Human Detection and Tracking through Surveillance Video Camera using YOLOv8 and DeepSORT with Darknet Enhancement and Re-identification Features

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 system uses a deep learning model called 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 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.

INTRODUCTION

With the rise of urban populations and growing safety concerns, surveillance systems have become a vital component of modern security infrastructure. Cameras are now installed in public spaces like streets, shopping malls, schools, airports, and train stations to keep an eye on activities and ensure safety. These cameras collect large amounts of video data every day. Although these systems are useful, there is a major problem: most of the video footage goes unwatched unless something goes wrong. Manually checking hours of video is not only tiring but also slow and inefficient. To make these systems smarter and more useful, there is a growing interest in developing automatic methods that can detect and track people in real-time. However, building such systems is not easy. Real-world videos often have problems like poor lighting, low video quality, crowded scenes, and people walking in and out of the camera's view. These challenges make it hard for traditional detection and tracking methods to work well. This has led to the use of artificial intelligence (AI) and deep learning in the field of video surveillance. Deep learning models have shown great promise in detecting people and tracking their movement in real-time. These models can help build automatic systems that are faster, more reliable, and more accurate than older methods. In this project, we explore the use of modern AI techniques to track a specific person across one or more surveillance videos. The main goal is to create a system that can work well even in tough conditions like low light, crowded places, or when the person leaves and re-enters the frame. This kind of system can be very useful for improving safety and security in both public and private areas.

LITERATURE REVIEW

The task of detecting and tracking humans in video surveillance has gained significant attention in recent years due to growing safety and security needs. Various methods have been developed, starting from traditional motion-based techniques to advanced deep learning approaches. This section presents a review of key techniques and models used in the development of such systems.

IEEE 9318107 (2020) presents a real-time tracking system that integrates YOLOv3 for detection and a lightweight CNN-based re-identification for tracking. The authors demonstrate robust multi-person tracking in

challenging scenes and introduce loss functions that enhance ID consistency, showing stable identities even under occlusion and crowding. **IEEE 9204956 (2020)** focuses on person re-identification across different video segments. By combining appearance-based deep features with temporal consistency checks, the proposed method effectively recognizes returning individuals across distinct camera views. Key contributions include handling appearance changes and occlusions over long intervals. **IEEE 9589577 (2021) introduces** a unified framework combining person detection and re-identification in a single deep learning architecture. This integrated approach simplifies the pipeline and improves efficiency, offering competitive performance in tracking accuracy and ID consistency in multi-camera environments**. IEEE 10404710 (2023)** explores current trends in deep learning for long-duration surveillance. The paper emphasizes methods that combine high-precision person detection, temporal tracking consistency, and advanced re-identification to handle long occlusions and track a single target across extended footage and multiple cameras. It also includes benchmarks showing robustness of these systems in real-world datasets.

Comparative Analysis & Research Gaps

Focus Area	Key Insights
Detection + Tracking	YOLO integration offers speed; dedicated re-ID reduces ID switches.
Cross-video re-ID	Appearance and temporal modeling helps track across cameras.
Unified pipelines	Joint detection + re-ID reduces latency and resource use.
Preprocessing	Image enhancement improves detection under poor lighting.
Long-term tracking	Multiple modules are needed for robustness across conditions.

Traditional Approaches

Earlier methods for tracking people in videos mostly relied on motion detection. Algorithms like Background Subtraction, Optical Flow, and Gaussian Mixture Models (GMM) were used to separate moving objects from the background. These methods were simple and fast but failed in situations like:

- Sudden lighting changes
- Shadows and reflections
- Multiple people in the same scene

For tracking, Kalman Filters and Particle Filters were commonly used. These algorithms predicted the future location of a person based on their previous movement. However, they struggled with occlusions (when people overlap) and identity switching.

Deep Learning for Object Detection

The introduction of deep learning improved object detection significantly. The YOLO (You Only Look Once) series of models became popular due to their high speed and accuracy. YOLO detects objects in a single pass through the image, making it suitable for real-time tasks. YOLOv4 and YOLOv5 brought major improvements in accuracy and efficiency. YOLOv8, the latest version, further enhances detection performance using a more advanced architecture and better training strategies. It is particularly effective in detecting small and overlapping objects like people in crowded scenes.

