Enhancing real human detection and people counting using YOLOv8

Tahreer Abdul Ridha Shyaa[1*] and *Ahmed* A. Hashim[2]

[1] The Iraqi Commission for Computers and Informatics, The Informatics Institute for Postgraduate Studies, The department of Computer Science Baghdad, Iraq
[2] University of Information Technology and Communications, College of Engineering Baghdad, Iraq

Abstract. The ability to accurately recognize and count persons is crucial in many real-world applications, including surveillance, security, and crowd management, making it one of computer vision's most fundamental tasks. You Only Look Once (YOLO) is one of the most effective deep learning models for object identification and counting in recent years. This research seeks to learn more about the YOLOv8 algorithm for precisely counting people in still photos and moving videos. The YOLO method has been at the forefront of computer vision due to its ability to recognize things in real time. People in a crowd typically overlap and block one other, and perspective effects can result in enormous changes in human size, shape, and appearance in the image, all of which make accurate headcounts challenging.The YOLO methodology and its adaptation for population census are the subject of this research. Results from experiments support the usefulness of the proposed approach. Surveillance, crowd control, traffic monitoring, retail analytics, event management, and urban planning are just some of the potential uses highlighted by the findings of this study. Mean Average Precision (MAP) numbers demonstrate that the identification procedure was successful, and the counting process was accurate to within 100%.

1 Introduction

The goal of computer vision-based crowd counting is to estimate how many individuals are visible in still or moving media. Researchers have been paying more attention to it as of late because of its usefulness in many practical contexts, including surveillance, public safety, traffic control, farm monitoring, and cell counting.

Methods developed for crowd counting also have promising applications in other areas, such as traffic congestion estimation [1] Microscopy image analysis for cell and bacterial count [2] ; animal population estimates for ecological survey[3] for several reasons:

- Crowds typically overlap and impede one other.
- perspective effects may create vast disparities in human size, form, and appearance [4].

Human identification and counting systems used hand-crafted features and heuristics, which required substantial topic expertise and calibration. However, deep learning allows data-driven algorithms to automatically train features and models from enormous datasets [5].

Indoor and outdoor positioning systems can be further divided. Indoor areas may create complex infrastructures, making localization more difficult than outside [6]. Crowd counting is a pixel-wise regression issue with several solutions. The tagging provided by unary all regression alone is typically noisy and unreliable [5].

*Corresponding author: phd202120683@iips.edu.iq

YOLO is a well-known deep learning model for object identification and counting, and its model has been demonstrated to reach state-of-the-art performance on a wide range of benchmark datasets, all while maintaining high accuracy and smooth frame rates on high-end GPUs [7] [8].

Our primary contributions are how to utilize yolov8 to recognize people using ROI and a simple method, Fig (1) demonstrates the primary work phases.

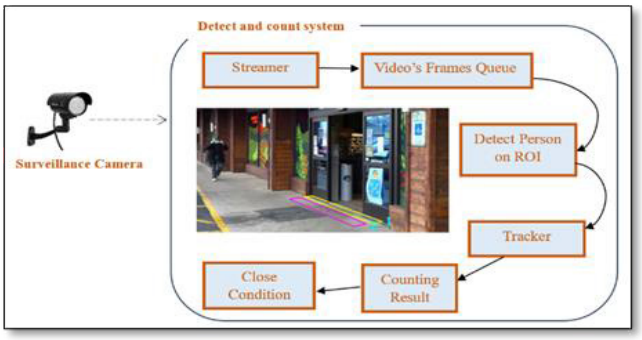

Fig. 1. The main steps of the proposed work

The count will include arrivals and departures. Numerous trials have determined embedded devices to govern occurrences, such as a warning or notification will be shown when a certain number of individuals arrive.

2 Related work

This review will focus on crowd counting and yolov8's application, which have improved recently. Summaries of recent key works on the topic are below. For video crowd counting, LSTNs use a Convolutional Neural Network (CNN) and an Adam optimizer to estimate the density map for each frame [9].

In[10] , the authors propose a Perspective Crowd Counting Network (PCC Net) that uses the DULR module to encode perspective changes in all four cardinal directions, as well as DME, R-HDC, and FBS.

Diverse density distributions reduce congestion. Neural network Scaling factors automatically adjust sub-region density forecasts to reduce local estimation error in ASNet. DANet pixel sizes ASNet's attention masks. Adaptive Pyramid Loss (APLoss) hierarchically estimates subregion losses to avoid training bias [11].

An end-to-end trainable deep architecture applies features from multiple receptive field widths to each picture pixel. Since it adaptively encodes the contextual information needed to estimate crowd density, this method outperforms current crowd counting methods, especially when perspective effects are large [12].

