Advancing Crowd Management through Innovative Surveillance using YOLOv8 and ByteTrack

J Cruz Antony[1] , Ch. Leela Sri Chowdary[2], Nanda Prabhu B[3], E Murali [4], Albert Mayan [5]

[1,4,5] Professors, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai.

[2,3]U.G Student, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai

[1]jcruzantony@gmail.com, [2]leelasric@gmail.com, [3]nanda96here@gmail.com, [4]emurali88@gmail.com, [5]albertmayan@gmail.com

2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET) | 979-8-3503-5084-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/WISPNET61464.2024.10533138

Abstract— **Crowd management is a cumbersome task that requires broad analysis and grasp of various constraints. The main hurdles include security issues, unexpected crowd dynamics over which there is typically minimal or no control, limited infrastructure to accommodate. This paper suggests a model that is custom trained to enhance people detection in crowd and monitor crowd density as well as tracking people in a video frame. The paper proposes using YOLOv8 which is known for its capabilities as a fast single shot detector and for being extremely customizable combined with ByteTrack, an advanced model that specializes in tracking people in crowded or congested areas. Together, the models help create a powerful system that can accurately locate and follow people in dense groups, therefore revolutionizes traditional approaches to crowd management. The people counting system also helps with crowd control by indicating how many people are in a certain area. In addition to addressing the issues associated with congested areas, this system provides trustworthy means of tracking people in a variety of scenarios, hence enhancing public safety and security.**

Keywords— ByteTrack, CrowdHuman, Crowd Management, YOLOv8

I. INTRODUCTION

Crowd management presents a significant challenge in urban and large-scale environments. Ensuring the safety and well-being of individuals in crowded environments is becoming increasingly complex as cities grow larger and larger [1]. As a result, we are seeing a surge in innovation within the field of people detection and surveillance, with researchers and technologists developing increasingly robust solutions for the detection and monitoring of individuals in crowded environments, in answer to an increasingly pressing need for such systems.

It is a highly inefficient process; monitoring and surveillance of individuals in such a manner requires a large workforce [2]. It's crucial to be able to do so autonomously. Therefore, it is critical to accomplish this process independently. Identifying humans can be challenging due to uncertainty in mobility. Fortunately, convolutional neural networks (CNNs) have addressed these issues. CNN-based approaches extract powerful features from diverse datasets through stacked convolution and pooling. They have outperformed humans in some areas of image recognition [3, 4].

The challenges of detecting and monitoring people are diverse, as are the proposed solutions. Akhtar and Malhotra [5] conducted a systematic analysis of crowd detection and counting techniques, highlighting challenges such as a lack of standard datasets and balancing speed and accuracy. These challenges underscore the complexity of designing systems that can operate seamlessly in diverse real-world scenarios. Alam et al. [6] introduce a people-tracking system in crowded environments through a mobile application, emphasizing the role of user-centric approaches in enhancing crowd management. However, the effectiveness of such solutions is contingent on factors like GPS signal availability as well as its accuracy, introducing concerns related to privacy and security. Yang et al. [7] delve into the development of an efficient deep neural model for video-based crowd anomaly detection. By incorporating an attention mechanism, their model demonstrates efficacy in identifying anomalies like stampedes or fights. Nevertheless, the study acknowledges the need for extensive labelled video data for training, raising questions about the generalizability of such models to diverse and unforeseen anomalies. This exploration extends beyond the confines of technological solutions.

You Only Look Once (YOLO) is a fully conventional neural network which improves the performance of object detection by directly training on complete images. YOLO predicts multiple bounding boxes and their corresponding class probabilities simultaneously by using a single convolutional network, hence it works extremely fast and can process the images at 45 frames per second [8]. At the forefront of recent advancements is the work of Gao et al. [9], who proposed an enhanced YOLOX for detecting

979-8-3503-5084-5/24/$31.00 ©2024 IEEE

pedestrians in densely populated areas. Leveraging a multi-scale approach, their system showcases a noteworthy improvement in accurately identifying individuals in dense and dynamic environments. Additionally, the study by Pervaiz et al. [10] introduces a Smart Surveillance System that utilizes particle flow and modified self-organizing maps for people counting and tracking, demonstrating the potential for innovative techniques in enhancing surveillance capabilities.

