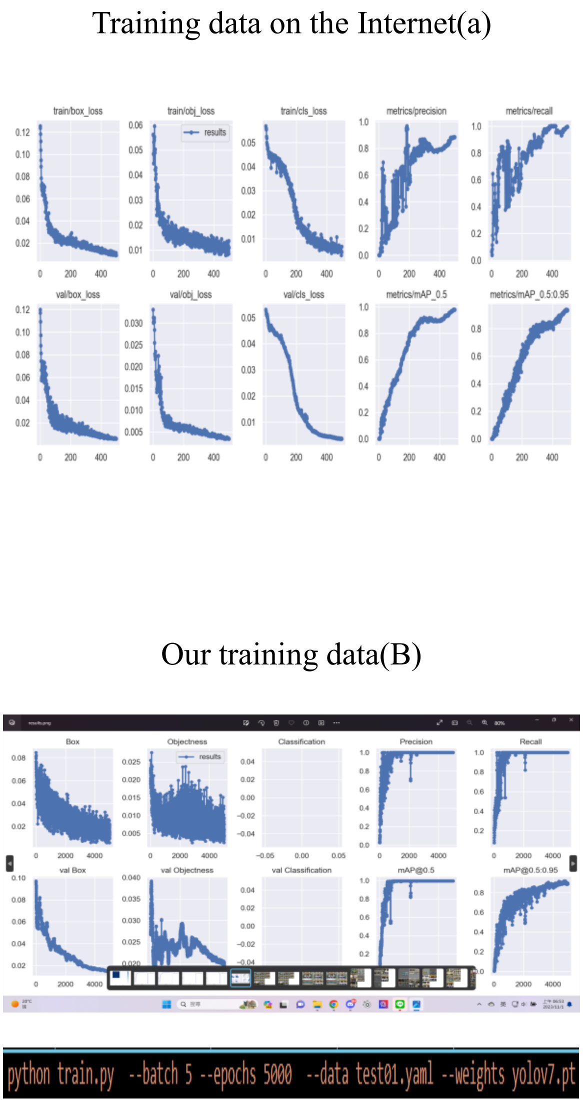
The obj loss［8］is used to supervise whether an object exists in a grid and calculates the network's confidence, considering both coordinate values and the Intersection over Union (IoU) area with the target object. Similar to box loss, our experimental data displays fluctuations in confidence values in the middle section.

Cls loss［9］is employed for classifying the target's category, ensuring accurate classification of objects such as cats and dogs. In our experiments, we lack cls loss data due to the limited number of categories in our labeling, in contrast to online data that uses around 400 categories for analysis.

Validation set loss (val Box_loss and val obj loss) data in our experiments exhibit some similarities with online data, particularly in the box loss segment. However, obj still shows some fluctuations in the middle section.

Metrics are indicators used to measure the performance of a developed model. Precision is calculated as TP/(TP+FP), representing the probability of correctly predicting True when the actual value is True. Our precision data consistently shows high values, indicating a high rate of correct predictions, which is expected given the predominant use of camouflage images in our training dataset.

MAP0.5:0.95 represents the average mAP across different IoU thresholds (from 0.5 to 0.95). The mAP is the area under the Precision-Recall curve, and mAP0.5 signifies the average mAP when the IoU threshold is greater than 0.5. Our recall data exhibits a gradual trend, and the fluctuation in mAP0.5:0.95 suggests variability in the overall model accuracy.



Training data on the Internet(a)

Our training data(B)

```
python train.py --batch 5 --epochs 5000 --data test01.yaml --weights yolov7.pt
```

Figure12: The comparison between the (a) online data and (b) our training data graphs

In conclusion, based on our testing results, we have identified several areas that require improvement. These include issues such as difficulty in discernment when the target is too close, occasional misjudgments for specific individuals or objects, and errors in identifying individuals wearing casual clothing. The root cause of these challenges lies primarily in the insufficient training of our machine, specifically the limited quantity of samples used. This inadequacy results in less-than-ideal performance during the current testing phase, preventing us from achieving the anticipated outcome of accurately distinguishing between friendly forces (wearing camouflage attire) and enemy forces (wearing casual clothing) in captured images.

4 Conclusions

Firstly, we encountered a challenge that sampling pictures of the national army from the internet might not reach the scheduled 10,000 images, hence the need to mix pictures of expected friendly forces.

Next, due to insufficient SSD and RAM capacity in the hardware devices, the batch size during model training can only be limited to 5. The storage space required for training the model is substantial. Currently, the device can only save two types of epochs to reach more than 5000 models. Due to capacity issues, various epochs cannot be stored to screen the best ones. Video recognition had to be temporarily suspended due to capacity problems, with future plans for hardware upgrades. Additionally, by upgrading the memory of the laptop, an additional team member's device was utilized to accelerate the training process. Furthermore, due to challenges in clothing recognition, casual clothing recognition was added, and casual clothing was marked as red to indicate enemy forces.

Then, the training analysis graph clearly indicates that several images may affect the execution results, as the current training set consists of approximately 800 images. Training volume will be increased to address this issue, and if there's an opportunity to upgrade hardware devices, better training outcomes can be achieved.

References

[1] Ivan “[Object Detection] S2: Introduction to Fast R-CNN” medium Aug 30, 2019.

[2] Lung-Ying Ling "R-CNN Learning Notes" Mar 23, 2019.

[3] Ivan “[Object Detection] S3: Introduction to Faster R-CNN” medium Sep 2, 2019.

[4] Ivan “[Object Detection] S9: Introduction to Mask R-CNN” medium Nov 2, 2019.

[5] Chinghai Ching Cheng “R-CNN Region Proposals” medium Aug 31, 2020.

[6] Chingi Lee “Overview of Object Detection” medium Jan 3, 2021.

[7] Jiaming Chang "The latest champion of object detection, YOLOv7"aiacademy Jul 14,2022.

[8] Tommy Huang"Machine/Deep Learning Fundamentals: Introduction to Loss Functions’ medium 27, 2018.

[9] Shai “What is loss_cls and loss_bbox and why are they always zero in training” stack overflow Nov,2019.

Yolov7人類檢測用敵我辨識系統

江政誼 胡乃申 劉宇哲 林磊 李政勳

陸軍軍官學校資訊系

摘要

本文旨在辨識在戰場上難以肉眼分辨的對手。更詳細的信息可被顯示，例如距離或敵人數量。隨著在現代戰爭中獲得優勢變得越來越具挑戰性，我們的目標是創造一種能提供早期優勢的識別技術，以便快速理解戰場動態。內容主要集中在非接觸式識別，並整合了 Fast R-CNN、YOLOv7和攝像機校準技術。系統利用計算機視覺技術辨識人員服裝，將友軍以綠色、敵軍以紅色區分，從而增強了辨識過程。此外，系統還提供了目標和攝像機的距離識別，以改善觀察者對戰場的實際理解。

關鍵詞：敵人，友軍，YOLOv7，檢測，Fast R-CNN