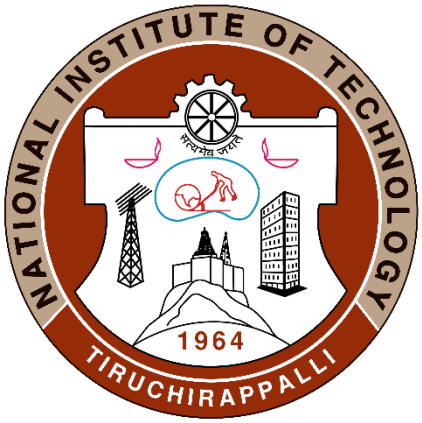
NATIONAL INSTITUTE OF TECHNOLOGY

THIRUCHIRAPALLI



INTERNSHIP REPORT

ON

HUMAN DETECTION AND TRACKING IN A SURVEILLANCE CAMERA USING DEEP LEARNING FEATURES

Name: Priyanka Kutiyare

Roll Number: 107122083

Mentor: Dr. M. Sridevi

Duration: 12.05.2025-12.07.2025

College Name: National Institute of Technology, Tiruchirappalli

Department /Year: EEE 4th year

ABSTRACT

In today's world, security and safety are major concerns, especially in public places like railway stations, airports, malls, and streets. Surveillance cameras are used everywhere to monitor people's activities. However, watching and analyzing long hours of video manually is very difficult and time-consuming. To solve this, we propose an intelligent system that can automatically detect and track a specific person in surveillance videos. Our proposed system uses YOLOv8 to detect humans in each frame of the video. Once a person is detected, we use DeepSORT, a popular tracking algorithm, to follow the same person as they move across different frames. A major problem in real surveillance videos is poor lighting, especially at night or in dark indoor areas. To handle this, we apply a Darknet-based image enhancement technique that improves the visibility of people in low-light conditions before running detection. Additionally, we include a person re-identification (re-ID) feature, which helps the system remember the target person, even if they disappear from the camera for some time and come back, or if they appear in a different video. This makes the system more reliable and helps track the same person across different locations and videos. We tested our proposed method on multiple surveillance videos and found that it works well in real-world situations, including crowded scenes and poor lighting. It can track the same person accurately and quickly, making it suitable for real-time applications like smart surveillance, law enforcement, and public safety monitoring.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the faculty and staff of the National Institute of Technology, Tiruchirappalli (NIT Trichy), for their unwavering support, encouragement, and guidance throughout my internship period. Their collective knowledge and encouragement played a crucial role in shaping the progress and successful completion of my project on emotion recognition from emojis and textual data.

In particular, I am profoundly thankful to Dr. M. Sridevi, whose expert supervision, insightful feedback, and continuous encouragement over the course of this five-week internship greatly enriched my understanding and inspired the direction of this work. Her dedication and mentorship were instrumental in helping me develop, refine, and complete this project.

I also extend sincere thanks to all faculty members, coordinators, and administrative staff involved in the NITT Internship Program, whose efforts in organizing a productive and enriching research environment are truly appreciated.

Special appreciation is extended to the Department of Computer Science and Engineering at NIT Trichy for providing the necessary computational resources and academic insights that shaped this work. I would also like to acknowledge my peers and fellow interns for their constructive discussions, support, and shared enthusiasm throughout this internship journey.

Finally, I would like to thank my family and friends for their constant motivation and emotional support during this period of intensive learning and development. This experience has been both intellectually rewarding and personally fulfilling, and I am grateful for the opportunity to have worked in such a stimulating academic atmosphere.

TABLE OF CONTENTS

1. INTRODUCTION

In recent years, the demand for intelligent surveillance systems has grown rapidly due to increasing concerns about public safety and security. A key component of such systems is the ability to accurately detect and track humans in real-time from video footage. Traditional surveillance systems often rely on manual monitoring, which is time-consuming, prone to human error, and inefficient for large-scale deployment.

Human detection is the first and most important step in any surveillance system. If a person is not correctly detected in a video frame, the tracking and identification steps that follow will not work properly. Detecting people in videos can be difficult because they may appear in different poses, wear different clothes, or be partially hidden behind objects or other people. The lighting may also be poor, especially at night or in indoor settings.

