CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.035

Detection of *Crowd concentrations* with YOLOv3

Ba Duy Nguyen[1], Thanh Nhan Dinh[1], Thanh Bach Nguyen[2], and Quoc Dinh Truong[3*]

[1]*Can Tho University of Technology, Viet Nam*

[2]*Tra Vinh Provincial Police Department, Viet Nam*

[3]*Can Tho University, Viet Nam*

**Corresponding author (tqdinh@cit.ctu.edu.vn)*

Article info.

Received 14 Jul 2023
Revised 10 Sep 2023
Accepted 22 Sep 2023

Keywords

Object detection, crowded scence, YOLOv3 model

ABSTRACT

Crowd detection using street cameras has attracted a lot of research in recent years. In this paper, we propose a simple, fast, and effective method using YOLOv3 model for crowd detection. Using image frames extracted from surveillance video, pedestrian objects are detected, counted and a warning signal is sent out when a crowd occurs. The obtained results on test data extracted from 2 data sets STCrowd, SmartCity, and our self-collected dataset confirm the feasibility of the proposed method.

1. INTRODUCTION

Currently, surveillance cameras have been deployed and used in various places such as offices, companies, train stations, shopping centers, bank offices, ATMs, and more. In these cases, video surveillance serves as the basis for observing and detecting behaviors and suspicious conditions that violate safety and security regulations in a specific area. Among these, the concentration of people at a location (e.g., in front of a military base, government office, etc.) in certain specific contexts is not allowed and needs to be promptly detected and appropriately alerted. Crowd concentration refers to a large number of people gathering together at a specific location and time for a common purpose, which cannot be automatically controlled and assessed by computer tools. Currently, with advancements in science and technology, particularly in the fields of computer vision and artificial intelligence, the detection of crowd concentrations can be addressed automatically through the following steps: detecting/marking human objects, counting the number of objects/classifying the crowded or non-crowded status, where the detection of human objects plays a crucial role.

Object detection techniques can be categorized into two approaches: non-neural network-based and neural network-based approaches (Ubale et al., 2021). The non-neural network-based approach first extracts object features from the images and then provides these features to a regression model to predict the position and label of the objects in the image. Some methods in this approach include Scale Invariant Feature Transform (SIFT), Haar-Like features, and Histogram of Oriented Gradients (HOG). On the other hand, the neural network-based approach has attracted significant research interest in recent times and has achieved promising results compared to the non-neural network-based approach. Here are some notable works in this approach.

Yizhou et al. (2019) proposed a method for detecting moving pedestrians by utilizing multiple frames with different scale ratios to detect objects (Single Shot Multibox Detector - SSD). The authors employed the MobileNet network model to enhance the performance of the VGG-SSD model and used

the Non-maximum Suppression algorithm to retain only one bounding box around each human object. This method can detect human objects in images with an accuracy of 93.34%.

Ahmad et al. (2019) utilized SSD for the localization and counting of people in images captured from a top-down (or overhead) view. The proposed system employed a pre-trained model on the COCO dataset and was tested on two datasets of indoor and outdoor images (with the number of people in the images ranging from 1 to 7), achieving accuracies of 95% and 94.42%, respectively.

Bhangale et al. (2020) employed a deep convolutional neural network (DCNN) to detect crowd concentration. The proposed DCNN architecture was based on the CSRNet architecture, comprising 10 convolutional layers and 3 pooling layers. The system was tested on the ShanghaiTech dataset and achieved high accuracy.

Kannadaguli (2020) proposed a system for detecting human objects in thermal infrared images using the YOLO model. In the preprocessing step, the author labeled human objects in the images using the Microsoft Visual Object Tagging Tool (VoTT) and removed noise from the thermal image using a median filter. The system was tested on four datasets and showed that by utilizing the YOLO model, it could successfully detect humans in overhead thermal images or videos.

Byju et al. (2021) improved the YOLOv4 model for detecting human objects in various challenging conditions. The transfer learning technique was integrated into the YOLOv4 model to extract features from a selectively curated training dataset. This enhanced model can detect human objects in weather such as snow, rain, and wind, with an accuracy improvement of over 7% compared to the original YOLOv4 model.