This study suggests using advanced models such as YOLO and ByteTrack for people monitoring and surveillance systems. The proposed system has the potential to detect individuals with higher accuracy in crowded environments. The robust ByteTrack is used for individual tracking, ensuring continuous detection and tracking of individuals, even in densely populated areas [11]. Notably, ByteTrack maintains tracking continuity in the presence of obstacles or partial facial interference by temporarily withholding judgment in a few frames to confirm the individual's status.

While many models use pre-trained networks, we propose a network custom trained on CrowdHuman dataset, which leads to a better performance in this application. The focus of crowd management is a novelty as many existing works focuses only on object detection, pedestrian tracking, and individual tracking. Through this paper we aim to address the complex problem of managing and monitoring crowd with an alert system based on customizable parameters.

II. Overview of existing systems

Zhang and Zhang [12] proposed a system for real-time crowd counting using a multi-scale head-shoulder detector. While the system performs well under certain conditions, it struggles with low-resolution or crowded scenes, handling multiple occlusions, or distinguishing between different types of people.

Xue et al. [13] created and integrated some core modules including sparse reconstruction layers, Multiple Dilated Convolution Branches, Adaptive Receptive Field Weighting, and Compressed Sensing based Output Encoding into an end-to-end trainable network for crowd analysis. The experiments revealed that it is imperative to address the issue of target size variation in crowd analysis.

Sharma et al. [14] conducted a systematic study on several deep learning techniques for crowd detection and people counting. Despite the comprehensive analysis, the study highlights challenges such as the lack of standard datasets, the trade-off between speed and accuracy, and the complexity of handling dynamic scenes.

Maddalena et al. [15] created a method that partitions a target scene's tracking areas using a counting line on an edge. Multiple cameras were used to capture the data, and background subtraction was employed for human identification. Motion prediction was used to track people, and the algorithm did a good job of segmenting and occluding moving and stationary items. However, the system’s limitation lies in the variance of an individual’s posture with multiple orientations, which can lead to inaccuracies in people tracking.

Yadav et al. [16] proposed a real-time crowd-monitoring system using deep learning techniques. The system leverages YOLO v4 for object detection and Deepsort for tracking. Despite its potential in enforcing social distancing measures, the system has limitations. It requires high-quality cameras and high-speed internet and may struggle with occlusion, illumination, or background clutter.

III. Proposal of the current system

A. Models

In this paper, we propose a combination of SOTA vision algorithms and techniques, each of which is chosen in accordance with the difficulties or major blockers that are to be handled by our expected and proposed system. When compared to classical CNNs and other object identification models like Faster R-CNN and SSD, YOLOv8 has an edge over them as it has a much faster inference time. Adding to it, YOLOv8 is also highly customizable/configurable. The hyperparameters which help YOLOv8 learn and generalize better can be tweaked as required [17]. The typical YOLOv8 training procedure includes using a large dataset, tweaking the learning rate, and iterating across multiple epochs.

ByteTrack specializes tracking small and fast-moving objects but moreover ByteTrack has a way to deal with occlusions. Instead of discarding an object when an occlusion occurs, ByteTrack keeps a track on it for several frames and checks if the occlusion is now no longer a blocker and if the object is an already identified entity. Its end-to-end trainable structure enables continual development and enhancement of tracking accuracy, even in heavily crowded environments [18].

Based on recent studies, YOLOv8 has shown significant advancements in object detection, particularly in crowded areas . The combination of YOLOv8 with ByteTrack enhances crowd surveillance and tracking capabilities, proving beneficial for crowd management and public safety. While YOLOv8 excels in fast and customizable detection, ByteTrack specializes in tracking individuals in congested environments.

YOLOv8 excels in the real-time performance compared to other state of the art models like EfficientDet, CenterNet and other older versions of Yolo. YOLOv8 also offers higher accuracy in comparison with RetinaNet and Cascade R-CNN along with ease of implementation. YOLOv8 also offers high customizability compared to other state of the art models.

