DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

Detection And Count of Human Bodies In a Crowd Scene Based on Enhancement Features By Using The YOLO v5 Algorithm

Mohammed Abduljabbar Ali[1], Abir Jaafar Hussain[2], Ahmed T. Sadiq[3]
[1, 3]*Department of Computer Science, University of Technology, Baghdad, Iraq*
[2] *School of Computer Sciences and Mathematics Liverpool John Moores University*
[1]*cs.19.21@grad.uotechnology.edu.iq*, [2]*A.Hussain@ljmu.ac.uk*, [3]*ahmed.t.sadiq@uotechnology.edu.iq*

Abstract— *Crowd detection has various applications nowadays. However, detecting humans in crowded circumstances is difficult because the features of different objects conflict, making cross-state detection impossible. Detectors in the overlapping zone may therefore overreact. The proposal uses the YOLO v5 (You Only Look Once) method to improve crowd recognition and counting. This algorithm is entirely accurate and detects things in real-time. The idea relies on edge enhancement and pre-processing to solve overlapping feature regions in the image and improve performance. The CrowdHuman data set is used to train YOLO v5. The system counts the number of humans in the image to detect a crowd. Before training, this model enhanced the image with several filters. The YOLO v5 algorithm distinguishes a person inside a crowd by utilizing the surrounding box on the head and overall body. Therefore, the number of head detection is x-coordinated compared to the body. Assume the detected heads outnumber the bodies. A square of the head will be extracted, but not a body square. Also, cropping the image reduces interference between human beings and enhances the edge features. Thus, YOLOv5 can detect it. The idea improves head and body detection by 2.17 and 4.1 percent, respectively.*

Index Terms — *human Detection, crowed, deep learning, YOLO v5, feature enhancement.*

I. INTRODUCTION

Detection of the human body is amongst the most significant study disciplines in computer vision, with many applications such as surveillance cameras, robots, and automated driving [1] [2]. Although there have been substantial advancements, detecting people in crowded situations using various gestures is tricky. Obstruction between human instances is typical in crowded environments, making detained person's visual patterns less selective and challenging to identify [3]. Two factors can cause poor object detection performance. First, compared to a somewhat or partially veiled item, an obscured object's visual signals are weak, making it harder to stand out against other backdrops or human bodies. Furthermore, the visual feature differences between the bodies of overlapping objects present a significant problem for one human to detect in the midst of both [4].

While Convolutional Neural Network (CNN) features include discriminative representations, deep Convolutional Neural Networks (CNN) are used to extract features for object identification. It has semantic characteristics that let it recognize things more accurately

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

[5]. CNN's are often used in a detection head and a backbone network (for image classification) [6]. The backbone uses the Convolutional Neural Network (CNN) architecture to learn the image's characteristics, and the detecting head uses these features to anticipate the bounding boxes. Several geographical aspects are taken into account to increase the network's efficiency and performance. Object detectors [7] are often two-stage techniques such Region-Based Convolutional Neural Networks (Faster R-CNN) [8] and Region-based Fully Convolutional Networks (R-FCN) [9], or single-shot detectors like You Only Look Once (YOLO) [10]. Single Shot Detector (SSD) [11].

In this paper, humans are detected in crowd scenes by optimizing the overlapping features between humans with each other and background frames in video with a series of improvements before proceeding with the YOLO v5 algorithm. In addition, it has been proposed to improve the edges in a proposed method so that the algorithm can object detection from the background and interference between features with another object. This method proved successful through its evaluation.

This paper is organized in the following manner: The second section contains related Work. Then the third section is the methodology, while Section 4 is the result and evaluation. Finally, the conclusion section .

II. RELETED WORK

This project focuses on the detection of humans in a crowded environment. As a result, this study reviews recent works on human detectors based on deep learning algorithms.