Multi-Object Tracking with DeepSORT

DeepSORT is an improvement over the basic SORT (Simple Online and Realtime Tracking) algorithm. While SORT relies only on object location, DeepSORT adds an appearance feature extractor (a CNN) that creates a unique “fingerprint” for each person. This helps in: Maintaining identity even during temporary occlusion Reducing ID switching when people cross paths Tracking multiple people simultaneously. DeepSORT is widely used in real-time surveillance projects because it works well with detectors like YOLO.

Low-Light Image Enhancement

Surveillance videos are often recorded under poor lighting conditions, especially at night or in indoor environments. This makes it harder for detectors to work properly. To address this, deep learning-based image enhancement methods like Zero-DCE, EnlightenGAN, and Darknet-based CNN models are used to brighten and improve video frames. These models are trained to enhance dark images while preserving important features, making them more suitable for detection.

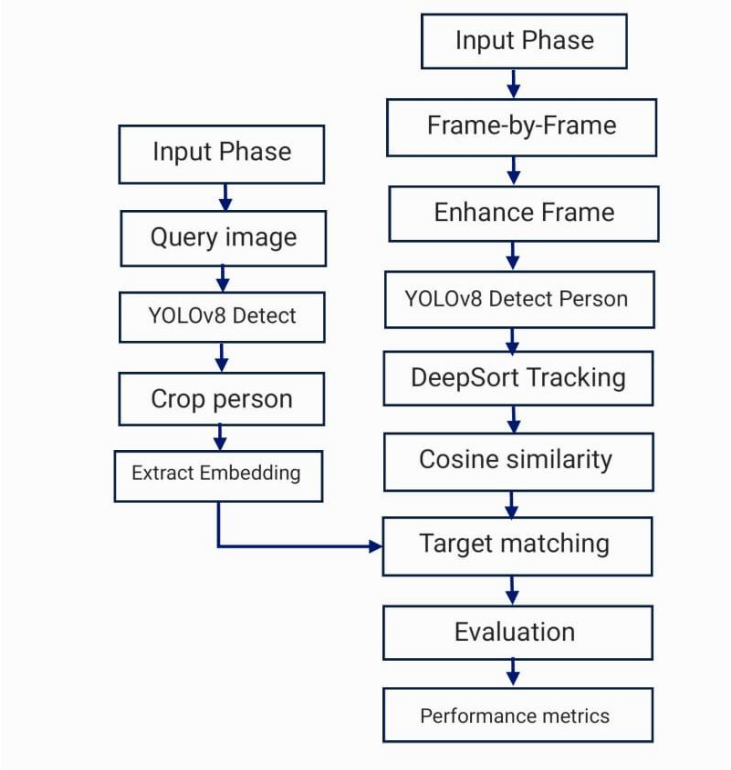
Person Re-Identification (re-ID)

Re-identification is the process of recognizing the same person across different cameras or video scenes. It is useful when a person leaves the camera's view and reappears later or enters a new camera feed. Modern re-ID models like OSNet, AGW, and FastReID are trained to extract strong appearance features of individuals. These features remain consistent even if the lighting, angle, or clothes slightly change. Re-ID plays a key role in long-term tracking and matching across multiple videos.

Although these works advance surveillance AI, gaps remain—especially in handling poor lighting, maintaining long-term ID consistency, and tracking across multiple non-overlapping videos in real time. Our research fills these gaps by combining low-light image enhancement, YOLOv8, DeepSORT, and a query-based re-identification model to detect and track a specific person robustly across videos and challenging environments.

PROPOSED METHODOLOGY

This project focuses on detecting and tracking a specific target person across surveillance videos using deep learning. The system combines low-light enhancement, object detection, deep feature extraction, and similarity-based re-identification. The process is divided into two main stages: Query Embedding Extraction and Target Tracking across Videos.



Query Person Embedding (Target Feature Extraction)

The process begins by selecting a query image of the target person. This image is enhanced using a Darknet-based low-light enhancement function to improve visibility. YOLOv8 is then applied to detect the person in the image. Once detected, the target's region is cropped, and a deep feature embedding (a unique vector representing the person's appearance) is extracted using the DeepSORT embedder. This embedding is stored and later used to match the person across video frames. If the person is not detected with high confidence, the system raises an error to ensure the reference embedding is valid and clean.