Unlike the aforementioned works, in this paper, we propose to build a system that enables the detection of crowded groups (where the number of people exceeds a predetermined threshold) or the number of people reaching a predefined crowd threshold (frame coverage ratio) in the observed frames. This system can detect crowded situations in monitored areas such as government agency entrances or disease treatment areas requiring social distancing…

2. MATERIALS AND METHOD

Based on the research results, we propose a model to use YOLOv3 model for crowd detection. The architecture of the proposed system is illustrated in Figure 1, comprising three main components: 1- Object detection of people in the frames; 2- Identification of groups of people; 3- Determination of crowd concentration status.

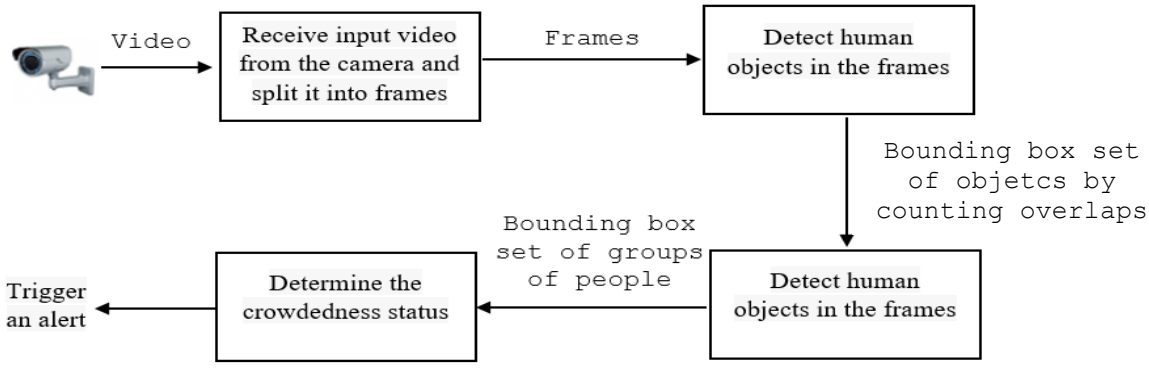

Figure 1. The architecture of the proposed system

Detecting human objects in photo frames

We use the YOLOv3 model (Redmon & Fahadi, 2018) to localize human objects within the image frames. The YOLO network was first introduced by Redmon et al. (2016) in a paper titled "You Only Look Once: Unified, Real-Time Object Detection". The model takes an image as input and predicts bounding boxes and class labels for each box. It operates by dividing the input image into a grid of cells, where each cell predicts bounding boxes if its center falls within the cell. Each grid cell predicts bounding boxes based on the coordinates (x, y) as the center, width, height, and confidence in containing an object inside.

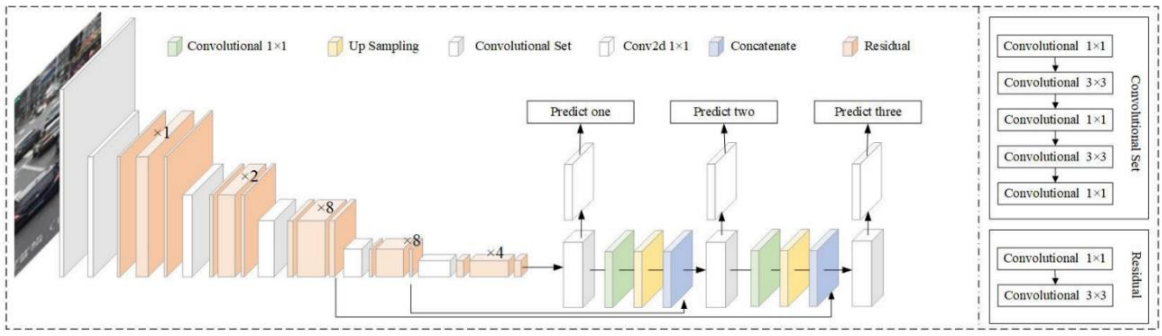

Figure 2. Architecture of the model YOLOv3

YOLOv3 has several improvements compared to its previous versions, such as using logistic regression for predicting the confidence of object labels, employing the Darknet-53 architecture as the backbone, and incorporating the Feature Pyramid Networks (FPN) architecture for generating predictions. In this proposed scope, we use a pre-trained YOLOv3 model, which can recognize 80 object classes, for detecting human objects in images provided at the address: https://pjreddie.com /darknet/yolo/.