Hyeok-June Jeong et al. [2] This study provided a method for pre-processing training images gathered by a web search engine Optimized for YOLO. This system has four steps to produce the proper training picture: Images picker, Size Modifier, Images Maker, and Annotation Creator. Thanks to this method, the crawling photos are now appropriate for YOLO training. The outcome was positive. The YOLO training diary has a high value as well. The findings of the object detection were similarly excellent. Zheng Ge et al. [4] PS-RCNN, a two-stage detector version, is introduced. PS-RCNN uses an R-CNN [1] module to identify slightly/none occluded objects (referred to as P-RCNN) and then uses human-shaped masks to suppress the discovered instances. The characteristics of substantially occluded instances can stand out. The rest of the items overlooked by P-RCNN are then detected by PS-RCNN using another R-CNN module specializing in strongly obstructed human detection (referred to as S-RCNN). The ensemble of these two RCNNs' outputs is the final result. Ruiqi Lu et al. [12] The detection of the head is formed by employing a particular section of a labeled box to detect the appropriate part of the human body, which enhances performance and resistance to occlusion. Furthermore, the head-body alignment structure is explicitly examined by adding Alignment Loss, a self-supervised function. Based on this, the head-body alignment net (HBAN) is proposed in this paper to improve pedestrian detection by exploiting the human head beforehand. Xing Hu et al. [13] This suggestion Detection of Abnormal Behavior and Localization in a weakly supervised framework is suggested. The Faster Regional Conventional Neural Network algorithm detects objects in the scene, such as cars, walkers, etc. Histogram of Large Scale Optical Flow (HLSOF) descriptor is used to describe object behavior. Finally, classification behaviors as normal or abnormal by training the Multiple Instance Support Vector Machine (MISVM).

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

Li Yuan et al. [14] This study aims to demonstrate a pipeline for tracking human poses in busy and complicated contexts. First, it assigns a human ID to each bounding box created by the detection model using a multi-object tracking mechanism. After that, a posture is created for each bounding box with ID. Finally, optical flow is employed to construct the final posture tracking result by using the temporal information in the movies.

III. PROPOSAL METHODS

Overlapping areas in an image affect object detection. This degrades the quality of classification of objects by the technologies due to this interference. Methods segmentation of the image is by regions or edges. Through experiments, it is found that edges further improve segmentation. Therefore, this proposal is used to improve edges by filters.

The basic idea is to detect crowd scenes by counting and body detection of humans in an image. A series of operations process the first step, the input image from the video. Next, detection of humans by YOLO v5 algorithm trained on Crowd Human data set. Count of heads detection in an image. If the number of heads is more significant than the number of human bodies, a box is cut around the head. Implementation of proposal edges enhancement method and accuracy improved on the part cut. Finally, the improved picture is re-examined by previous processes of identifying humans in crowding. *Fig. 1* shows an overview of the structure of the proposed procedure detection system.

A. Pre-processing

The video is processed by splitting it into many frames in this step. Each frame is an image scaled according to the YOLO v5 algorithm input. The image is also improved by removing noise and blurring with the Gaussian and sharpening filter. In multi-scale edge detection algorithms, 2D Gaussian filters are often used for three primary reasons. This is an essential consideration since only 2D Gaussian filters do not produce false edges as the scale increases when used with a Laplacian operator. Secondly, when appropriately used, Gaussian filters offer the optimal compromise between spatial and frequency localization. Third, because only rotationally invariant 2D filters can be separated in horizontal and vertical directions, spatial convolution with Gaussian filters is highly efficient [15]. The smoothing filter is based on the Gaussian function of discrete the double by zero-meant. This proposal uses a Laplacian filter to sharpen the image after removing noise by a Gaussian filter. An image's Laplacian filter reveals areas of significant intensity change. As a result, edge sharpening is a common application for the LF. This operator excels in spotting the essential details in an image. LF will enhance the sharpness of any features that have sharp discontinuities. *Fig. 2* shows enhanced images by using filters.