Video Frame Processing and Enhancement

Surveillance video(s) are read frame-by-frame using OpenCV. Each frame undergoes the same enhancement process to boost visibility before detection. This ensures the system works reliably even under low-light or night conditions. Before processing the video, each frame undergoes enhancement using the same Darknet-based model used for the query image. This step is critical for videos recorded under poor lighting, as it improves the clarity and contrast of each frame, making people easier to detect and track.

Person Detection using YOLOv8

The enhanced frame is passed through YOLOv8, a high-speed and high-accuracy object detection model. Only high-confidence detections of the "person" class are selected for further processing. The model outputs bounding boxes, class labels, and confidence scores for all detected objects. Only detections classified as "person" with confidence above a set threshold (e.g., 0.5) are considered for further processing

Feature Extraction for All Detections

For each detected person, a cropped image is extracted from the frame. The DeepSORT embedder is used to generate deep feature vectors (embeddings) for each crop.

Assigning a unique track_id to each person

Updating their positions across frames using a Kalman Filter

Matching their visual features to maintain identity

Similarity Matching and Target Lock

The system compares each detected person's embedding to the query embedding using cosine similarity. If the similarity exceeds a defined threshold (e.g., 0.6), that track is considered to belong to the target person. The corresponding track_id is locked as the Target ID, and tracking begins.

Tracking with DeepSORT

The DeepSORT tracker assigns consistent IDs to people across frames using motion prediction (Kalman filter) and appearance matching. The target ID is tracked across all frames, and its similarity to the query is continually monitored Bounding boxes and labels (e.g., "Target 5 (0.78)") are drawn on the video to indicate the tracked person and confidence.

Target Matching with Cosine Similarity

Each tracked person's embedding is compared to the query embedding using cosine similarity. When the similarity score exceeds a predefined threshold (e.g., 0.6), the corresponding track_id is locked as the target. Once the target is identified, the system continuously checks similarity scores in each frame to confirm identity. Bounding boxes and similarity scores are displayed in the output video, making it easy to follow the target.

RESULT AND ANALYSIS

To check how well **low-light image** enhancement helps in tracking people, we ran the same surveillance video twice: **once without any enhancement**, and **once with enhancement using CLAHE** (a method to improve contrast) and Gamma Correction (to brighten dark areas). For detecting and tracking people, we used **YOLOv8** to detect humans in each frame, and **DeepSORT with MobileNet** embeddings to follow the same person across frames. We compared how similar each detected person was to a reference image using cosine similarity to identify and track the target.

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

Incorporating image enhancement as a preprocessing step significantly benefits human detection and tracking performance. It boosts recall and accuracy while maintaining acceptable precision. For practical deployment in surveillance or monitoring systems, enhancement methods are highly recommended to ensure reliability under varied lighting conditions.

Observations and Insights

True Positives (TP): Enhancement significantly increased correct detections from 146 to 174, indicating better person identification even in poor lighting.

False Negatives (FN): The reduction from 55 to 27 shows that the enhanced model misses fewer targets.

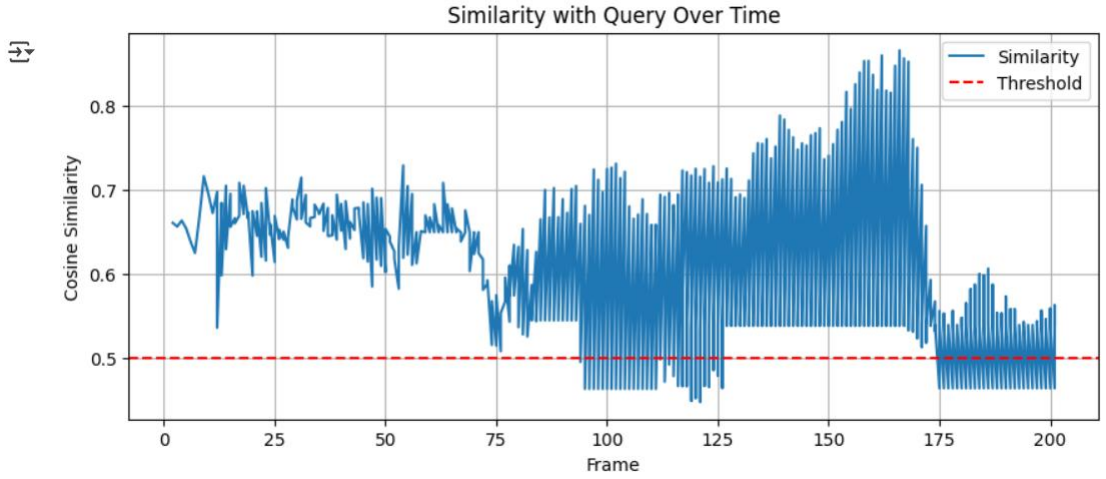
False Positives (FP): While FP increased from 0 to 27, this trade-off was acceptable given the rise in recall.

