



From Pixels To Presence: A YOLO-Based Deep Learning Model For Human Detection

¹Dr M. Sindhana Devi, ²Dr. R K Kavitha, ³Devadarshini P, ⁴Midhunesh G

¹Assistant Professor, ²Assistant Professor, ³Students, ⁴Students

Department of Data science,

Kumaraguru college of Liberal Arts and Science, Coimbatore, India

Abstract

Object detection plays a vital role in modern computer vision applications by enabling automated understanding of visual scenes. This paper presents an efficient deep learning-based human detection system using a YOLO-based object detection architecture. The proposed model is trained on a large-scale, single class annotated dataset to ensure robustness against variations in scale, position, and environmental conditions. Data preprocessing and augmentation techniques are employed to improve generalization and reduce overfitting. The model's performance is evaluated using standard metrics including Precision, Recall, mAP@0.5, and mAP@0.5–0.95. Experimental results demonstrate steady loss convergence and consistent improvement in detection accuracy across training epochs, indicating reliable learning behavior. The framework achieves accurate localization of human instances while maintaining computational efficiency, making it suitable for real-time applications such as surveillance, crowd monitoring, and intelligent vision systems. The proposed approach provides a scalable foundation for future extensions to multi-class detection and deployment on edge-based platforms.

Keywords— Object Detection, YOLO, Deep Learning, Computer Vision, Convolutional Neural Networks (CNN), PyTorch, Model Training, Precision and Recall and MAP Evaluation.

INTRODUCTION

Object detection has recently become a basic task in computer vision, allowing computers to automatically detect and locate objects of interest in images and videos. As a result of the fast development of intelligent surveillance systems, smart cities, and autonomous monitoring, accurate human detection has become a crucial requirement for real-world applications. In the past, computer vision methods were dominated by handcrafted features and heuristic-based approaches, which were not robust to complex lighting conditions, occlusions, and scale and pose variations.

However, recent breakthroughs in deep learning, especially in convolutional neural networks (CNNs), have led to improvements in the performance of object detection tasks. Among these, single-stage object detection models like the You Only Look Once (YOLO) series have gained popularity because of their capability to detect objects in real-time with high accuracy. YOLO-based models treat object detection as a regression task. Although considerable progress is already achieved in recognizing people in a group effectively, the task is still challenging because of many people in a group, a more complex background, and a varying appearance

of people. Yet, the issue is to balance how accurate recognition is and how fast it is to make the system efficient. As a motivation to deal with these hurdles in a YOLO-based system for efficient detection, the presented paper aims to fulfill the task.

RELATED WORK

Object Detection is an active research field in the field of computer vision, which started from the trick-based feature-based approaches to the end-to-end-based "deep learning." Early approaches such as Haar cascades, HOGs, and Deformable Part-Based Models are typically based on hand-crafted features and the "sliding window." They are known to work well in a highly controlled setting, but they are not easy to scale or to maintain their robustness when the scene changes its lighting, or when there are a lot of unpredictable "backgrounds".

Later came the power of deep convolutional networks. Two-stage object detection in models such as R-CNN, Fast R-CNN, and Faster R-CNN improved accuracy by dividing process and prediction in two steps: object proposal and subsequent classification of these objects. Even though this increased computation held us back from making object detection real-time ready, a new class of object detection models in models such as SSD and YOLO shed a new light on object detection as a problem of object regression.

The YOLO series of detection models have developed through multiple versions, with further improvements to the extraction of features, the use of anchors, and the development of loss functions. Various investigations have proven the strength of the YOLO-based models for human detection with the aid of real-world applications such as surveillance, crowd detection, and pedestrian detection.