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

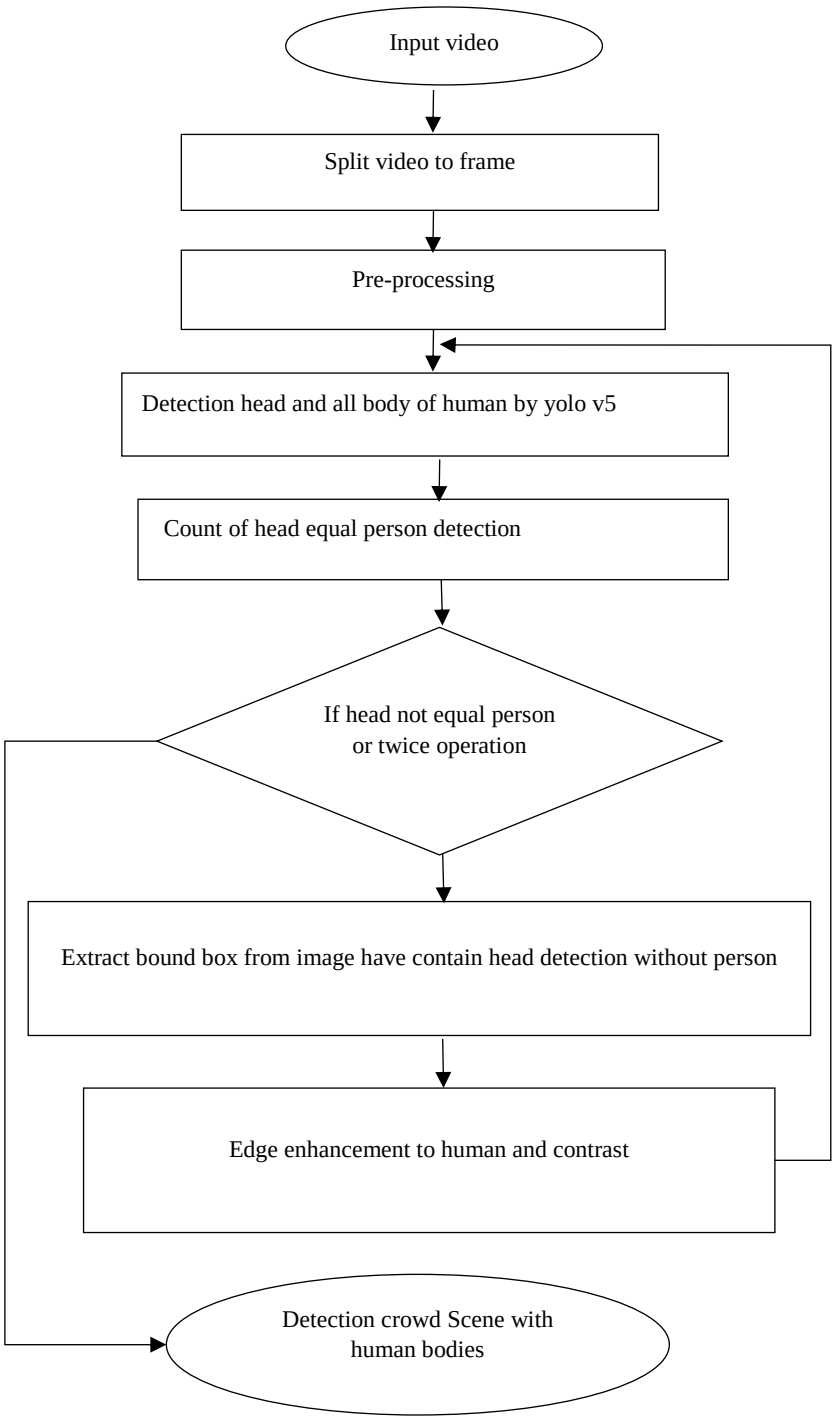

FIG. 1. FLOWCHART PROPOSAL.