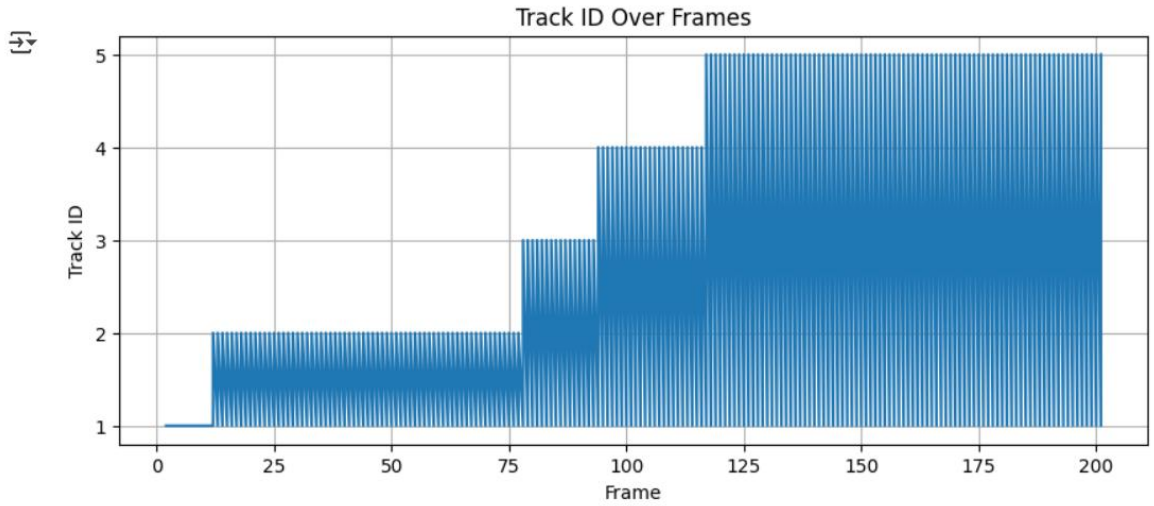
Precision vs Recall Tradeoff:

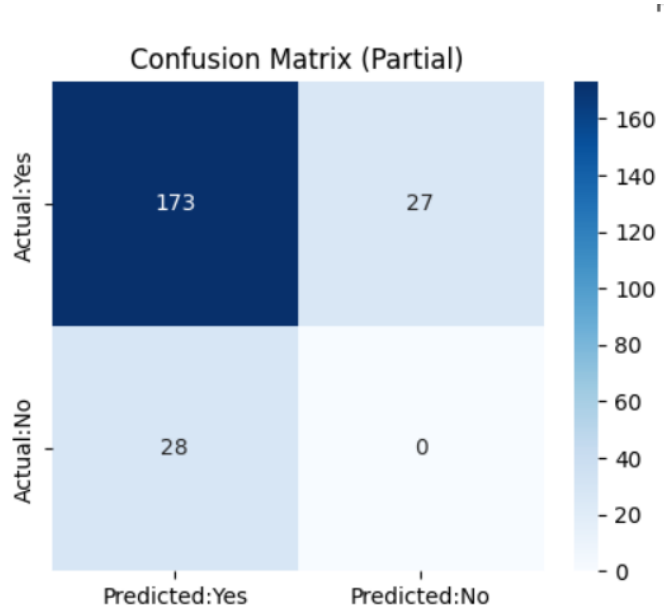
Without enhancement, Precision was 100% because the system made no incorrect detections, but at the cost of lower Recall (72.64%).