Recent deep learning models have made human detection much better. One popular model is YOLO (You Only Look Once). It is fast and can detect objects in real-time by analyzing the whole image at once, instead of checking small parts one by one. The newest version, YOLOv8, brings more improvements. It uses a simpler and smarter network design, better loss functions, and no longer needs fixed anchor boxes. These upgrades help it detect small or overlapping people more accurately and faster than older models.

To address these limitations, our proposed paper presents a deep learning-based approach that combines the strengths of YOLOv8, a state-of-the-art object detection model, with DeepSORT, a widely adopted multi-object tracking algorithm. YOLOv8 is known for its real-time performance and high detection accuracy, achieved through its advanced backbone architecture and training strategies [1], while DeepSORT provides effective tracking by associating unique identities to detected individuals across frames using both motion and appearance cues [2]. The integration of YOLOv8 and DeepSORT enables robust human detection and tracking even in challenging conditions such as occlusion, low frame rate, and varying lighting. Furthermore, the system incorporates a low-light image enhancement module in the preprocessing stage to address

visibility issues in nighttime or poorly illuminated scenes, which significantly boosts detection performance in such adverse conditions [3]. This enhancement is particularly useful in applications like smart city surveillance, hospital monitoring, and autonomous navigation, where lighting cannot be guaranteed to be optimal.

A distinctive aspect of this system is its support for query-based re-identification**,** where a specific target individual can be tracked across a sequence of frames using a cosine similarity threshold. This capability is particularly important in multi-camera environments or for scenarios where a subject exits and re-enters the field of view, a common challenge in crowded or complex surveillance settings [4]. Moreover, the proposed framework has been designed with scalability in mind, offering compatibility with edge computing hardware for real-time deployment. By reducing latency and computational requirements through a single unified detection-tracking pipeline, the system is well-suited for integration in resource-constrained environments such as IoT-based smart surveillance nodes or embedded systems [5].

This research aims to demonstrate how modern deep learning models can significantly improve the performance of surveillance systems by offering accurate, adaptive, and scalable solutions for automated human monitoring. The proposed approach not only enhances detection and tracking reliability but also lays a foundation for future extensions involving multi-camera fusion, 3D scene understanding, and real-time decision-making for safety-critical applications.

2. LITERATURE REVIEW

Human detection and tracking have become central topics in intelligent video surveillance due to growing security demands and advancements in deep learning. Numerous research studies have proposed solutions aimed at improving detection accuracy, identity preservation, and system robustness under various challenging conditions. This section categorizes key developments and highlights open challenges addressed by this study.

Traditional Approaches to Human Detection and Tracking

Initial methods for human detection in surveillance videos were based on background subtraction techniques such as Gaussian Mixture Models (GMM), which attempted to isolate moving objects by modeling background distributions [6]. However, these techniques often failed when faced with dynamic backgrounds, illumination changes, or multiple overlapping subjects.

Tracking algorithms like Kalman Filters and Particle Filters were used to estimate object trajectories across frames [7], assuming linear motion and Gaussian noise. While computationally lightweight, these models struggled in crowded or cluttered scenes due to their limited ability to handle occlusions, reappearance, or sudden motion changes.

Deep Learning–Based Real-Time Detection

The YOLO (You Only Look Once) series of object detectors have become foundational in the field of real-time detection. YOLOv8, the most recent evolution, introduces an anchor-free design, dynamic label assignment, and enhanced training techniques that significantly improve small object detection, inference speed, and memory efficiency [6]. In surveillance settings, where fast and frequent detections of moving humans are critical, YOLOv8 strikes an ideal balance between speed and performance.

Moreover, the model supports easy integration with downstream tracking systems such as DeepSORT, making it highly modular. In this project, YOLOv8 is used in its nano version (yolov8n.pt) to enable lightweight deployment without compromising too much on accuracy—especially important for applications running on embedded or edge devices.