Received 11/March/2022; Accepted 5/May/2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

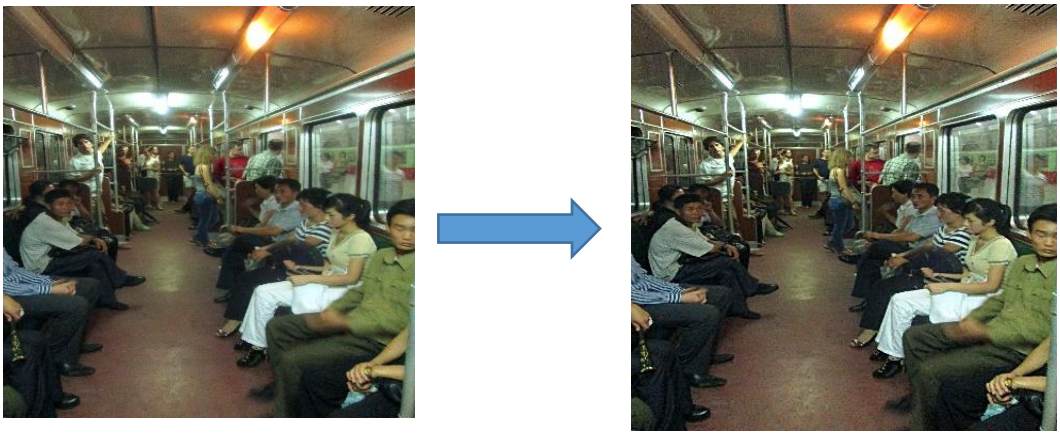

FIG. 2. EXPLAIN THE APPLICATION OF FILTERS TO IMAGE ENHANCEMENT.

B. Human head and body detection by YOLO v5

The YOLO v5 algorithm is a one-stage detector using a region-based object detection network. The YOLO redefines object detection as a regression problem with a high processing speed. Recent applications of YOLO v5 include real-time person searches and a vision system[16][17]. The three fundamental components of YOLO v5 are the backbone, the head, and the detection. A Convolutional Neural Network (CNN) is the system's backbone, which collects and shapes image properties at different granularities. The YOLO v5 employs the Center and Scale Prediction (CSP) Bottleneck to establish visual attributes. Layers make up the head, which combines visual properties before transmitting them to a prediction algorithm. The PA-NET is also used for feature aggregation in Yolo v5[18]. The detection is a technique that incorporates the box and class prediction stages and head features. The YOLO v5 architecture is shown schematically in *Fig. 3*.

The Crowd Human data set improved the human detection body in the model using YOLO v5. In the test model process. *Fig. 4* illustrates the head and human bodies by the YOLO v5 algorithm. These steps following to train YOLO v5 on these data sets:

- Data preparation for the detection of the "head" (0) and the "person" (1) classes, where the "person" class corresponds to "complete body" (including occluded body regions) in the original "CrowdHuman" annotations. But to train the Yolo-v5 model, we need to organise our dataset structure and it requires images (.jpg/.png, etc.) and their matching labels in.txt format that comprises all the (labels x center, y center, width, height) values for that single image file.
- Second, download the pre-trained weights by copying over all the necessary files. Changes must be made to this file after it has been saved to be compatible with Yolo-v5.
- Yolo-v5 model's *.yaml file has to be edited. We simply need to change the number of classes in our model's YAML file to match the number in the model. Two classes were employed when it comes to a head and body (head and body).

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

- Divide the dataset into training and validation and save the related photos and labeled.txt files. After training the "yolov5" model, we were able to run this model through its paces.