However, there are many areas that are yet to be addressed, such as dealing effectively with unbalanced classes, changes in the number of objects, as well as handling densely crowded scenes that contain many objects. Although a significant number of researchers have focused on multi-class detection, there has been less emphasis on single-class human detection, which remains a significant aspect in human detection in practice. Following previous efforts like YOLO, this research desires to effectively resolve the detection issue, which faces the limitations in efficiency and accuracy, as well as computing

SYSTEM ARCHITECTURE

The proposed system architecture is intended to be efficient for human detection tasks using a deep learning-based object detection system. The system architecture is divided into a modular pipeline that includes data preparation, training, inference, and evaluation. The system architecture also includes a YOLO-based single-stage detector that acts as the main component of the system, which supports real-time object detection.

3.1 Overall Workflow

The system starts with a large-scale annotated image dataset with human instances annotated using bounding boxes. The annotations are in a normalized format that is compatible with the YOLO format. Before the training process, the dataset is preprocessed to include resizing, normalization, and data augmentation. The preprocessed dataset is then split into a training and a validation set.

During training, the YOLO approach trains the network to perform tasks that involve coordinate estimation of bounding boxes, confidence in an object, and estimation of class. It specifically targets the problem of detecting humans in an image and reduces ambiguity by making the network concentrate more on the accuracy of the boxes while minimizing class ambiguity. The YOLO loss combines bounding box regression loss, classification loss, and distribution focal loss.

3.2 Model Architecture

The YOLO-based method can be divided into three core components: a backbone module for feature abstraction, a neck module focused on feature expansion, and a head module for the final predictions. The backbone module filters out the hierarchical features from the images, and the features are broadened and

sharpened in the neck module to deal with the scaling problem. The head module predicts the bounding boxes with different scales to ensure the reliable detection of humans with different sizes.

3.3 Inference and Evaluation Pipeline

In inference, the trained model takes in new images and produces bounding box predictions in one pass. Non-Maximum Suppression (NMS) is used to remove duplicate detections. The performance of the system is measured using standard metrics like Precision, Recall, mAP@0.5, and mAP@0.5-0.95. The architecture provides robust training, efficient inference, and scalability for real-time human detection.

DATASET DESCRIPTION AND PRE-PROCESSING

4.1 Dataset Description

The dataset employed in this research work is a large-scale dataset of images with human instances annotated using bounding boxes. The dataset is intended for single-class object detection, where all the annotated instances are of the person class. The dataset comprises images taken in different environmental conditions, lighting, camera perspectives, and human poses and scales. This ensures that the model is well generalized when applied in real-world applications.

Every image in the dataset comes with annotation files in YOLO format, in which the bounding boxes are described by the normalized coordinates of the center point, width, and height. The dataset has almost a million instances, which is sufficient for learning robust spatial and contextual features. The class distribution is uniform since the dataset only focuses on human detection, thus not having the class imbalance problem that is common in multi-class detection problems.

The data is divided into training and validation sets to facilitate an unbiased assessment of the performance. The training set is utilized for the optimization of the model parameters, while the validation set is used to control overfitting.

4.2 Data Pre-processing

To facilitate a uniform representation of the input, all the images are resized to a fixed resolution that is compatible with the YOLO model. The pixel values are normalized to enhance numerical stability during training. Various data augmentation methods like random scaling, horizontal flipping, translation, and mosaic augmentation are used to improve the robustness of the model with respect to size, location, and orientation of the object.

These preprocessing and augmentation techniques are helpful in improving the generalization performance, preventing overfitting, and allowing the model to learn discriminative features for human localization. The combined dataset preparation pipeline is a crucial component in ensuring stable training convergence and detection performance.

MODEL ARCHITECTURE AND TRAINING DETAILS

5.1 Model Architecture

The human detection method relies on a YOLO-inspired object detection approach. The object detection approach is selected for its ability to yield competitive accuracy without compromising real-time speed. The object detection approach consists of a conventional three-part configuration: backbone, neck, and detection heads. The backbone of our object detection approach operates on our input images using convolutional layers to extract hierarchical spatial features of our images that are informative for detecting human objects. Here, the combination of information across various scales is facilitated through feature fusion, allowing the model to recognize individuals of various sizes and shapes due to its robustness in tricky scenes where objects are either distant, partially hidden, or where many objects are present in one scene. After this is accomplished, the detection head processes this information and produces bounding box coordinates, object confidence scores, and class probability simultaneously. As our objects only consist of one class, our output layer is more specific to the person class, allowing us to have high precision while retaining simplicity in classification.