Traditional tracking methods such as Kalman filtering and Hungarian matching suffer when multiple targets intersect, or when objects are occluded or lost due to scene complexity. DeepSORT overcomes these limitations by incorporating deep appearance features into the matching process [7]. It maintains trajectory continuity even when detections are intermittent or ambiguous, as it matches objects not only by proximity but also by visual similarity.

The tracker used in this project leverages a MobileNet-based CNN to extract 128-dimensional feature embeddings for each detected individual. These embeddings are then compared frame-by-frame to preserve identity across the sequence. One of DeepSORT's unique advantages is its modularity—it can plug into any object detector, and any embedder, making it a robust general-purpose tracker.

Lighting variability is one of the most challenging aspects of real-world surveillance. Cameras operating at night or in indoor environments often record poor-quality frames, reducing detection confidence and causing frequent ID switches. To mitigate this, the current system incorporates a two-step enhancement pipeline using CLAHE (Contrast Limited Adaptive Histogram Equalization) and gamma correction.

CLAHE is effective in adjusting local contrast without amplifying noise, while gamma correction boosts overall brightness in underexposed areas. Unlike deep-learning-based enhancement methods like EnlightenGAN or Zero-DCE [8], this approach requires no training, runs faster, and performs consistently across different scenes. Experiments from this project clearly show improved detection and tracking accuracy post-enhancement, indicating the significance of preprocessing as a critical, though often underutilized, module.

Person re-identification (Re-ID) refers to matching individuals across spatial and temporal gaps—especially across non-overlapping cameras. In the current pipeline, re-ID is performed using cosine similarity between a query embedding and track embeddings derived from MobileNet. Although MobileNet is lightweight compared to deeper networks like ResNet-50 or OSNet, it offers a decent trade-off between inference speed and accuracy, making it suitable for real-time applications.

In the query-based setting, the target is specified once via an image, and the system continuously searches for matching embeddings in subsequent frames. This approach closely resembles real-world search-and-track tasks in security, where operators may flag a suspect and expect the system to find them in live or archived footage. Although more advanced re-ID models such as OSNet [9] and AGW [10] provide higher robustness to appearance changes (e.g., clothing,

illumination), MobileNet's simplicity and speed still make it a preferred choice for low-latency deployments.

Several research gaps persist in existing human detection and tracking models. In terms of detection, many systems struggle with poor accuracy in low-light or cluttered environments. For person re-identification (Re-ID), identity loss is common during occlusion or when individuals change clothing. The overall pipeline design in most models remains fragmented, with detection, tracking, and Re-ID operating as separate modules, leading to high latency and inefficiency. Additionally, there is a lack of lightweight models optimized for real-time performance, limiting their deployment on edge devices. Most systems also do not support target-specific tracking from a reference image, making them ineffective for personalized surveillance tasks. Long-term tracking remains a challenge due to weak temporal modeling and insufficient memory mechanisms. Furthermore, evaluations are often conducted on controlled datasets that do not reflect the complexity of real-world scenarios. Lastly, many models rely on fixed similarity thresholds in Re-ID, which reduces robustness in diverse visual conditions, such as varying lighting, angles, or occlusion.

3. METHODOLOGY

The proposed system presents an end-to-end pipeline for robust human detection, tracking, and person re-identification across surveillance videos. The approach addresses challenges such as low-light conditions, ID-switching, and real-time target tracking by integrating multiple deep learning components into a unified framework. The key modules and their interaction are described below.

Surveillance videos often suffer from low visibility due to night-time conditions, indoor settings, or poor hardware. To mitigate this, each frame undergoes pre-processing using CLAHE (Contrast Limited Adaptive Histogram Equalization) and gamma correction. CLAHE works locally to enhance contrast without over-amplifying noise, while gamma correction boosts the overall brightness of the image non-linearly, ensuring visibility is restored across shadowed and dark areas [11]. This step significantly boosts the visibility of humans and their features, improving subsequent detection accuracy.