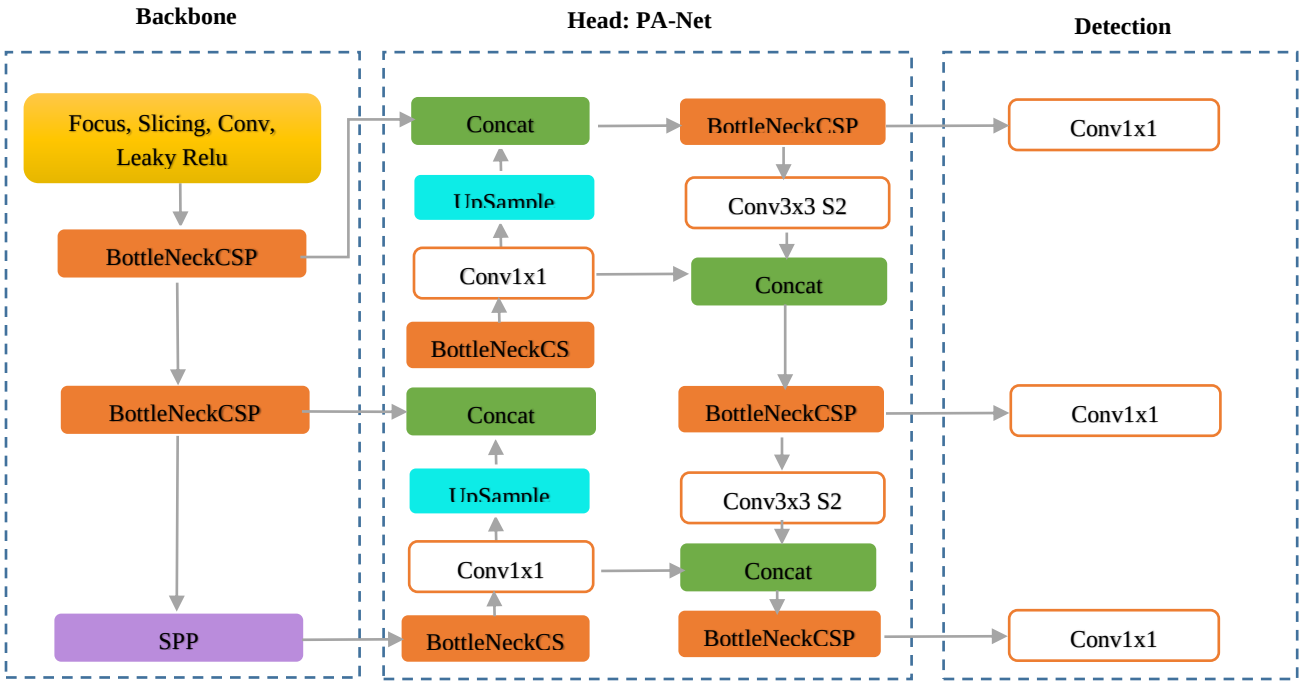

FIG. 3. ARCHITECTONOCS OF THE YOLOV5[18].

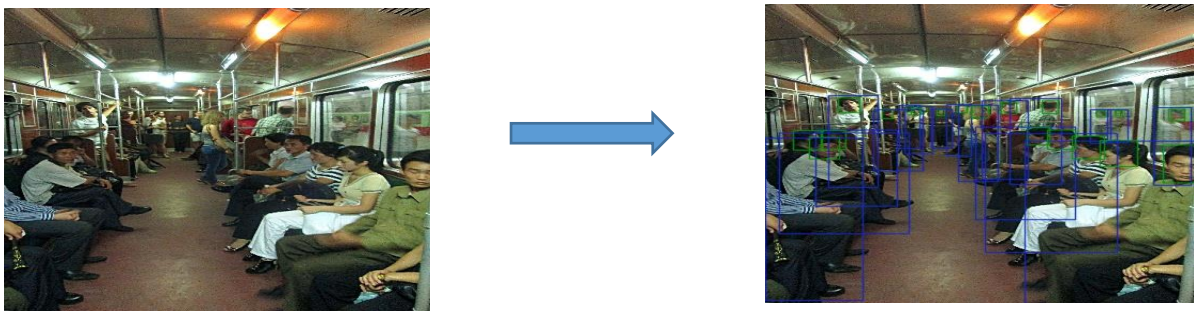

FIG. 4. IMPLEMENTATION OF THE YOLO V5 ALGORITHM ON CROWDHUMAN DATASET.