B. *Training*

In the proposed system, YOLOv8 was trained on the CrowdHuman dataset. This dataset is a rich collection, specifically curated for human detection tasks. It provides a diverse range of crowd scenarios, including various types of occlusions commonly encountered in crowded environments [19].

The original dataset is vast, but for our training, we utilized a subset of approximately 4,000 images. Out of these, 3,486 were used to train our model, 304 to validate the trained model, and 105 to test the model. This decision was influenced by the annotation tool we employed, RoboFlow. Despite its limitation of handling only 5,000 images per account, RoboFlow was our tool of choice due to its ability to automate the annotation process. It allowed us to manually annotate a few images, after which it took over and annotated the rest based on our inputs. This significantly reduced the manual effort required for annotation.

After the annotation process, these images underwent several preprocessing and augmentation steps in order to further enhance the quality of the dataset thereby aiding in improving the performance of the models.

The models were trained on a Kaggle GPU. Due to a 15GiB GPU memory allocation constraint on Kaggle, we could only accommodate up to 16 batches. We experimented with different epochs, batch sizes, and image sizes to find a teeter between the hardware constraints and performance accuracy. The final parameters were set to 25 epochs, a 256-pixel image, and 16 batches.

Looking ahead, with the availability of TPUs, we would be able to raise batch size to 32-64 batches. An important note to make is that 64 is a rather normal batch size which works for most models. The preliminary results indicate that this could significantly improve our model's performance if we can move to a rather standard batch size. We employed the wandb.ai tool to visualise metrics and monitor the evolution of our models. This tool offered us a concise view of our models' performance metrics, allowing us to make more informed judgements regarding future optimisations.

C. *System Design*

The system proposed in the paper is a real-time solution which is capable of monitoring and analyzing crowd activity. The system uses carefully selected and trained computer vision techniques and deep learning models to detect, track, and recognise individuals in a crowded setting. We begin the process by identifying all the people in each frame of a video feed. The video feed is processed into multiple frames and then each frame undergoes selected preprocessing and augmentation techniques, as depicted in Table 1, such as noise removal, image resizing, frame-rate conversion, and de-interlacing to increase video data quality which in turn enables YOLOv8 to produce finer results as YOLOv8 works its single shot detection algorithm on the frames as depicted in Fig 1.

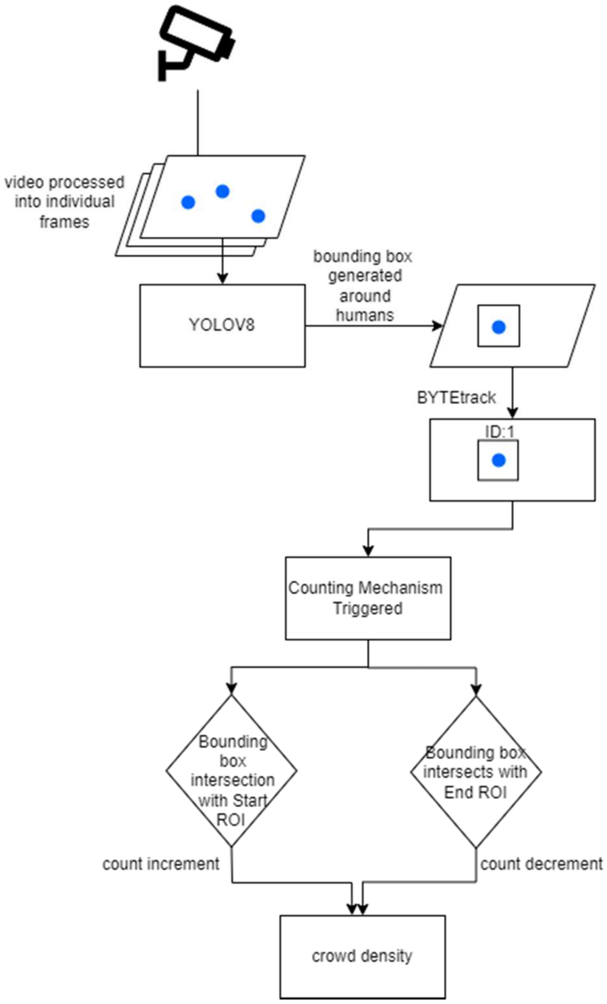
video processed into individual frames
bounding box generated around humans
YOLOV8
BYTEtrack
ID:1
Counting Mechanism Triggered
Bounding box intersection with Start ROI
Bounding box intersects with End ROI
count increment
count decrement
crowd density

