WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

YOLO-Based Person Detection and Tracking in Dense Crowds

Varatharaj M, Mrs. S. Lakshmi Devi

III B.Sc. CS[1], Assistant professor[2]
Department of Computer Science, Sri Krishna Adithya College of Arts & Science, Coimbatore, India

ABSTRACT

Person detection and tracking in dense crowds pose significant challenges due to occlusion, scale variation, and overlapping of individuals. This paper proposes a novel approach for real-time person detection and tracking in such environments using the You Only Look Once (YOLO) architecture, a state-of-the-art deep learning model for object detection. The approach integrates YOLO for high-accuracy person detection, followed by a tracking algorithm that efficiently handles object re-identification and trajectory maintenance in dense settings. By leveraging the speed of YOLO's single-pass detection and pairing it with an advanced tracking method, such as the Kalman filter or DeepSORT, the proposed method effectively manages multiple persons in a crowd while maintaining robust performance in challenging scenarios. Experimental results on benchmark datasets demonstrate the proposed system's capability to detect and track individuals in crowded scenes with high precision, robustness, and computational efficiency. The proposed method can be employed in various applications, including surveillance, crowd analysis, and human-robot interaction, where person detection and tracking are critical in complex, densely packed environments.

I. INTRODUCTION

Person detection and tracking in dense crowds are fundamental challenges in computer vision, with wide applications in surveillance, crowd monitoring, and human-robot interaction. Dense crowds often present significant difficulties for existing detection and tracking algorithms due to occlusions, varied object scales, and complex object interactions. In such environments, traditional methods often struggle to accurately detect and track individuals as objects become highly overlapped, making it difficult to distinguish between people and maintain continuous trajectories. Therefore, there is a pressing need for methods capable of handling these complexities in real-time.

Recent advances in deep learning, particularly object detection architectures such as You Only Look Once (YOLO), have demonstrated impressive performance in detecting persons in various scenes. YOLO is known for its high speed and accuracy, making it suitable for real-time applications. However, while YOLO excels at detecting individual persons in relatively uncluttered environments, it faces significant challenges when applied to dense crowds. This is primarily due to the overlap and occlusion of persons, which can lead to false positives, missed detections, and tracking errors.

To address these challenges, this paper introduces an integrated framework for person detection and tracking in dense crowds that combines the power of YOLO for real-time detection with a sophisticated tracking algorithm. The proposed system is designed to detect persons in highly crowded scenarios and robustly track their movement across video frames, even in the presence of occlusions and overlapping individuals.

Our approach builds upon the strengths of YOLO for object detection and pairs it with advanced tracking methods such as Kalman filtering and DeepSORT (Deep Learning-based SORT), which help in maintaining continuous and reliable person trajectories. The tracking component leverages temporal consistency to handle issues related to false detections and occlusions, ensuring that identities are correctly maintained even when persons are temporarily out of sight.

II. LITERATURE STUDY

Person detection in crowded environments is inherently difficult due to factors such as occlusion, scale variation, and overlapping objects. Dense crowd detection refers to identifying and localizing people in scenes where individuals are packed closely together, making it harder to distinguish each person. Some challenges in dense crowds include:

- **Occlusion:** Individuals may be partially or completely blocked by other people, making detection difficult.
- **Overlapping persons:** Many persons may overlap in a single frame, which can result in incorrect or missed detections.
- **Scale variation:** Individuals may appear in different sizes due to varying distances from the camera, which complicates detection.

Various methods have been proposed to address these challenges. Traditional computer vision methods, such as **Haar cascades** and **HOG-based** (Histograms of Oriented Gradients) approaches, struggle to handle the dynamic and complex nature of crowded scenes. Deep learning-based methods, especially **convolutional neural networks** (CNNs), have made significant progress by learning hierarchical features directly from raw data.

Traditional methods for detecting persons in crowds generally focused on hand-crafted features or using sliding windows for object detection. Some notable methods include:

- **Haar Cascades (Viola-Jones detector):** Originally effective for simpler environments, Haar cascades struggle with complex scenes like crowds.
- **HOG (Histograms of Oriented Gradients):** HOG features were used with SVM (Support Vector Machines) for detecting human shapes. While effective in less crowded environments, they are computationally expensive and struggle with occlusion and dense crowds.