In the previous step, detection of the head and human bodies in the YOLO v5 algorithm. In this step, the number of head detectors is compared with the number of people. This model calculates and checks if the heads are more than the number of bodies for image enhancement. The process is done by storing the coordinates of the head (X and Y), calculating the center, and comparing it with the center of the body. If there is no stock center of the body, nobody detects his head. Therefore, the value of the x of the center of the head and the body is approximately equal.

Received 11/March/2022; Accepted 5/May/2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

C. Enhancement segmentation

A standard tool in image processing is segmentation. When the digital image is partitioned into multiple regions, and each region is given a unique label, the visual characteristics of the pixel in each region tend to be similar. In the image segmentation process, several different approaches and mechanisms can be used, such as thresholding, border tracking, clustering, and techniques based on regions. Image segmentation augmentation based on the edge is the subject of this paper.

When the number of heads and bodies is checked in the previous step with the YOLO v5 algorithm, the number of people is less than the number of heads, especially in crowds. A method has been proposed to improve the detection of people by improving the features of the body in the image. The operation is done by subtracting a square whose coordinates depend on the coordinates of the head and its center, which are processed by the code. The cropped part of the image is entered into several filters to improve the display of features so that YOLO v5 can detect the human body. The suggested filter for edge enhancement of a color image is based on the canny filter. The cut part of the image is processed by converting it to grayscale and applying a canny filter. In the next step, the locations of the values that contain a value that is considered an edge are stored. The clarity of the pixels of the color image is increased depending on the locations of the stored values of the edges, and the increase in clarity is according to the neighboring values. Finally, the part that has been improved is tested with the YOLO v5 algorithm and merged with the original image. This optimization aims to detail the object's properties further and improve the edges to enable YOLO v5 to explore the object for the explored vertices. The results showed an improvement in the characterization of people in crowding.

IV. RESULTS AND EVALUATION

A. Data set

The CrowdHuman dataset is a testbed for human detectors in crowded environments. It's broken down into three sections: training (15000 photos), verification (4370 photos), and testing (5000 photos). There are 470k human cases from the training and verification subsets, with 22:6 people per picture and other occlusions [19]. Each human instance has a head-bounding box, a human visible-region bounding box, and a full-body bounding box. Because the online assessment server for the testing dataset is presently unavailable, yolov5 was trained on the CrowdHuman dataset and assessed on the validation dataset. During training, the input photo is resized such that the short edges are 800 pixels long and the long edges are no longer than 1333 pixels. The anchor aspect ratios for the CrowdHuman dataset are 0:5, 1, and 2 [20].

B. Evaluation method

Object detection has two tasks: one is to identify whether an item exists in the picture, and the other is to locate the object. A typical dataset will also include numerous classes with non-uniform distributions. With a simple accuracy-based measure, bias will develop since it is also vital to quantify the risk of misclassification. Consequently, a definitive score was provided to evaluate the model at varying degrees of confidence by identifying each BBox. The Mean Average Precision (mAP) [21] was chosen as the assessment criterion in this Work. The mAP

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

score is calculated by multiplying the mAP by all classes and all Intersection over Union (IoU) criteria.

In most cases, the IoU value is between 1 and 0. The detection is categorized as True Positive if the object's value is greater than the threshold (TP). It's a False Positive if the value is less than the threshold (FP). The detection would be classified as False Negative if the ground truth failed to identify a value (FN). As a result,. (1) and (2) are used to derive precision and recall [5]:

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

And

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

The (mAP) numbers represent the object detection sets' accuracy compared to ground truth. The (IoU) is utilized to construct the (mAP), which specifies the intersection between the predicted image and the ground truth..(3) can calculate mAP after computing the interpolated precision's mean for each information's recall level. In this experiment, the IoU threshold was set to 0.5, and the mAP was averaged across the human detection classes.