Detecting groups of people

We rely on the set of bounding boxes provided by the person detection component in the previous step to establish groups of people for identifying crowded groups. Overlapping bounding boxes (considering a discrete threshold determined by the user) are grouped together. The algorithm details are described:

Input: List of bounding boxes of people.

Output: Set of groups of people.

Algorithm:

- Sort the list of bounding boxes in ascending order of the x-coordinate value of the top-left corner point(SBB)
- numberGroup = 1
- $G_{numberGroup} = \{S_{BB1}\}$

For each bounding box S_{BBi} **do**
  For each group G_j **do**
    If S_{BBi} overlaps with at least one bounding box belonging to G_j **then**
      $G_j = G_j + \{S_{BBi}\}$
    End If
  End For
  numberGroup++
  $G_{numberGroup} = \{S_{BBi}\}$
End For

Identifying crowdedness status

The problem we aim to address is detecting crowded groups of people in restricted areas, where the usual number of people should not exceed 4 or 5 individuals. Therefore, after identifying the groups of people, determining the crowdedness status becomes relatively simple by counting the number of people within each group. However, in order to increase processing speed, during the grouping stage, we did not consider the distance from the individuals to the surveillance camera. This may cause some individuals being assigned to the same group even though they do not belong together. To address this limitation, we propose introducing a threshold for the coverage ratio to determine whether or not a group is crowded. This ratio helps mitigate the impact of small bounding boxes within a group that may still be counted as a single individual.

3. RESULTS AND DISCUSSION

3.1. Experimental Environment

To evaluate the performance of the proposed system, we used 200 outdoor images from two datasets, STCrowd and SmartCity. These images have a resolution of 1280 x 720 pixels. The datasets comprise frames extracted from outdoor surveillance cameras at different locations. We specifically selected representative frames for different situations at each location (Figure 3). For each image, we recorded information about the

number of people in the frame (ranging from 2 to 15 individuals), the number of groups (ranging from 1 to 6 groups), and the number of crowded groups (ranging from 0 to 2 groups, but mostly 1 group).

Figure 3. Illustration of some crowded scenarios in the test image set

The system was implemented and tested on the following hardware configuration:

Operating system: Windows 11 Home Edition.

CPU: Intel® Core(TM) i5-1035G1

RAM: 8GB

Programming language: JAVA

3.2. Experimental results

We evaluate the performance of the proposed system based on two criteria: accuracy and execution time. For accuracy, we assess three metrics: the rate of correctly identifying individuals within the image, the rate of correctly identifying groups of people, and the rate of correctly identifying crowded groups. The execution time for the set of 200 images is 28 seconds, averaging 0.14 seconds per image. The average accuracy of the system on the experimental dataset is presented in Table 1.

Table 1. Experimental results on the set of 200 images

	Accuracy	Number of correct instances/50 frames
The number of people in the frame	91,59%	183
The number of groups in the frame	96,33%	193
The number of crowded groups in the frame	98%	196

We can observe that although the phase of detecting human objects in the images has a relatively high average accuracy rate, the system fails to accurately identify and count the number of people in most frames. Through the analysis of the image set, the results show that most of the occluded or small-sized human objects are not recognized by the pre-trained YOLOv3 model we used. Figure 4 illustrates some situations where human objects are not fully detected.

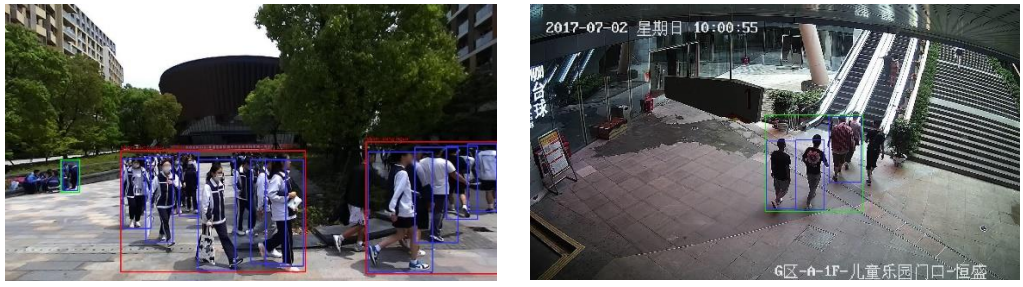

Figure 4. People not detected due to occlusion (left), small size (right)

The proposed system accurately identifies almost the entire number of people groups in the image (average accuracy of 96.33%) and, in particular, the number of crowded people groups (average accuracy of 98%). This meets the requirements of the initial research problem we set out to solve. Figure 5 illustrates cases where crowded people groups are correctly identified by the system (with different scenarios), and Figure 6 demonstrates the crowded people alert function with video footage captured from surveillance cameras at various locations in Trà Vinh City, Trà Vinh Province.