Fig. 1. System Architecture for people tracking and crowd density

After YOLOv8 detects an individual, a bounding box is created to depict that entity detected. When an individual crosses a specific region, the head count of the region is increased by 1 as shown in Fig 1. The mechanism behind this involves incrementing the count whenever a bounding box intersects with entry line of the region under surveillance as decrementing whenever the bounding box intersects with the exit area of the region. This helps track the crowd density.

After identifying the people in a frame, it is important that we are able to track their movements across the frames. This could prove a tedious task as there could be temporary occlusions and most models discard the entity as a previously being tracking one hence proving to be inefficient. A workaround for this hurdle is using ByteTrack algorithm and training it on a crowd dataset. ByteTrack specializes tracking small and fast-moving objects but moreover ByteTrack has a way to deal with occlusions. Instead of discarding an object when an occlusion occurs, ByteTrack keeps a track on it for several frame and checks if the occlusion is now no longer a blocker and if the object is an already identified entity. The ByteTrack algorithm detects individuals and tracks their movements across frames. ByteTrack assigns a unique ID to each individual, allowing the system to follow their movements even when they move out of the frame. Its end-to-end trainable structure enables continual development and enhancement of tracking accuracy, even in heavily crowded environments .

An alert system/mechanism is integrated to the system as shown in Fig 2. that is set to be triggered when certain conditions are met. In the context of our application, the pre-conditions would be when crowd density in an area surpasses a pre-defined limit as mentioned in Fig 2. This feature improves the system's utility in public safety applications, since the system works with video data from CCTV cameras and using deep learning models such as YOLOv8 and ByteTrack, the system would require a computing environment, including a powerful GPU, for training and inference. It would additionally require enough storage to hold the video data and trained models.

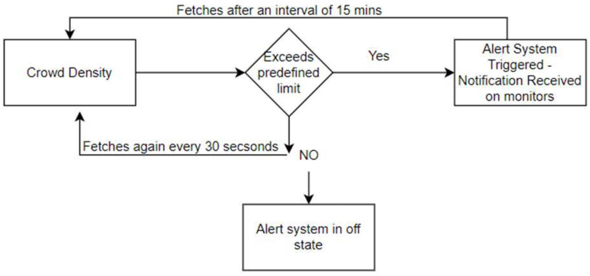
Fetches after an interval of 15 mins
Crowd Density
Exceeds predefined limit
Yes
Alert System Triggered - Notification Received on monitors
Fetches again every 30 secsonds
NO
Alert system in off state

Fig. 2. Alert System

D. *Evaluation*

The performance of the system can be evaluated using a variety of metrics and scenarios. Metrics such as precision, recall, confidence and F1-score were employed. The system was tested in a variety of real-world scenarios to guarantee that it works well under varied conditions. This involved testing it in different contexts (for example, crowded vs. uncrowded regions) and with different types of video data like good and low-quality feeds.

IV. Results and discussions

A. *Data Collection Results*

The data was collected from the CrowdHuman dataset, which is a large and rich-annotated people detection dataset. We utilized a subset of approximately 4,000 images. Out of these, 3,486 were used to train our model, 304 to validate the trained model, and 105 to test the model. The dataset is exhaustively annotated and contains diverse scenes.

B. *Data Preparation*

The data used in training underwent extensive preprocessing and augmentation steps as mentioned in Table 1. These methods help to generate more consistent images and improve model performance. They help the models become invariant to multiple transformations, improving their capacity to generalize from training data to new examples. And these measures helps us to reduce overfitting by, increasing the size of training dataset.