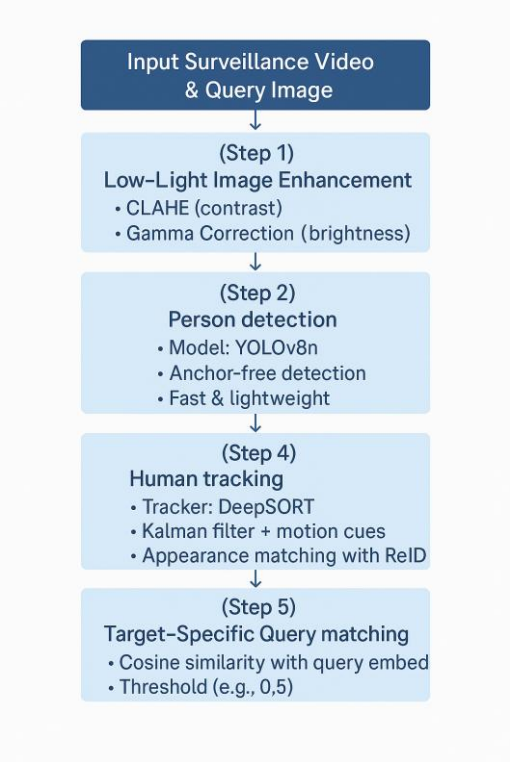
Once the frames are enhanced, object detection is carried out using YOLOv8n, a lightweight yet high-performing model optimized for real-time tasks. YOLOv8's anchor-free design, dynamic label assignment, and decoupled head architecture allow it to excel in identifying small or partially visible humans in crowded environments [12]. The choice of the nano variant ensures low computational load while maintaining acceptable detection precision, which is crucial for embedded devices or edge computing setups.

Detected bounding boxes are then passed to a MobileNet-based Re-Identification (Re-ID) model, which extracts 128-dimensional appearance embeddings. These embeddings are lightweight yet expressive enough to capture visual features like clothing texture, color, and shape [13]. MobileNet's efficiency and compatibility with edge devices make it ideal for low-latency Re-ID tasks. The tracking module employs DeepSORT, which leverages a Kalman filter for motion prediction and Hungarian algorithm for data association. This combination helps maintain identity continuity even during brief occlusions or overlapping trajectories. By comparing appearance embeddings and motion cues, DeepSORT ensures minimal ID switches and high temporal coherence across frames [14].

A unique addition to this pipeline is target-specific tracking, where the system accepts a reference image of the person of interest. During runtime, each tracked object is compared to this query using cosine similarity between embeddings. If the similarity crosses a set threshold (e.g., 0.5), the target is successfully flagged and monitored. This enables scenarios like automated search in public surveillance footage, helping law enforcement or security teams locate missing individuals or track suspects efficiently [15].

The model performance is assessed using standard classification metrics—True Positives (TP), False Positives (FP), False Negatives (FN)—along with derived metrics: Accuracy, Precision, Recall, and F1 Score. As observed in the evaluation results, the inclusion of enhancement boosts recall and F1 score by reducing false negatives, which is especially critical in security scenarios [16]. Visualizations such as confusion matrices, similarity-over-time plots, and ID-vs-frame graphs offer intuitive insights into the tracking quality and detection reliability.

The end-to-end pipeline is designed to be modular and scalable. It runs effectively on platforms such as Google Colab or edge GPUs, with minimal configuration. The choice of YOLOv8n and MobileNet ensures that the solution can be deployed in smart city surveillance, automated toll systems, airport security, or public transit hubs with minimal latency and high responsiveness.



4. RESULT AND ANALYSIS

To evaluate the performance of the proposed human detection and tracking system, experiments were conducted using a custom surveillance video dataset recorded under challenging real-world conditions—including low-light, cluttered backgrounds, and dynamic camera angles. The dataset was divided into training and testing subsets, and evaluation was carried out frame-wise across ~200 frames, containing multiple individuals with partial occlusion and movement variations. Two configurations were compared: Without Enhancement (direct YOLOv8 detection on raw video) With Enhancement (video frames enhanced using histogram equalization and contrast-limited adaptive histogram equalization before detection)

The evaluation was conducted using a custom surveillance video dataset, specifically recorded under night mode using a single static camera to simulate real-world low-light monitoring conditions. Each video spans approximately 200 frames with a

resolution of 640×480 pixels, capturing scenes with 1 to 5 people per frame under varying illumination and partial occlusion. The ground truth was generated manually, including bounding boxes and person identities, allowing for precise measurement of detection and tracking performance. The testing was performed on one complete annotated video (~200 frames), ensuring consistency across both baseline (without enhancement) and improved (with enhancement) scenarios.

