International Journal of Current Engineering and Technology
E-ISSN 2277-4106, P-ISSN 2347-5161
©2026 IJCET ®, All Rights Reserved
Available at https://ijcet.evegenis.org/index.php/ijcet/index

Research Article

Crowd Monitoring by Video Surveillance System through Real-Time Object Detection and Tracking using AI Techniques

Vaibhav Katiyar*, Dr. Mamta Tiwari

CSE Department, CSJMU, Kanpur, India

Received 10 May 2026, Accepted 31 May 2026, Available online 01 Jun 2026, **Vol.16, No.3 (May/Jun 2026)**

Abstract

Real-time video surveillance is an essential component of the modern security system, particularly in the context of monitoring public areas, particularly in densely populated areas such as airports, train stations, and urban junctions. Occlusion, quick movement, and high item density are some of the issues that traditional surveillance systems face when accurately detecting and effectively monitoring objects. YOLO, which stands for "You Only Look Once," is a framework that includes DeepSORT, which stands for "Simple Online and Realtime Tracking with a Deep Association Metric," for effective object tracking. This framework is proposed in this study as a robust multi-object identification and tracking system. By utilizing appearance traits and Kalman filtering, YOLO guarantees that objects are localized and classified quickly and accurately, whereas DeepSORT ensures that identification is maintained consistently between frames. Under various lighting and crowd density situations, the combined model was implemented and tested on video feeds captured from the real world. Given that the system can handle occlusions and re-identify lost targets with high accuracy, low latency, and resilience, as demonstrated by the results of the experiments, it is a feasible option for applications that include intelligent surveillance. When applied in dynamic and complex public contexts, this technique dramatically improves situational awareness and assists proactive security management.

Keywords: *Multi Object, Detection, Model, YOLO, Video*

Introduction

The rapid urbanization and growing concerns over the safety of the general population have led to a greater dependence on intelligent video surveillance systems. Conventional surveillance systems, which are highly reliant on manual monitoring, are inefficient and prone to errors, particularly in high-density places such as airports, retail malls, train stations, and public festivals and events. Occlusions, overlapping objects, rapid motion, and the requirement for consistency in identity tracking over time are some of the issues that are present in these settings. Automated multi-object detection and tracking systems have become indispensable for maintaining real-time situational awareness and effective threat responses. This is accomplished by addressing the identified constraints. Recent developments in computer vision and deep learning have significantly improved the performance of real-time object recognition and tracking. With a single forward pass of the neural network, YOLO (You Only Look Once), which is one of the most common object identification models, provides real-time speed and excellent accuracy.

*Corresponding author's ORCID ID: 0000-0000-0000-0000
DOI: https://doi.org/10.14741/ijcet/v.16.3.5

It does this by predicting item classes and bounding boxes. Because of its capacity to handle high-resolution video frames with low latency, YOLO is particularly well-suited for use in surveillance applications. YOLO, on the other hand, is particularly good at recognising many things at the same time; but, it does not automatically follow these items across several frames. With the intention of bridging this gap, YOLO has been merged with DeepSORT, which stands for Simple Online and Realtime Tracking. DeepSORT improves tracking performance by integrating motion estimates using the Kalman Filter with appearance descriptors generated by a deep neural network. This integration results in a more accurate tracking performance. Because of this, it is possible to achieve more strong object association and re-identification, even in situations where objects are briefly obscured or leave and then return to the scene repeatedly. In situations that are both complicated and busy, the combination of YOLO with DeepSORT results in a strong framework that is able to recognise and track many items simultaneously. The ability to do thorough movement analysis and behaviour monitoring is made possible by this system, which not only recognises and categorises objects in each frame but also keeps identification labels consistent over subsequent frames. The purpose

of this work is to demonstrate a real-time video surveillance system that makes use of YOLO for the purpose of detecting objects in a quick and accurate manner, and DeepSORT for the purpose of accomplishing reliable multi-object tracking. Performance parameters like as accuracy, recall, Multiple Object Tracking Accuracy (MOTA), and processing speed (FPS) are analysed in order to assess the system using video datasets that are representative of the cluttered situations that exists in the real world. The objective is to provide a surveillance system that is both intelligent and automated, and that is capable of functioning well under a wide variety of adverse environmental circumstances. This strategy makes a substantial contribution to the construction of urban settings that are both more intelligent and safer by enhancing the dependability and responsiveness of surveillance systems.