TABLE I

Preprocessing and Augmentation Steps

STEP	VALUE
Auto-Orient	landscape
Resize	480x480 pixels
Flip	Horizontal
Rotation Angle	-10° a+10°
Grayscale	Applied to 20% of images
Blurring	1.5 px
Noise Addition	Up to 3% of pixels

C. *Evaluation*

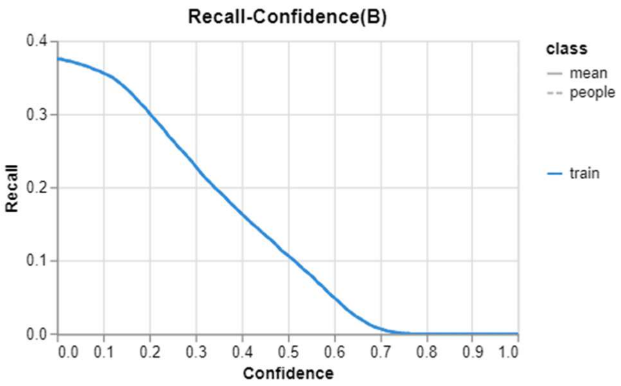

Fig. 3.Recall Graph

A recall-confidence graph is a common way to visualize how the recall of a model changes as we change the confidence threshold (the value that the estimate needs to exceed in order for the model to predict positive). Essentially, if we slid this threshold left and right across the data, we will see recall of the model change. From Fig 3, we see that as we increase confidence, we decrease recall, which means this model, at lower confidence thresholds, is identifying a larger fraction of relevant instances.

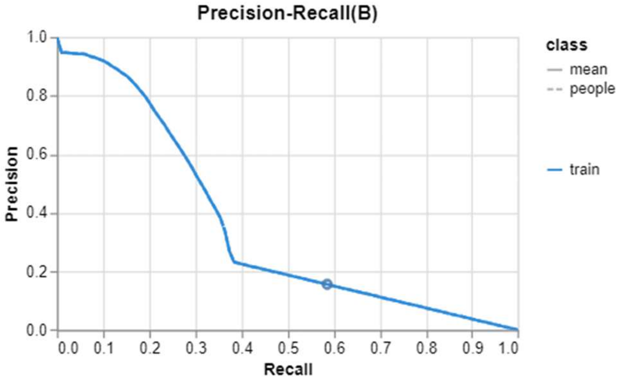

Fig. 4. Precision-Recall Graph

Recall and precision are two of the most important measures for evaluating the effectiveness of a classification model, particularly in regard to binary classification scenarios. The precision-recall curve is often used to visualize them, as this one is doing. This graph shows the trade-off between recall and precision at different classification thresholds. It's doing this to gauge the quality of the model, where in higher values are indicative of superior performance. The graph depicted in Fig 4 also shows that as recall begins to rise, precision will begin to drop. This means that the model is doing a better job of identifying a larger proportion of the relevant instances at lower recall thresholds.

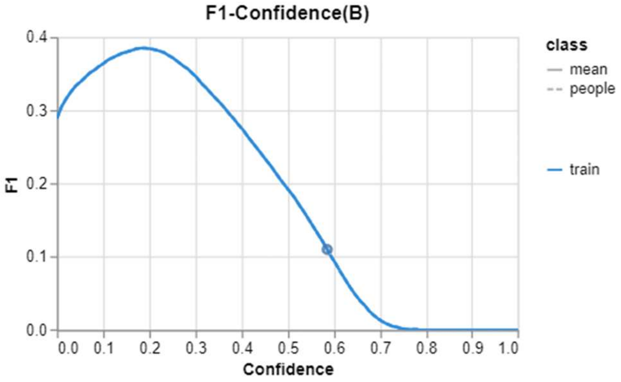

Fig. 5. F1 Graph

The F1 score is the harmonic mean of precision and recall taking both metrics into account. The F1 score is the best if there is some sort of balance between precision and recall. From the graph, we can observe that as the model's confidence is increasing, the recall is decreasing. As we can see in Fig 5, when the model identifies 9 positive instances, those instances were found at a confidence threshold of about 0.7.