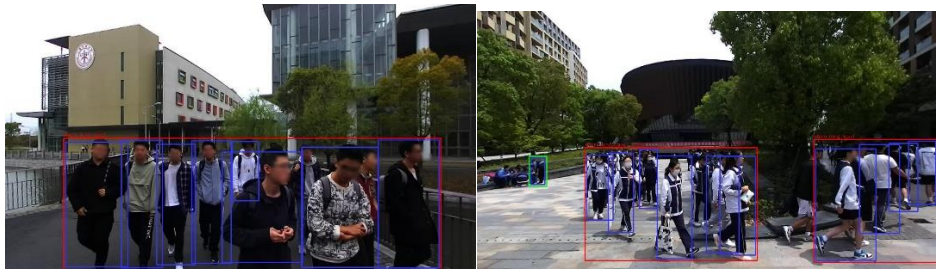

Figure 5. Crowded people groups detected

Figure 6. Surveillance video (left) and crowded people concentration alert (right)

By comparing with the work of Zhang et al. (2018) which proposed the use of SaCNN network, which has shown outstanding performance in pedestrian recognition within frames. However, Zhang et al. (2018)'s method only stops at estimating the number of pedestrians in the image by counting the detected pedestrian heads, and this method does not provide a way to detect crowds for alerting in densely populated areas. In contrast, our proposed solution can detect crowds for alerting through the algorithm presented.

4. CONCLUSION

the detection of crowded groups of people, which can identify gatherings in restricted areas such as government premises or quarantine zones. The proposed system has been tested on real-world data from publicly available standard datasets and real-time data collected from surveillance cameras in various public facilities in Tra Vinh City, Tra Vinh Province, Viet Nam. The results obtained have showed the feasibility of the proposed system. The uniqueness of the approach lies in its ability to construct an algorithm for crowd detection to alert in densely populated areas. This forms a crucial foundation for alerting to overcrowding in restricted or isolated zones, facilitating control over gatherings of people. The solution is highly workable when implemented in practical scenarios. The proposed method can detect groups of people, whereas other works only focus on counting individuals.

However, the system still has some limitations that need to be addressed in the future. These include the inability to fully detect small-sized objects (distant from the camera) or objects that are occluded. Individuals at different distances are sometimes grouped together. The person detection model needs to be retrained with samples of smaller-sized objects, and improvements can be made by considering the size factor of bounding boxes in the grouping stage to overcome these limitations.

In the future, we will compare our method with similar works to show the effectiveness of the proposed approach.

REFERENCES

Ahmad, M., Ahmed, I., Ullah, K., & Ahmad, M. (2019, October). A deep neural network approach for top view people detection and counting. *In 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)* (pp. 1082-1088). IEEE.

Bhangale, U., Patil, S., Vishwanath, V., Thakker, P., Bansode, A., & Navandhar, D. (2020). Near real-time crowd counting using deep learning approach. *Procedia Computer Science*, *171*, 770-779.

Byju, J., Chitra, R., Pranesh, P. E., Pavan, R. S., & Aravinth, J. (2021, March). Pedestrian Detection and Tracking in Challenging Conditions. *In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 399-403). IEEE.

Kannadaguli, P. (2020, November). YOLO v4 based human detection system using aerial thermal imaging for uav based surveillance applications. *In 2020 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 1213-1219). IEEE.

Ubale, P., Surve, A., Srivastava, A., Vishwakarma, R., & Malik, S. (2021). *ML Based Crowd Detection System–A review*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Yizhou, F., Junyan, C., Tianmin, X., Rongfeng, C., Xinyu, L., Yongqing, T., & Xiaochun, L. (2019, December). Application of the ssd algorithm in a people flow monitoring system. *In 2019 15th International Conference on Computational Intelligence and Security (CIS)* (pp. 341-344). IEEE.

Zhang, L., Shi, M., & Chen, Q. (2018, March). Crowd counting via scale-adaptive convolutional neural network. *In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1113-1121). IEEE.