Crowd counting is limited by diversity. The innovative Deep Structured Scale Integration Network (DSSINet) beats state-of-the-art approaches on four hard benchmarks utilizing structured feature representation learning and hierarchically structured loss function optimization to account for people's scale diversity [13].

A perspective-aware convolutional neural network (PACNN) counts crowds effectively. Density regression using perspective information improves our knowledge of picture subjects' sizes. Perspective maps contain two viewpoint-aware weighting layers to dynamically mix multi-scale density outputs. The combined density map resists crowd picture perspective distortion [14].

Crowd counting outdoors is difficult due to the unpredictable climate and diverse population. Researchers created a data collector and labeler to produce synthetic crowd situations and annotate them without human labor [15].

With a total of 2,133,375 annotated heads with points and boxes [16], the dataset is intended to solve the constraints of previous small-scale datasets and to aid the development of supervised CNN-based algorithms for crowd analysis.

A single-column convolutional neural network (CNN) called a perspective-guided convolutional network (PGCNet) has been presented as a solution to the problem of large-scale changes within a single scene caused by the perspective effect [17] .

Proposed online or mobile-based applications may be built and placed at needed locations [18] as a path for future study in the domain of crowd counting and to stimulate the creation of genuine applications that can be utilized in real settings.

The FishEye8K dataset is a public benchmark for road object detection tasks using fisheye cameras had been used with (SoTA) model; it consists of 8,000 images recorded in 22 videos by 18 fisheye cameras for traffic monitoring and features 157K bounding boxes across five classes [19].

A new dataset for multi-camera people tracking, built with actual and synthetic data for training and assessment with multi-target multi-camera (MTMC) people tracking and other methods, was recently released by a researcher in preparation for the 7th AI City Challenge [20].

3 The YOLO Algorithm

The YOLO model, created by University of Washington researchers Joseph Redmon and Ali Farhadi, is widely used for real-time object recognition and picture segmentation. See fig. (2) for an illustration of YOLO's rapid ascent to fame after its 2015 debut. For this task, we used YOLOv8, a cutting-edge object identification model created by Ultralytics and released on the 10th of January 2023 [21]. In YOLOv8, a single neural network is used for object identification across the board. Its performance is superior to that of competing models, and it has several architectural enhancements that make it more reliable [22].

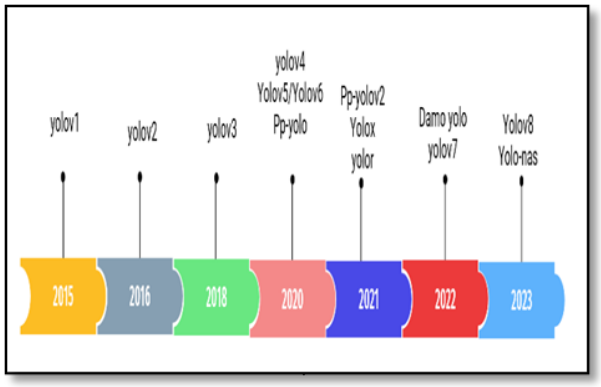

Fig. 2. YOLO versions

3.1 YOLOv8 Architecture

With minor tweaks to the CSPLayer (now known as the C2f module), YOLOv8 is structurally identical to YOLOv5. Improve detection precision with the use of a cross-stage partial bottleneck with two convolutions (C2f module). A decoupled head anchor-free model lets YOLOv8 execute abjectness, classification, and regression separately. Focusing on specialized nodes improves model accuracy. A sigmoid function triggered the abjectness score in YOLOv8's output layer. The SoftMax function represents class probabilities. YOLOv8 uses binary cross-entropy for classification loss and CIoU and DFL for bounding box loss [21] [23] [24]. see fig (3) for general architecture of yolo [25].

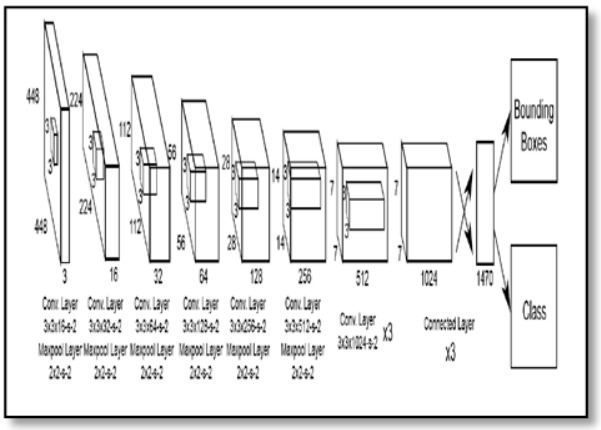

Fig.3. YOLO architecture

3.2 Grid Cell & Bounding box

YOLO is a single deep convolutional neural network that divides the input image into a grid of cells; unlike image classification or face detection [26], the YOLO algorithm's output contains a vector for each grid cell that indicates whether an object is present in that cell, as well as its class and a prediction of its bounding box, see fig (4). YOLO segments the input picture into SS grid cells, and then predicts B bounding boxes and a score for each of the C classes inside each grid cell. Five predictions—center x, center y, width, height, and confidence of the bounding box make up each bounding box [25]. Many studies have shown that YOLO variants of this processing mechanism are faster than alternatives [27] .