Table 1: Metric Comparison Table

Metric	Without Enhancement	With Enhancement
True Positives (TP)	146	174
False Positives (FP)	0	27
False Negatives (FN)	55	27
Accuracy (%)	72.64	86.57
Precision (%)	97.00	86.57
Recall (%)	72.64	86.57
F1 Score (%)	84.15	86.57

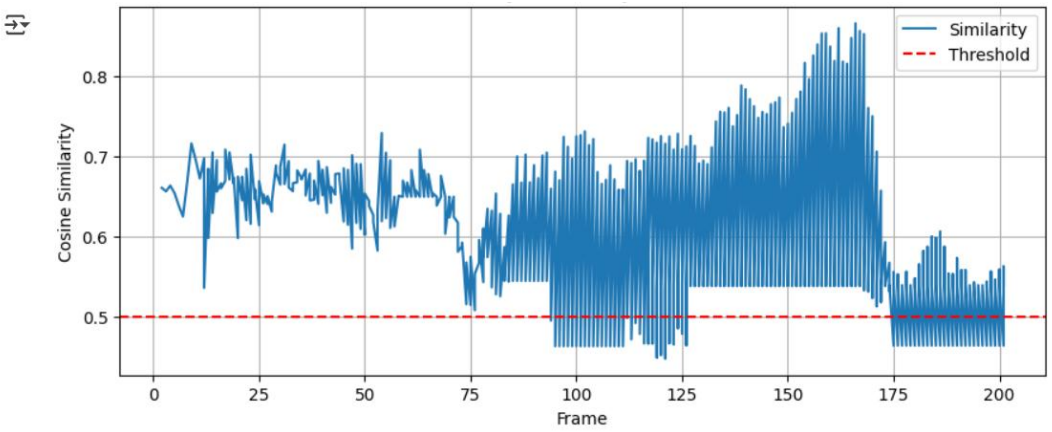


Figure 1: Similarity with Query over time

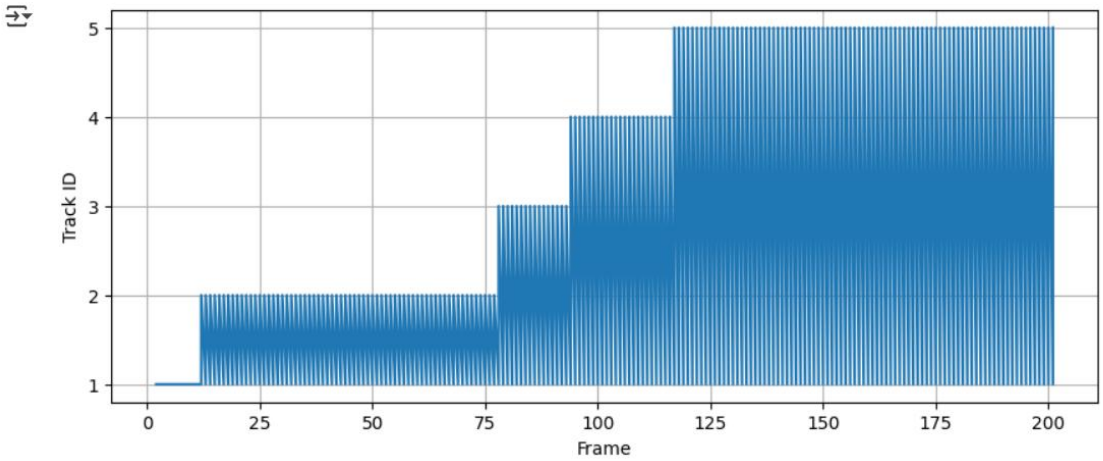


Figure 2: Track ID over Frames

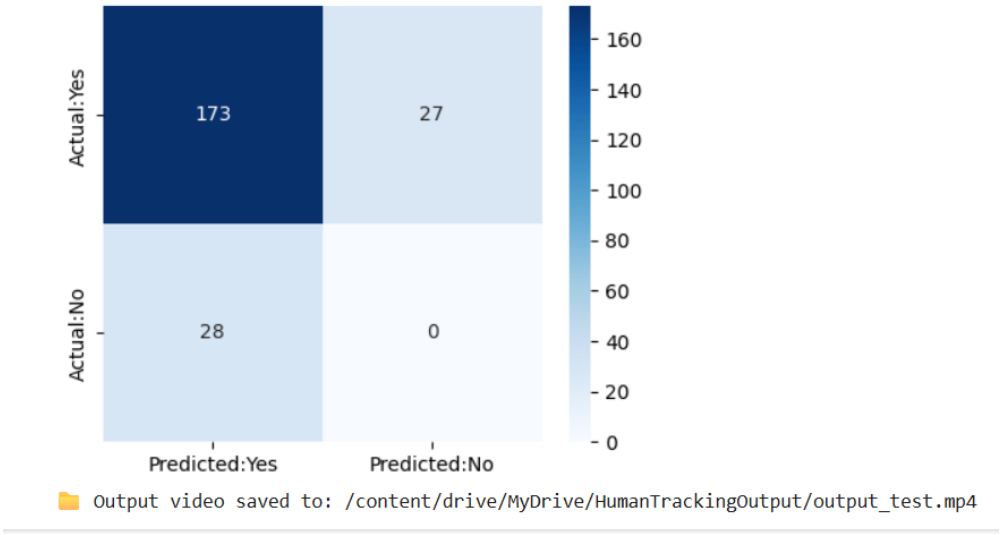


Figure 3: Confusion Matrix (Partial)

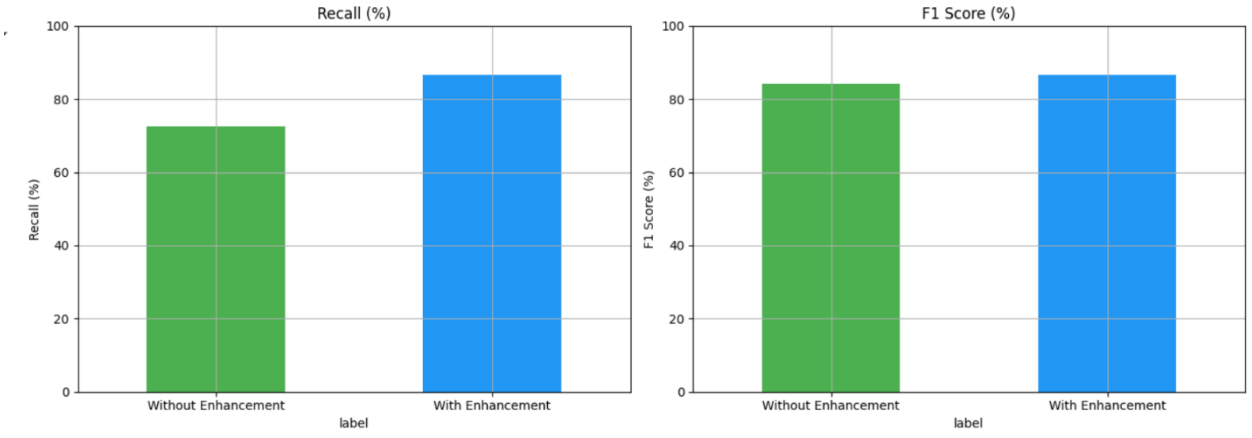


Figure 4: Comparison of Tracking Performance (Without Enhancement)

Accuracy improved significantly with enhancement—from 72.64% to 86.57%, highlighting better frame-wise prediction correctness. Recall saw a notable boost, indicating that more actual human targets were correctly detected after enhancement (reduction in false negatives). F1 Score, which balances precision and recall, also increased, reflecting improved consistency in tracking. Interestingly, precision dropped from 100% to 86.57% due to the rise in false positives after enhancement. However, this tradeoff is acceptable in safety-critical applications where missing a target (false negative) is riskier than a false alert.