Table I displays the detection model results and findings from the CrowdHuman benchmark, providing the mean average accuracy.

$$mAP = \frac{1}{n} \sum_{r\in\{0,0.1\ldots1\}} P_{interp(r)} \quad (3)$$

Where n is the number of interpolations employed, and P (interp(r)) denotes the interpolated precision that takes the highest precision.

C. Comparison

The three approaches' performance is compared to Yolo's data set training in the Table I. Yolov5 with CrowdHuman dataset training has an mAP of 0.771, 0.166 lower than the proposed approaches. On the other hand, Yolov5 with Coco dataset map has the lowest mAP at 0.041. As a result, the proposed approaches provide a better balance of accuracy and speed in algorithm development. These models were human body detection and a count of persons in a crowd scenario.

TABLE I .RESULT FROM THE DETECTION MODEL

model	mAP	
	Head	**Body of human**
Yolov5+coco dataset training	-	0.771
Yolov5+ CrowdHuman dataset training	0.934	0.896
proposed methods	0.956	0.937

To some extent, our method is accurate and easy to implement. The results are presented in Table II, which compares our approach to human detection in the crowd by listing methods, Features, and mAP.

Received 11/March/2022; Accepted 5/May/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

TABLE II. COMPARING OUR APPROACH TO OTHERS

Methods	Features	mAP
novel Representative Region NMS (R2NMS)[20]	Paired-Box Model (PBM)	89.29
YOLO v3[21]	three-dimensional feature space	81.03
Multiple Instance Support Vector Machine (MISVM)[13]	Histogram of Large Scale Optical Flow (HLSOF)	88.3
R-CNN + human-shaped masks[4]	High Resolution RoI	87.94
proposed methods	Edge + interest region	93.7

V. CONCLUSIONS

In this proposal, the YOLO v5 algorithm optimizes the detection of the human body within a crowd. The previous algorithm failed to detect human bodies in several crowd scenes. The CrowdHuman data set was used in training YOLO v5. The comparison between the proposal with the previous algorithm and the Coco and CrowdHuman data set showed that the proposal has an accuracy of 95.6% and 93.7% for the head and body, respectively. The main conclusion for why the proposal performed better in detecting the human body with the CrowdHuman dataset. The reason back to the focus was on improving the overlapping traits between background and humans in others. Several filters were applied in and hyper in the pre-process. A method has also been proposed to enhance the edges for detecting the human body based on the detection of heads. In addition to identifying crowd scenes in an image by calculating the human detection.

In order to further develop the possibility of object detection algorithms. This proposal shows through the results, that efforts to enhancement the data set for training increase the efficiency of the detector.

REFERENCES

[1] M. Abduljabbar Ali, A. Jaafar Hussain, and A. T. Sadiq, "Deep Learning Algorithms for Human Fighting Action Recognition," Int. J. Online Biomed. Eng., vol. 18, no. 02, pp. 71–87, Feb. 2022, doi: 10.3991/ijoe.v18i02.28019.

[2] H. J. Jeong, K. S. Park, and Y. G. Ha, "Image Preprocessing for Efficient Training of YOLO Deep Learning Networks," Proc. - 2018 IEEE Int. Conf. Big Data Smart Comput. BigComp 2018, pp. 635–637, 2018, doi: 10.1109/BigComp.2018.00113.

[3] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "PedHunter: Occlusion robust pedestrian detector in crowded scenes," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 10639–10646, 2020, doi: 10.1609/aaai.v34i07.6690.

[4] Z. Ge, Z. Jie, X. Huang, R. Xu, and O. Yoshie, "PS-RCNN: Detecting secondary human instances in a crowd via primary object suppression," Proc. - IEEE Int. Conf. Multimed. Expo, vol. 2020-July, pp. 1–6, 2020, doi: 10.1109/ICME46284.2020.9102793.

[5] M. M. Mahmoud and A. R. Nasser, "Dual Architecture Deep Learning Based Object Detection System for Autonomous Driving," Iraqi J. Comput. Commun. Control Syst. Eng, vol. 21, no. 2, pp. 36–43, 2021.