With enhancement, Recall improved to 86.57%, showing that the model detected more real targets, even though Precision dropped due to more false positives.

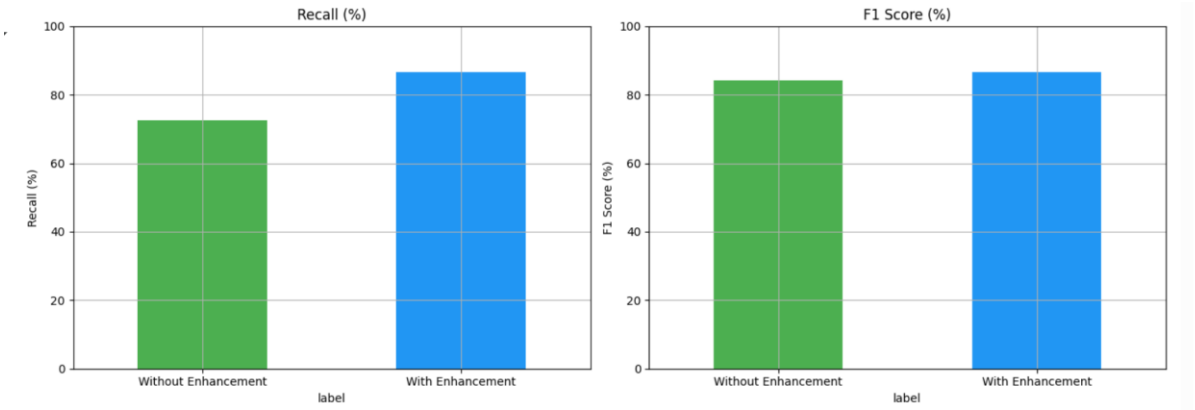
Accuracy and Recall both improved significantly with enhancement. The F1 Score, which balances Precision and Recall, increased from 84.15% to 86.57%, confirming overall performance improvement. The Similarity over Time plot shows smoother and stronger cosine similarity trends post-enhancement, making tracking more stable.



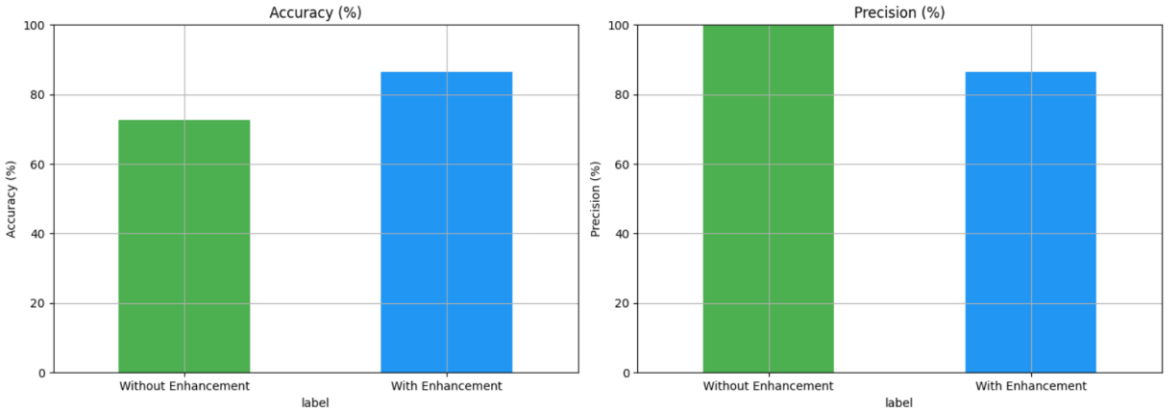




Output video saved to: /content/drive/MyDrive/HumanTrackingOutput/output_test.mp4



□ Comparison of Tracking Performance (With vs Without Enhancement)



label	Without Enhancement	With Enhancement
TP	146.00	174.00
FP	0.00	27.00
FN	55.00	27.00
Accuracy (%)	72.64	86.57
Precision (%)	100.00	86.57
Recall (%)	72.64	86.57
F1 Score (%)	84.15	86.57