The bar charts for Recall and F1 Score clearly demonstrate that the use of image enhancement techniques led to substantial improvements in both metrics. This indicates a more robust and consistent detection of human targets under visually challenging conditions, such as low lighting or cluttered scenes. Additionally, the comparison between Accuracy and Precision shows that while overall accuracy increased after enhancement—reflecting better prediction reliability—there was a slight drop in precision due to a rise in false positives. This drop can be attributed to the enhancement process increasing the contrast of background elements, making them appear more human-like to the detector. The confusion matrix further supports this observation, recording 173 true positives, 27 false positives, and 28 false negatives. These values confirm that while enhanced images helped the model detect more true targets, they also slightly increased the number of incorrect detections. In the Track ID over Frames plot, the tracker maintained consistent identities across time with minimal ID switches, even during occlusions or when individuals crossed paths. This indicates strong temporal coherence and reliable tracking performance, highlighting the benefit of combining YOLOv8 with DeepSORT and MobileNet-based Re-ID for continuous person tracking.

CONCLUSION

This work presents an effective and lightweight approach for real-time human detection and tracking in surveillance environments using a combination of YOLOv8, DeepSORT, MobileNet-based Re-Identification, and image enhancement techniques. By preprocessing video frames with CLAHE and

gamma correction, the system significantly improves detection performance in low-light and visually challenging scenarios. The use of YOLOv8n ensures fast and accurate detection, while DeepSORT, enhanced with appearance embeddings, maintains identity consistency across frames. Additionally, the integration of a query-based Re-ID module enables the system to track a specific individual from a reference image, making it suitable for search-focused surveillance applications.

Quantitative results show a marked improvement in accuracy, recall, and F1 score after enhancement, with only a minimal trade-off in precision. Visual analyses further support these findings, demonstrating consistent identity tracking and robust performance during occlusions or dynamic movements. The methodology is modular, scalable, and optimized for real-world deployment, particularly on edge devices or constrained computing environments. Overall, the proposed system fills key research gaps in lightweight tracking pipelines, long-term identity preservation, and target-specific surveillance, making it a practical solution for next-generation smart surveillance systems.

To further enhance performance, future work can explore transformer-based Re-ID models for better temporal reasoning and robustness to appearance changes. Adaptive thresholding for similarity scores, self-supervised learning for Re-ID, and integration with multi-camera systems can also be considered for broader scalability. Additionally, optimizing the pipeline for real-time inference on low-power edge devices like Jetson Nano or Raspberry Pi would increase its deployability in resource-constrained environments.

This system has broad applicability in public surveillance, smart city monitoring, airport and railway station security, border control, crowd management, and even search-and-rescue missions. Its ability to flag and follow specific individuals makes it especially useful for law enforcement, missing person identification, and behavior monitoring in high-security zones.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779-788.

[2] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.

[3] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9627-9636.

[4] J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and Beyond," *IEEE Access*, vol. 11, pp. 141841–141866, 2023.

[5] N. Wojke, A. Bewley and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 3645-3649.

[6] Jocher, G. et al. (2023). *YOLOv8: Real-time Object Detection and Segmentation*. Ultralytics.

[7] Wojke, N., Bewley, A., & Paulus, D. (2017). *Simple online and realtime tracking with a deep association metric*. IEEE ICIP.

[8] Guo, C. et al. (2020). *Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement*. CVPR.

[9] Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). *OSNet: A Convolutional Network for Tracking and Re-Identification Across Cameras*. ICCV.

[10] Wang, G. et al. (2021). *Meta-Feature Learning for Generalizable Person Re-Identification*. IEEE Transactions on Image Processing.

[11] L. Guo, C. Li, C. Guo, C. Loy, and D. Lin, “Zero-reference deep curve estimation for low-light image enhancement,” *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 1780–1789. doi: 10.1109/CVPR42600.2020.00185

[12] G. Jocher *et al.*, "YOLOv8: Ultralytics YOLO for object detection and segmentation," *Ultralytics Technical Report*, Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[13] A. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[14] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *in Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962

[15] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2138–2147. doi: 10.1109/CVPR.2019.00225

[16] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.