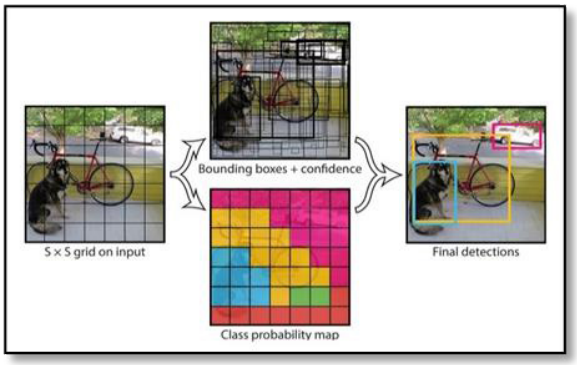

Fig.4. Grid Cell & Bounding box in yolo

Besides the principle of it work, YOLOv8 had many models as shown in table (1) below, where each model has it own feature and parameters.

Table 1. Models of YOLOv8

Model	size (pixels)	mAP$^{val}_{50-95}$	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)
YOLOv8 n	640	37.3	80.4	0.99	3.2
YOLOv8 s	640	44.9	128.4	1.20	11.2
YOLOv8 m	640	50.2	234.7	1.83	25.9
YOLOv8l	640	52.9	375.2	2.39	43.7
YOLOv8 x	640	53.9	479.1	3.53	68.2

4 Experiment and Results

4.1 dataset

A test-dev 2017 MS COCO dataset was utilized for the experiment. This dataset employs a more involved approach to AP calculation, and it has 80 item types. It employs a 101-point interpolation, which is more precise than an 11-point interpolation since it accounts for a wider range of recall thresholds, from 0 to 1. Also, unlike a typical AP statistic known as AP50, which is the AP for a single IoU threshold of 0.5, the AP is generated by averaging over numerous IoU values instead of just one. With image size of 640 pixels, YOLOv8x was able to get an AP of 53.9%.

4.2 Test and Results

The YOLOv8 algorithm can be used to count people in images and videos by simply counting the number of bounding boxes with a confidence score above a certain threshold. This threshold can be adjusted to control the accuracy of the counting. the YOLOv8s is the type that been used in the testing.

The YOLOv8 /Small model as can be seen from the table (1) has a suitable speed to train and test each patch of dataset, which is 128.4 Ms. If the application has a requirement of more speed, we can use another model of yolov8. All the weights are the same as the ones found in the original coco dataset. If the score comparing the ground-truth bounding box to the detected box is higher, then the model is more accurate in the detections task, and the MAP values for single-model single-scale on the COCO val2017 dataset are 44.9, indicating an accurate performance.

For implementing the research idea, a nearly one-minute video of many people attempting to enter a specific location was used. After applying the proposed algorithm to the selected video, the process began by defining two regions of interest (ROI) next to the door, and according to the movement of the people and interaction with region (1) and region (2), the counter began to count the people into two types of counts, one for people going inside the building and the other for people leaving the, see fig (5).

Fig. 5-a. detects the person inside both IOR and counting the entering person.

Fig. 5-b. counting the exiting person.

Fig (5-c). counting the existence person.

As seen above, both the detection and counting of persons were completed effectively and with great performance. The alternative option for this test is to regulate the event based on the number of people inside the area; for our testing, we opted to issue a notification and warning throughout the test when the number of people inside the area reaches 10, see fig (5-c).

This option can be used in a variety of situations, such as the integration of Counting People Using YOLO Model with IoT devices and artificial intelligence (AI) techniques, which consider a current area of research that has received significant attention in recent years, such as CO2 calculation or closing the door for a specific purpose.

During the testing of the movie, it was discovered that the model can recognize each object in each frame, see fig (6), and that even when it detects several objects, there is no confusion in the outcome, and the model successfully detects and counts the person without any errors.