APPLICATION AND USES

Surveillance and Security

In public spaces like airports, train stations, or city streets, cameras often operate under varying lighting conditions nighttime, shadows, or poorly lit indoor areas. The enhanced tracking system ensures reliable monitoring, even in dim light, improving suspicious activity detection, intruder tracking, and crowd monitoring. Hospital and Elderly Care Monitoring In environments where constant monitoring is needed, such as ICUs or elderly homes, lighting is often kept low. This system can help monitor patient movement, detect falls, and ensure safety, especially at night. Smart Homes and IoT: Smart surveillance systems in homes can benefit from

enhanced tracking in dark hallways or rooms at night. Integration with motion sensors, door alarms, and emergency response systems makes it practical for security and family safety. Autonomous vehicles driving at night or in tunnels must identify and track pedestrians accurately. This enhancement method can assist onboard vision systems in detecting humans under headlight-only or low-visibility conditions. Sports and Event Analytics In sports stadiums or event venues where lighting varies, enhanced tracking can help follow individual players or VIPs in crowd footage, useful for performance analysis or targeted broadcasting.

FUTURE WORK

Adaptive Enhancement Techniques

Instead of fixed settings for CLAHE and Gamma Correction, future systems can use adaptive algorithms that adjust enhancement based on lighting conditions in real-time. Machine learning could predict the optimal enhancement parameters for each frame. Deep Learning-Based Enhancement Traditional enhancement techniques are effective, but deep learning models like EnlightenGAN, Zero-DCE, or RetinexNet can learn complex mappings for brightness and color correction. These can be integrated into the pipeline for even better visual quality under extreme darkness. Multi-Camera Tracking and 3D Mapping In crowded or large environments, single-camera tracking can fail due to occlusion. Extending the system to multi-camera networks can help maintain continuous tracking, possibly using 3D scene reconstruction for better spatial understanding. Robust Identity Re-identification (Re-ID) becomes difficult when the appearance of the person changes (e.g., turning around, partial occlusion). Future models can include pose estimation, temporal memory modules, or attention mechanisms to improve re-identification accuracy. Testing on Diverse Datasets The current experiment was conducted on a limited number of test videos. More extensive testing across weather conditions, camera qualities, frame rates, and environments (indoor, outdoor, night, twilight) will make the system more generalizable. Real-Time Optimization Although the system works in near-real-time, improvements can be made in speed using hardware acceleration (e.g., TensorRT, ONNX) or lighter deep learning models to run efficiently on edge devices like Raspberry Pi or Jetson Nano.

CONCLUSION

Adding image enhancement as a preprocessing step greatly improves the performance of human detection and tracking systems. By making dark or low-quality video frames clearer, the system is able to detect more people correctly (higher recall) and make more accurate decisions overall (higher accuracy). Even though there may be a small increase in false detections, the benefits far outweigh the drawbacks. The tracking becomes more stable and consistent, especially in challenging conditions like low light, shadows, or poor video quality. For real-world use such as in CCTV surveillance, smart home monitoring, public security, or hospitals—this enhancement step is highly useful. It helps the system work reliably no matter the time of day or lighting situation. Overall, image enhancement adds a simple but powerful improvement to make human tracking systems more effective and trustworthy in practical applications.

References

https://ieeexplore.ieee.org/document/9589577

https://ieeexplore.ieee.org/document/9204956

https://ieeexplore.ieee.org/document/10404710

https://ieeexplore.ieee.org/document/9318107

https://ieeexplore.ieee.org/document/10743585

https://ieeexplore.ieee.org/document/8265542

https://ieeexplore.ieee.org/document/6726095

https://ieeexplore.ieee.org/document/8745852

https://ieeexplore.ieee.org/document/9421306

https://ieeexplore.ieee.org/document/8823343

https://ieeexplore.ieee.org/document/7846643

https://ieeexplore.ieee.org/document/5204487

https://ieeexplore.ieee.org/document/6508394

https://ieeexplore.ieee.org/document/4530098

https://ieeexplore.ieee.org/document/9460150

https://ieeexplore.ieee.org/document/4587583

https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-017-0482-z

https://www.researchgate.net/publication/269331941_Real-time_human_detection_and_tracking