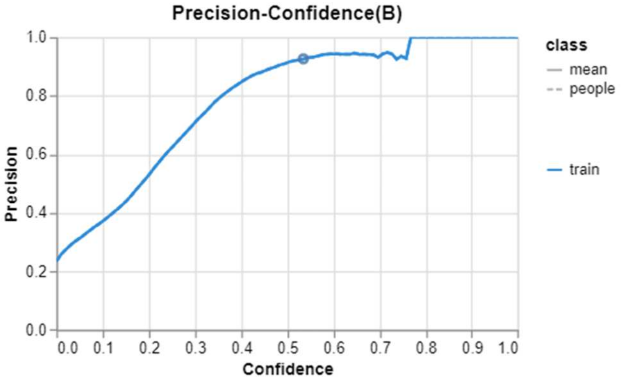

Fig. 6. Precision-Confidence Graph

From Fig 6, we can observe that as confidence threshold increased, the precision got better. This it typically true as the precision in very low confidence predictions tend to be low, as True positives remain constant at a certain point, but False positives reduce as confidence is increased. This graph shows the model can identify 9 relevant instances with a precision of around 90%, when the confidence threshold is about 80%.

D. *Discussions*

Below depicted in Table 2 is a confusion matrix. A confusion matrix is used as a metric to evaluate the performance of a model by drawing a matrix against predicted and actual values.

Confusion Matrix	Values
True Positive (TP)	10,905
False Negative (FN)	7,712
False Positive (FP)	386
True Negative (TN)	40,105

TABLE 2. CONFUSION MATRIX TABLE

Below provided in Table 3 are the calculations for the performance metrics such as Recall, Precision, F1-Score, and accuracy.

Measure	Value	Derivations
Precision	0.9658	TP / (TP + FP)
Accuracy	0.863	(TP + TN) / (P + N)
F1 Score	0.7292	2TP / (2TP + FP + FN)
Recall	0.5858	TP / (TP + FN)

TABLE 3. PERFORMANCE METRICS

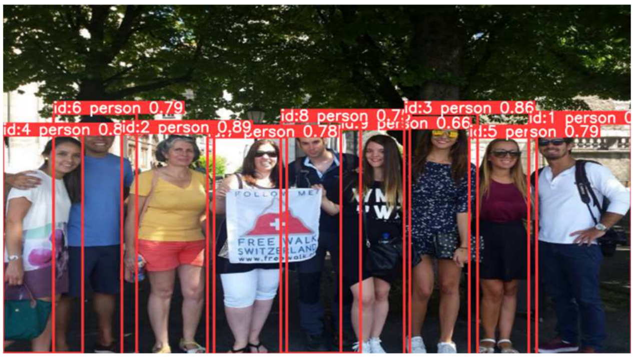

Fig. 7. Model Output

V. Conclusion

The proposed system, here combines both YOLOv8 and ByteTrack, effectively to tackle real-time crowd monitoring challenges. The system performs well in identifying and tracking individuals across different crowd scenarios, as indicated by comprehensive evaluation metrics such as recall-confidence, precision-recall, and 86.6% accuracy from the confusion matrix. While proficient in real-time operations and occlusion management, limitations include persistent occlusion scenarios and reliance on high-performance computing resources. Future improvements aim to address limitations and increase the system's capabilities. This involves dealing training the model with an increased batch size as well as integrating features such as fall detection and stampede detection to enhance crowd management solutions. Making the transition to Tensor Processing Units (TPUs) could improve speed and accuracy, leading to efficient model training and inference.

References

[1] Weaver, K. & Liu-Lastres, B. (2021). Applying Crowd Risk Mitigation Technologies in Urban Sport Events: A Case Analysis of the Collegiate Football Event in Indianapolis, IN., *Events and Tourism Review*, 4(2), 14-27.

[2] Ravipati, A., Kondamuri, R. K., Posonia, M., & Mayan, A. (2023, April). Vision Based Detection and Analysis of Human Activities. In 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1542-1547). IEEE.

[3] Zhou, X., Yi, J., Xie, G., Jia, Y., Xu, G., & Sun, M. (2022). Human detection algorithm based on improved YOLO v4. Information Technology and Control, 51(3), 485-498.

[4] Saravanan, M. S., Antony, J. C., Kumar, V. P., Veeramanickam, M. R. M., and Upadhyaya, M. (2022),"A new framework to classify the cancerous and non-cancerous pap smear images using filtering techniques to improve accuracy",International Conference on Artificial Intelligence Trends and Pattern Recognition pp. 1-6. IEEE.