[6] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[7] P. Soviany and R. T. Ionescu, "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," Proc. - 2018 20th Int. Symp. Symb. Numer. Algorithms Sci. Comput. SYNASC 2018, pp. 209–214, 2018, doi: 10.1109/SYNASC.2018.00041.

[8] K. H. Shih, C. Te Chiu, J. A. Lin, and Y. Y. Bu, "Real-Time Object Detection with Reduced Region Proposal Network via Multi-Feature Concatenation," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 6, pp. 2164–2173, 2020, doi: 10.1109/TNNLS.2019.2929059.

Received 11/March/2022; Accepted 5/May/2022

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.2.11

[9] I. Idrees, S. P. Reiss, and S. Tellex, "RoboMem: Giving Long Term Memory to Robots," ICRA2019 Work., 2020, [Online]. Available: https://arxiv.org/abs/2003.10553v1.

[10] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," Procedia Comput. Sci., vol. 199, pp. 1066–1073, 2022, doi: 10.1016/j.procs.2022.01.135.

[11] A. Arora, A. Grover, R. Chugh, and S. S. Reka, "Real Time Multi Object Detection for Blind Using Single Shot Multibox Detector," Wirel. Pers. Commun., no. 0123456789, 2019, doi: 10.1007/s11277-019-06294-1.

[12] R. Lu, H. Ma, and Y. Wang, "Semantic head enhanced pedestrian detection in a crowd," Neurocomputing, vol. 400, pp. 343–351, 2020, doi: 10.1016/j.neucom.2020.03.037.

[13] X. Hu et al., "A weakly supervised framework for abnormal behavior detection and localization in crowded scenes," Neurocomputing, vol. 383, pp. 270–281, 2020, doi: 10.1016/j.neucom.2019.11.087.

[14] L. Yuan et al., "A Simple Baseline for Pose Tracking in Videos of Crowed Scenes," MM 2020 - Proc. 28th ACM Int. Conf. Multimed., pp. 4684–4688, 2020, doi: 10.1145/3394171.3416300.

[15] M. Abduljabbar Ali, A. Jaafar Hussain, and A. T. Sadiq, "Human Fall Down Recognition Using Coordinates Key Points Skeleton," Int. J. Online Biomed. Eng., vol. 18, no. 02, pp. 88–104, Feb. 2022, doi: 10.3991/ijoe.v18i02.28017.

[16] Y. Li, K. Yin, J. Liang, C. Wang, and G. Yin, "A Multi-task Joint Framework for Real-time Person Search," no. 2006, 2020, [Online]. Available: http://arxiv.org/abs/2012.06418.

[17] W. Cai, C. Wang, H. Huang, and T. Wang, "A Real-Time Smoke Detection Model Based on YOLO-SMOKE Algorithm," 2020 Cross Strait Radio Sci. Wirel. Technol. Conf. CSRSWTC 2020 - Proc., no. 1, pp. 1–3, 2020, doi: 10.1109/CSRSWTC50769.2020.9372453.

[18] J. Ieamsaard, S. N. Charoensook, and S. Yammen, "Deep Learning-based Face Mask Detection Using YoloV5," Proceeding 2021 9th Int. Electr. Eng. Congr. iEECON 2021, pp. 428–431, 2021, doi: 10.1109/iEECON51072.2021.9440346.

[19] S. Shao et al., "CrowdHuman: A Benchmark for Detecting Human in a Crowd," pp. 1–9, 2018, [Online]. Available: http://arxiv.org/abs/1805.00123.

[20] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 10747–10756, 2020, doi: 10.1109/CVPR42600.2020.01076.

[21] N. S. Punn, S. K. Sonbhadra, S. Agarwal, and G. Rai, "Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques," pp. 1–10, 2020, [Online]. Available: http://arxiv.org/abs/2005.01385.

Received 11/March/2022; Accepted 5/May/2022

 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*