5.2 Training Configuration

We train the model using supervised learning on the cleaned dataset, optimizing a combined loss regarding how good the placing of boxes is, how well the objects are classified, and how distributional predictions are handled. This setting helps the model precisely localize objects and reliably converge. We adopt stochastic gradient descent with carefully chosen hyperparameters-learning rate, batch size, and number of epochs-to strike a good balance between fast convergence and good generalization.

The training cuts across several epochs, where both training and validation losses are traced. We also periodically check on metrics such as Precision, Recall, and Average Precision for progress. A consistent drop in loss, complemented by consistent gains in such metrics, indicates stable learning and correct model development.

EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The proposed YOLO-based human detection system was evaluated on both training and validation metrics with the aim of measuring how learning is progressive, at what rate it converges, and how accurate the detection of people is. Evaluation is based on the standard object detection measures-loss components, Precision, Recall, mAP@0.5, and mAP@0.5–0.95.

6.1 Training and Validation Loss Analysis

Figure X shows training and validation loss as a function of epochs. During training, the trend seems to be that of going downwards for all three losses: bounding box regression loss, classification loss, and distribution focal loss, which in turn indicates that there has been effective optimization and stable learning. Validation losses generally take on a similar trend of sliding downhill, tattling well about good generalization into data not seen before with little sign of overfitting. The close alignment of the training and validation curves reinforces the robustness of the training process.

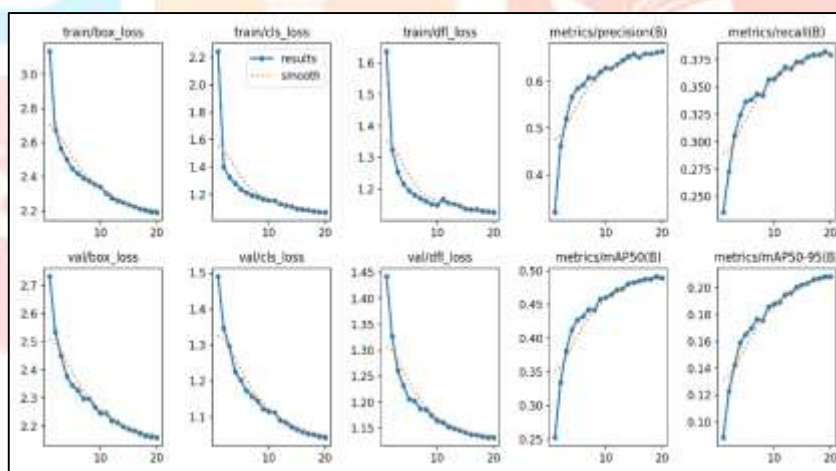


Figure 1: Training and validation loss curves with precision, recall, and mAP across epochs.

6.2 Detection Performance Metrics

As training progresses, the detection performance increases gradually. Precision and Recall increase steadily with the epochs since this model develops its powers of correctly identifying human instances while simultaneously reducing false positives and false negatives. Precision stabilizes at a higher mark, reflecting dependable confidence in the predictions, while Recall increases to encompass more of the ground-truth human instances.

Mean Average Precision, or mAP, is the major metric here. The model steadily improves mAP@0.5, which signals already quite reliable localization in the case of moderate overlap. At the same time, the mAP@0.5–0.95 grows consistently-to say nothing of its growth-already showing the model's robustness under tighter localization demands.

6.3 Result Interpretation

Taken together, the experiments demonstrate that the proposed framework converges smoothly and maintains a balanced precision–recall trade-off while delivering dependable detection accuracy. These results support the model’s suitability for real-time human detection in practical vision-based systems.