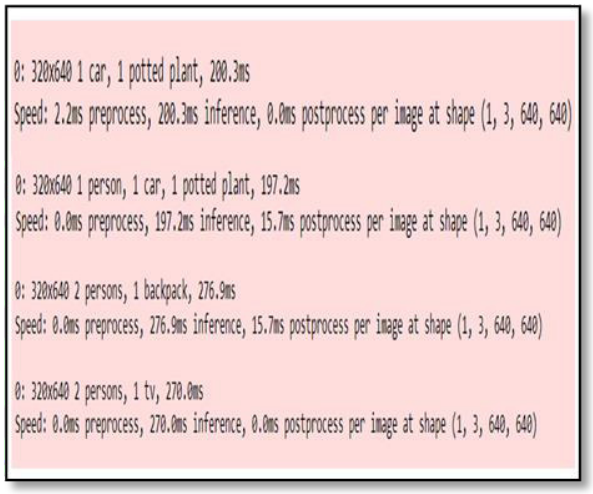

Fig. 6. Model Multiple Detection

5 Conclusion & Future Applications

YOLOv8 is a great tool for identifying and counting people since it's easy to use and has numerous features that may be utilized for diverse applications, especially when the effort, resources, and customization are specialized. The strategy outperforms people counts and video analytics tools for counting people. The idea might be used to count people during festivals, celebrations, and Pilgrimages in the future. The other trend is integrating COUNTING PEOPLE USING THE YOLO MODEL with a security system to open or close a door, IOT devices to compute CO2 to reduce shortness of breath, people flow analysis to find bottlenecks, and daily reports for people entering and leaving.

References

1. T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 785–800.
2. V. Lempitsky and A. Zisserman, "Learning To Count Objects in Images."
3. Z. Ma, L. Yu, and A. B. Chan, "Small Instance Detection by Integer Programming on Object Density Maps."
4. Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian Loss for Crowd Count Estimation with Point Supervision." [Online]. Available: https://github.com/ZhihengCV/
5. A. Zhang et al., "Relational Attention Network for Crowd Counting."
6. A. A. Hashim, M. M. Rasheed, and S. A. Abdullah, "ANALYSIS OF BLUETOOTH LOW ENERGY-BASED INDOOR LOCALIZATION SYSTEM USING MACHINE LEARNING ALGORITHMS."

7. A. Kumar Suhane, A. Vani, and U. Raghuwanshi, "HUMAN DETECTION AND CROWD COUNTING USING YOLO." [Online]. Available: https://www.researchgate.net/publication/370341591
8. H. Gomes, N. Redinha, N. Lavado, and M. Mendes, "Counting People and Bicycles in Real Time Using YOLO on Jetson Nano," Energies (Basel), vol. 15, no. 23, Dec. 2022, doi: 10.3390/en15238816.
9. Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, "Locality-constrained spatial transformer network for video crowd counting," in Proceedings - IEEE International Conference on Multimedia and Expo, IEEE Computer Society, Jul. 2019, pp. 814–819. doi: 10.1109/ICME.2019.00145.
10. J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3486–3498, Oct. 2020, doi: 10.1109/TCSVT.2019.2919139.
11. X. Jiang et al., "Attention Scaling for Crowd Counting."
12. W. Liu, M. Salzmann, and P. Fua, "Context-Aware Crowd Counting." [Online]. Available: https://sites.google.com/view/weizheliu/home/
13. L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd Counting with Deep Structured Scale Integration Network."
14. M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting Perspective Information for Efficient Crowd Counting."
15. Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from Synthetic Data for Crowd Counting in the Wild." [Online]. Available: www.youtube.com/watch?v=Hvl7xWkIueo.
16. Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 6, pp. 2141–2149, Jun. 2021, doi: 10.1109/TPAMI.2020.3013269.
17. Z. Yan et al., "Perspective-Guided Convolution Networks for Crowd Counting."
18. A. Dalwadi et al., "Detecting and Counting People In Dense Crowd," 2012. [Online]. Available: www.ijfans.org
19. M. Gochoo et al., "FishEye8K: A Benchmark and Dataset for Fisheye Camera Object Detection." [Online]. Available: https://github.com/MoyoG/FishEye8K
20. M. Naphade et al., "The 7th AI City Challenge."
21. J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO: From YOLOv1 and Beyond," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.00501
22. S. Manzoor, Y. C. An, G. G. In, Y. Zhang, S. Kim, and T. Y. Kuc, "SPT: Single Pedestrian Tracking Framework with Re-Identification-Based Learning Using the Siamese Model," Sensors, vol. 23, no. 10, May 2023, doi: 10.3390/s23104906.
23. X. Li et al., "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection."
24. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," 2016. [Online]. Available: https://github.com/Zzh-tju/DIoU.
25. M. H. Putra, Z. M. Yussof, K. C. Lim, and S. I. Salim, "Convolutional Neural Network for Person and Car Detection using YOLO Framework".
26. Dr. S. Gothane, "A Practice for Object Detection Using YOLO Algorithm," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 268–272, Apr. 2021, doi: 10.32628/cseit217249.
27. L. Qi et al., "Ship target detection algorithm based on improved faster R-CNN," Electronics (Switzerland), vol. 8, no. 9, Sep. 2019, doi: 10.3390/electronics8090959.