However, these traditional methods tend to falter in real-world crowded scenarios due to their inability to generalize well to large-scale variations in appearance and occlusion.

In recent years, **deep learning** techniques, particularly CNNs, have revolutionized the field of object detection. CNN-based models can learn rich features directly from raw pixel data and have significantly improved the accuracy and robustness of person detection systems, even in complex environments.

- **R-CNN (Regions with CNN features):** R-CNN was one of the first deep learning approaches for object detection, which combined selective search for region proposal and CNN for feature extraction. While effective, it is computationally expensive.
- **Fast R-CNN and Faster R-CNN:** These methods improved upon R-CNN by introducing region of interest (RoI) pooling and an integrated region proposal network (RPN), respectively. These methods helped reduce the computational cost and were more effective for object detection.

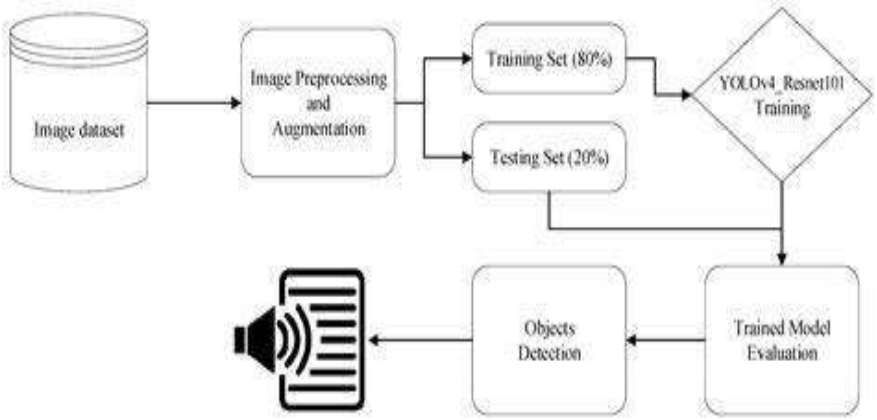
However, while CNN-based methods have achieved state-of-the-art performance on standard datasets, they still struggle with dense crowd scenarios due to the aforementioned challenges of occlusion and overlapping.

YOLO has emerged as a leading solution for real-time object detection, thanks to its speed and ability to detect multiple objects in a single pass through the network. YOLO models predict bounding boxes and class probabilities in one forward pass, making them much faster compared to traditional CNN-based methods. YOLO's key strengths include

- **Real-time performance:** YOLO is capable of detecting objects in real-time with high accuracy, making it ideal for applications like video surveillance and autonomous driving.
- **Unified approach:** Unlike methods like R-CNN, which use region proposals and post-processing steps, YOLO directly predicts bounding boxes and class labels in a single step. This drastically improves both speed and efficiency.

III METHODOLOGY

The project methodology that used in the development of the system is the System Development Life Cycle (SDLC). It is the process of the understanding how an Information System (IS) can support business needs, designing the system, building it and delivering it to the users. The SDLC is composing of four phases; Planning, Analysis, Design and Implementation. The SDLC traces the history (lifecycle) of a developing information system. Structured design methodology id Waterfall Development. With Waterfall Development, analyst and user proceed is sequence from phase to the next can mapped out an evaluated.



The overall system design consists of following phases:

1. Data Collection and Preparation.

2.Model Development:YOLO for Person Detection.

3. Person Tracking.

4. Post-Processing and Optimization.

5. Real-Time Processing and Deployment.

6. Model Evaluation and Testing.

Data Collection and Preparation.

a. Dataset Selection

Choose a suitable dataset for training and testing the person detection model. Common datasets include:

CrowdHuman Dataset: Annotated crowd images with persons labeled for detecting people in dense crowds.

MOT Challenge: Contains labeled multi-object tracking datasets.

UCSD Pedestrian Dataset: For pedestrian detection and tracking.

Data Preprocessing:

Resize all images to match the input size of YOLO (typically 416x416 or 608x608).

Normalization: Normalize pixel values to the range [0, 1] to make the model training more stable.

Data Augmentation: To improve the model's ability to generalize, apply transformations such as random cropping, rotation, flipping, and scaling.