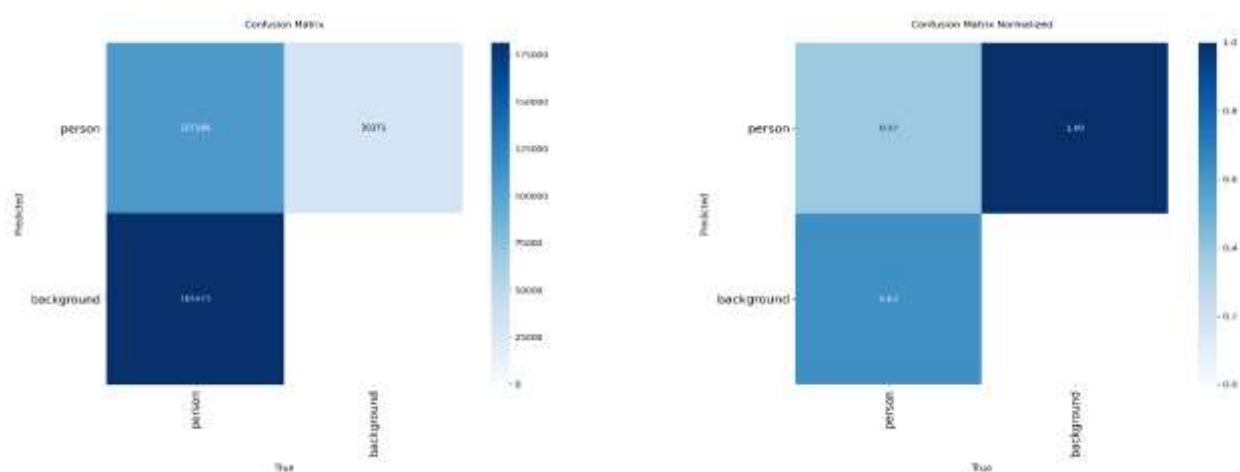


Figure 2: (a) Confusion matrix and (b) normalized confusion matrix of the proposed model.

DISCUSSION

All these curves indicate that our YOLO-based human detector is learning steadily. Three of these (loss metrics) are trending downwards with each iteration of epochs: this is a sure sign of optimization and convergence. The near-identical trend of training and validation losses suggests we are getting generalization with minimal chances of overfitting.

The rise in Precision and Recall will eventually allow the model to detect more humans while sending fewer false warnings or alarms. With increased Precision, it will become more confident in its predictions or imagery it portrays to users, while increased Recall means it will identify more of the actual objects or entities which it has been trained to detect in various contexts or scenes.

We also witness the steady increase in mAP@0.5, validating the stable localizations at the standard overlap conditions. We also observe an improvement in mAP@0.5–0.95, verifying the consistent performance of the model at a more stringent evaluation criterion. This one appears to be relatively small and leaves room for tuning. In the end, we confirm the consistency and efficacy of the approach for the purpose of human detection.

CONCLUSION AND FUTURE WORK

It proposes a lightweight deep learning framework for detecting humans with YOLO. It accentuates precise localization and stable training behavior for dependable detection performance. Adopting the single-class detection setup supported by a large, richly annotated dataset, the system demonstrates steady convergence and a balanced precision-recall profile. The experiments reveal a consistent drop in both training and validation losses and ongoing gains in mAP, underlining the effectiveness of the model for real-time human detection tasks.

The framework yields a good trade-off between estimation accuracy and computational efficiency, making it practical for vision-based monitoring applications in real-world settings such as surveillance, crowd analysis, and smart environments.

Looking to the future, we want to expand the model into multi-class detection, enhance performance in dense or occlusion-heavy scenarios, and increase inference speed for edge and resource-constrained devices. We further expect improvements by advanced data augmentation, hyperparameter optimization, and testing on real-world video streams to further validate the robustness and scalability aspects.

REFERENCES

- [1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
- [2] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271, 2017.
- [3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” *arXiv preprint arXiv:2004.10934*, 2020.
- [4] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” *arXiv preprint arXiv:2207.02696*, 2022.
- [5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