[5] Akhtar, R., & Malhotra, D. (2022). Intelligent Techniques for Crowd Detection and People Counting—A Systematic Study. In Computer Vision and Robotics: Proceedings of CVR 2021 (pp. 119-130). Singapore: Springer Singapore.

[6] Alam, T., Hadi, A. A., & Najam, R. Q. S. (2022). Designing and implementing the people tracking system in the crowded environment using mobile application for smart cities. International Journal of System Assurance Engineering and Management, 13(1), 11-33.

[7] Yang, M., Tian, S., Rao, A. S., Rajasegarar, S., Palaniswami, M., & Zhou, Z. (2023). An efficient deep neural model for detecting crowd anomalies in videos. Applied Intelligence, 53(12), 15695-15710.

[8] 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).

[9] Gao, F., Cai, C., Jia, R., & Hu, X. (2023). Improved YOLOX for pedestrian detection in crowded scenes. Journal of Real-Time Image Processing, 20(2), 24.

[10] Pervaiz, M., Ghadi, Y. Y., Gochoo, M., Jalal, A., Kamal, S., & Kim, D. S. (2021). A smart surveillance system for people counting and tracking using particle flow and modified SOM. Sustainability, 13(10), 5367.

[11] Cheng, P., Xiong, Z., Bao, Y., Zhuang, P., Zhang, Y., Blasch, E., & Chen, G. (2023). A Deep Learning-Enhanced Multi-Modal Sensing Platform for Robust Human Object Detection and Tracking in Challenging Environments. Electronics, 12(16), 3423.

[12] Zhang, X., & Zhang, L. (2014). Real time crowd counting with human detection and human tracking. In Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part III 21 (pp. 1-8). Springer International Publishing.

[13] Sudhakar.M, Mayan J.A., Srinivasan.N, "Intelligent data prediction system using data mining and neural networks", Proceedings of the International Conference on Soft Computing Systems, Advances in Intelligent Systems and Computing 398.

[14] Xue, Y., Liu, S., Li, Y., & Qian, X. (2020). Crowd Scene Analysis by Output Encoding. arXiv preprint arXiv:2001.09556.

[15] Sharma, V., Gupta, M., Pandey, A. K., Mishra, D., & Kumar, A. (2022). A review of deep learning-based human activity recognition on benchmark video datasets. Applied Artificial Intelligence, 36(1), 2093705.

[16] Maddalena, L., Petrosino, A., & Russo, F. (2014). People counting by learning their appearance in a multi-view camera environment. Pattern Recognition Letters, 36, 125-134.

[17] Yadav, S., Gulia, P., Gill, N. S., & Chatterjee, J. M. (2022). A real-time crowd monitoring and management system for social distance classification and healthcare using deep learning. Journal of Healthcare Engineering, 2022. N. Cahyadi and B. Rahardjo, "Literature Review of People Counting," in 2021.

[18] Albert Mayan J, Karthikeyan S, Nikhil Chandak, Bharat Mundhra and Padmavathy J, "Facial attendance system technology using Microsoft Cognitive Services" ,International Journal of Engineering Systems Modelling and Simulation, Vol. 12, Nos. 2/3, pp. 180-187 ,2021

[19] Sohan, M., Sai Ram, T., Reddy, R., & Venkata, C. (2024). A Review on YOLOv8 and Its Advancements. In International Conference on Data Intelligence and Cognitive Informatics (pp. 529-545). Springer, Singapore.

[20] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., & Sun, J. (2018). CrowdHuman: A Benchmark for Detecting Human in a CrowdarXiv preprint arXiv:1805.00123.

[21] Murali E,Anouncia S.M(2022),"A Survey on Computational Aptitudes towards Precision Agriculture using Data Mining",International Conference on Smart Electronics and Communication, pp. 952–956.

[22] Erfani, Seyed Mohammad Hassan. "Developing a Vision-Based Framework for Measuring and Monitoring Water Resource Systems Using Computer Vision and Deep Learning Techniques", University of South Carolina, 2023

[23] Divya Seshadri M, Indhuja U S, Amutha R. "Pedestrian Detection in Extreme Weather", 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021