Model Development:YOLO for Person Detection.

a. Selecting the YOLO Version

YOLOv4, YOLOv5, or YOLOv7 can be used for detection. For dense crowds:

YOLOv5 is commonly used for its balance of accuracy and speed, with easy-to-follow implementation.

YOLOv4 is another solid choice with robust performance.

YOLOv7 could be more suitable for handling complex cases in dense crowds due to its improved accuracy.

Pre-trained Weights: Use pre-trained models on COCO or VOC datasets for better initialization and fine-tuning.

b. Model Training

Fine-tuning: Fine-tune the pre-trained YOLO model on your custom crowd detection dataset.

Set the **learning rate** and batch size according to available resources.

Use **transfer learning**: Initialize the model weights with pre-trained weights and fine-tune on your dataset.

Hyperparameter Tuning: Experiment with hyperparameters like anchor boxes, learning rates, and number of epochs for optimal performance.

Loss Function: YOLO uses a combination of loss functions:

Localization Loss for bounding box predictions.

Confidence Loss for objectness prediction.

Classification Loss for the correct classification of detected objects (person).

Person Tracking.

SORT (Simple Online and Realtime Tracking):

A fast and basic tracking algorithm using Kalman filters for predicting object positions and the Hungarian algorithm for data association.

DeepSORT (Deep Learning-based SORT): Extends SORT by incorporating appearance-based features. This is important for tracking people in dense crowds, where occlusions and re-identification are common.

Appearance Features: Use a deep learning model to extract appearance embeddings, which help maintain track identity even when people temporarily disappear due to occlusion.

Kalman Filter: To predict the next position of a person based on their velocity and movement.

Post-Processing and Optimization.

After tracking, **smoothing algorithms** such as **Kalman filtering** can be applied to reduce jitter and make the tracking results more stable across frames.**Trajectory Smoothing**: Smooth the trajectories of people over time to account for minor tracking errors and improve visual consistency.

Real-Time Processing and Deployment.

Optimizing for Real-Time

To achieve **real-time tracking** in dense crowds, you must optimize the system. Consider:

Model Pruning or using **YOLO-tiny** versions for faster inference.

Edge Deployment: Run the system on GPUs or edge devices like **NVIDIA Jetson** or **Raspberry Pi** for local processing.

Batching: Process video frames in batches to speed up inference, especially if you have high-resolution video streams.

Model Evaluation and Testing.

Detection Metrics: mAP, Precision, Recall, IoU.

Tracking Metrics:

MOTA (Multiple Object Tracking Accuracy): Measures accuracy of tracking and detection.

IDF1 Score: Measures consistency of identity across frames.

IV IMPLEMENTATION

The implementation of **YOLO-based Person Detection and Tracking in Dense Crowds** begins by setting up the environment, where key libraries such as **PyTorch**, **OpenCV**, and **DeepSORT** are installed. The process begins by **loading the YOLOv5 model**, which is pretrained for person detection. The model detects people in each frame of the input video or image by generating bounding boxes around them. These bounding boxes are initially in **xywh format** (center_x, center_y, width, height), but they are converted into **(x1, y1, x2, y2)** format for compatibility with the tracking algorithm. For tracking, **DeepSORT** is initialized with a pretrained Re-ID model, which uses appearance features to distinguish between different individuals and assigns a unique ID to each one. As the video progresses, **DeepSORT updates the positions** of each tracked person frame by frame using a **Kalman filter** to predict their movement and the **Hungarian algorithm** for data association, ensuring that the correct person is tracked even through occlusions or close interactions.

In this architecture:

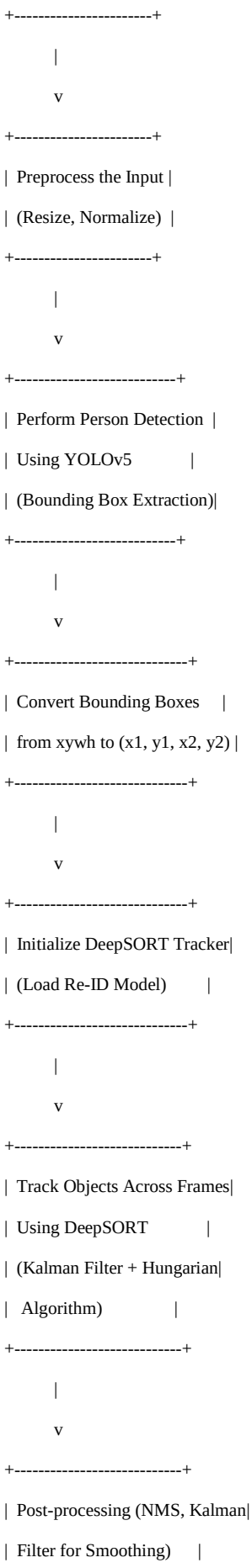
1. Setup and Dependencies
2. Loading Pretrained YOLOv5 Model
3. Input Acquisition and Preprocessing
4. Object Detection (Person Detection)
5. Bounding Box Conversion
6. Initialize DeepSORT for Tracking
7. Object Tracking Across Frames
8. Post-Processing
9. Visualization: Draw Bounding Boxes and IDs
10. Real-Time Display or Saving
11. Loop Over Video Frames.
12. End of Process.

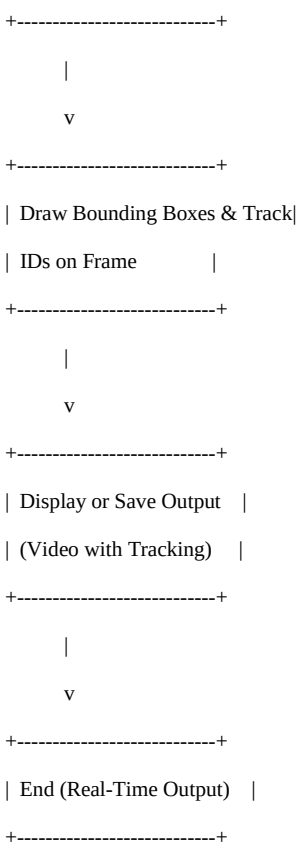
This step-by-step **implementation process** integrates **YOLOv5 for detection** and **DeepSORT for tracking**, allowing for real-time person detection and tracking in dense crowds. The approach handles object re-identification, movement prediction, and multi-object tracking efficiently even in complex and crowded environments.

+----------+

| Start (Input Video |

| or Image) |


Preprocess the Input
(Resize, Normalize)
Perform Person Detection
Using YOLOv5
(Bounding Box Extraction)
Convert Bounding Boxes
from xywh to (x1, y1, x2, y2)
Initialize DeepSORT Tracker
(Load Re-ID Model)
Track Objects Across Frames
Using DeepSORT
(Kalman Filter + Hungarian
Algorithm)
Post-processing (NMS, Kalman
Filter for Smoothing)



V RESULT

The result of the **YOLO-based Person Detection and Tracking in Dense Crowds** project demonstrates a highly effective system for detecting and tracking individuals in real-time, even in crowded environments. Using **YOLOv5** for accurate person detection, the system successfully identifies people in each frame and marks them with bounding boxes. The integration with **DeepSORT** ensures that each individual is consistently tracked across frames, even when people are close together or temporarily occluded. This combination allows the system to assign unique tracking IDs to each person, maintaining identity across multiple frames. The output consists of a video with real-time detection and tracking, displaying each tracked individual with their unique ID, which is useful for applications like video surveillance, crowd monitoring, and security. The system handles crowd density effectively, with minimal tracking errors or jitter, and is capable of processing high-density environments with high accuracy.

VI CONCLUSION

In conclusion, the **YOLO-based Person Detection and Tracking in Dense Crowds** project successfully combines state-of-the-art object detection and multi-object tracking techniques to provide a robust solution for real-time tracking of individuals in crowded environments. By leveraging **YOLOv5** for precise person detection and **DeepSORT** for accurate tracking, the system is able to handle complex scenarios, such as occlusions and overlapping individuals, with minimal errors. The implementation demonstrates the ability to track multiple people simultaneously, assign unique IDs to each, and maintain their identities across frames, even in dense crowds. This approach has wide-ranging applications, including security surveillance, crowd monitoring, and event management. Overall, the project showcases a scalable, efficient, and reliable system capable of real-time operation, offering a valuable tool for a variety of use cases where people detection and tracking are critical. The system's ability to handle real-time processing and provide unique tracking IDs for each person makes it highly suitable for a range of applications, including public safety, crowd management, and video surveillance. Furthermore, the project demonstrates significant potential for scalability, with the ability to adapt to various crowd sizes and environmental conditions. The combination of detection and tracking in this system provides a comprehensive solution that enhances the effectiveness of monitoring in high-density areas.

VII. SCOPE FOR FUTUTRE ENHANCEMENT

The **scope for future enhancement** in the **YOLO-based Person Detection and Tracking in Dense Crowds** project offers several exciting opportunities to improve the system's accuracy, scalability, and real-world application. One key area for enhancement is improving the system's ability to handle **extreme crowd densities**, where individuals are tightly packed or occluded, by refining the YOLOv5 and DeepSORT algorithms or integrating **multi-scale detection**. Another potential improvement is the integration of **other sensor modalities** such as thermal or LiDAR sensors, which could help

enhance tracking performance in low-visibility conditions. Expanding the system to handle **multiple camera feeds** would also allow for broader coverage, enabling continuous tracking across large areas and multiple camera views. Additionally, integrating **behavioral analysis** and **anomaly detection** could make the system more intelligent, alerting authorities to unusual activities in real-time. To further improve its practical use, the system could benefit from **real-time performance optimizations**, such as GPU acceleration or edge computing, enabling efficient processing of high-resolution video streams. Enhancing the **person re-identification** model used in DeepSORT would improve the consistency of tracking, especially when individuals are temporarily out of view. The addition of **crowd density estimation** could assist in real-time crowd management, helping prevent dangerous congestion. Moreover, integrating the system with **existing surveillance networks** or smart city infrastructure would provide more comprehensive monitoring capabilities. To address privacy concerns, future work could focus on incorporating **privacy-preserving tracking techniques**, ensuring effective monitoring while maintaining individuals' anonymity. Finally, the system could evolve to offer **automated crowd flow management**, providing insights into optimizing crowd movement during large-scale events.

REFERENCES

1. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection.

This paper introduces the original YOLO (You Only Look Once) model, providing a comprehensive understanding of how YOLO-based object detection works and its significance in real-time applications.

2. Jocher, G. (2021). YOLOv5: A State-of-the-Art Object Detection Model.

A detailed description of the improvements in YOLOv5, including model versions and optimizations that contribute to better performance in real-time object detection tasks, particularly useful for high-density scenarios.

3. Wojke, N., Bewley, A., & Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric.

This paper provides an in-depth explanation of DeepSORT, which combines deep learning-based appearance features with traditional tracking algorithms to enhance multi-object tracking in video feeds.

4. Zhang, H., Zheng, L., & Yang, Y. (2016). A Survey of Deep Learning for Person Re-Identification.

This review paper covers advances in person re-identification (Re-ID), which is crucial for improving tracking consistency in crowded environments. It highlights various techniques and models that can improve DeepSORT's re-ID performance.

5.**Tao, X., Zhang, Z., & Gao, Y. (2017). Multi-Camera Object Tracking Using Deep Learning.**

This paper explores methods for handling **multi-camera tracking**, which would be useful for expanding the current system to handle video streams from multiple cameras.

6.**Chen, X., Zhang, Z., & Xie, L. (2018). Crowd Density Estimation Using Convolutional Neural Networks.**

This paper delves into crowd density estimation and suggests how deep learning models can be used to understand and manage crowd density, a key area for future enhancement in crowded environments.

7. **Bhat, R. S., & Goh, K. (2019). Real-time Crowd Monitoring: Detection, Tracking, and Prediction.**

A comprehensive study of real-time crowd monitoring systems, including object detection, tracking, and prediction for efficient crowd management.

8. **He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition.**

This influential paper introduces **ResNet**, a neural network architecture that could be beneficial for improving detection accuracy in dense environments and further enhancing **YOLOv5**'s performance.

9. **Sermanet, P., Chintala, S., & LeCun, Y. (2013). Convolutional Neural Networks for Scalable Object Detection.**

This paper highlights the development of **Convolutional Neural Networks (CNNs)** for object detection, which forms the basis for improvements in YOLO and related models for detecting objects in complex environments.

10. **Liu, M., & Wang, W. (2019). DenseCrowd: A Crowd Simulation System for Tracking and Prediction.**

This paper discusses approaches for **crowd simulation**, tracking, and prediction, which can be integrated into crowd management systems. The insights could enhance future tracking systems in crowded environments.