Object Detection Techniques

Earlier object identification methods, such as R-CNN (Girshick et al., 2014), presented a two-stage pipeline that consisted of region proposal followed by classification. This pipeline was successful in achieving correct results, but it was also sluggish. Additionally, this was enhanced by Fast R-CNN and Faster R-CNN, which added Region Proposal Networks (RPNs) in order to decrease the amount of time required for calculation while preserving accuracy. The multi-stage structure of these systems, on the other hand, limited their usefulness in real-time applications. With the release of YOLO (You Only Look Once) by Redmon et al. (2016), a breakthrough occurred. This was the first time that object identification was reformed as a single regression issue. It predicted bounding boxes and class probabilities directly from entire pictures. In order to make them appropriate for real-time surveillance, YOLOv3 (Redmon & Farhadi, 2018) and later versions (YOLOv4 and YOLOv5) brought about considerable improvements in terms of speed, accuracy, and scalability. When it comes to recognising objects in a wide range of sizes and lighting situations, these models perform exceptionally well, which is essential in busy surveillance environments.

Object Tracking Approaches

Single-object tracking (SOT) and multi-object tracking (MOT) are the two categories that are commonly used to classify object tracking methodologies. Motion-based tracking techniques, such as Kalman filters and particle filters, comprised the foundation of early MOT systems. These techniques were frequently used in conjunction with background removal or optical flow. In situations when there were obstruction or sudden shifts in motion, these techniques performed poorly. The proliferation of deep learning has led to the development of more robust solutions. Using Kalman filters and the Hungarian method for frame-to-frame data association, SORT (Simple Online and Realtime Tracking) developed a lightweight and quick tracking framework (Bewley et al., 2016). This framework was named after the acronym for the acronym. On the other hand, SORT did not include a re-identification module, which meant that it was vulnerable to ID-switching whenever objects briefly detached themselves from the frame. DeepSORT (Wojke et al., 2017) is an extension of SORT that incorporates a deep appearance descriptor. This descriptor improves data association by utilising feature similarity, which in turn enables better tracking in scenarios when there is a lot of background noise or a lot of people. Due to the fact that it strikes a compromise between accuracy and real-time speed, DeepSORT gained widespread use.

YOLO + DeepSORT Integration in Surveillance

For real-time multi-object tracking, the combination of YOLO and DeepSORT has emerged as a solution that has gained widespread use. A study that was conducted by Chen et al. (2018) indicated that the use of YOLOv3 in conjunction with DeepSORT is an excellent method for monitoring pedestrians in real-time surveillance systems. Similarly, Yilmaz and Aksu (2020) conducted tests on the MOT16 dataset using YOLOv4 and DeepSORT. The results demonstrated gains in tracking accuracy, particularly in scenes that had a high number of objects and a high level of occlusion. Other investigations, such as the ones conducted by Nagrath et al. (2021), concentrated on the monitoring of the COVID-19 protocol by identifying individuals and determining the distances that separated them through the use of YOLO and DeepSORT. Through the use of these actual applications, the flexibility and robustness of the integrated architecture are brought to light in order to fulfil a variety of surveillance objectives.

Challenges in Crowded Environments

Crowded scenes present a number of issues, including the possibility of item overlap, frequent occlusion, and backdrops that are crowded. Research conducted by Milan et al. (2016) highlights the importance of sophisticated appearance modelling and the maintenance of identity over an extended period of time in situations like these. The recent developments in transformer-based object trackers and spatial-temporal attention mechanisms (such as FairMOT and TrackFormer) have shown promise; nevertheless, these gains frequently come at the expense of overall computing efficiency. For real-time implementation, the YOLO + DeepSORT pipeline continues to be a solution that is both feasible and efficient, particularly when it is optimised for GPU acceleration. In order to further minimise ID-switches and improve tracking robustness, research is continuing to concentrate on developing re-identification models and adding scene context.

Methodology

In recent years, there has been a significant advancement in the capabilities of object identification and tracking, both of which are essential characteristics of computer vision. The objective of object detection is to locate occurrences of a certain thing through the use of photographs or videos. In order to accomplish the objective of object tracking, which is to follow a target over several frames of a movie, the algorithm depicted in Figure 1 illustrates a system that is capable of tracking objects in videos. In the beginning, tracking-by-detection techniques are used to identify objects in each frame. These approaches then associate the observed objects throughout the course of time. It is possible for powerful algorithms to track objects even when they move in complicated movements, are occluded, or alter their angle of view or appearance. During the research on object detection and tracking, the challenging task of identifying and tracking items inside video sequences was taken on by YOLO (You Only Look Once) and DeepSORT (Simple Online and real-time tracking with a Deep Association Metric). Both of these methods were utilised.

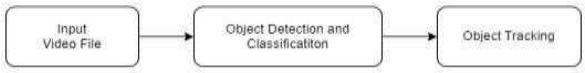

Figure 1: This is the model for detecting and tracking

A number of steps are included in the technique, including the collection of data, the preprocessing of that data, the selection of a model, and the customisation of that model accordingly. In order to complete the process of data collection, it was necessary to compile a broad variety of video footage from surveillance cameras, traffic monitoring, and robotics situations. The dataset accurately captured the real challenges that were encountered, including variations in lighting, weather, occlusions, and object kinds. It was necessary to convert raw video data into frames in order to construct ground truth annotations that identify item locations and classes after the data had been translated. Before the dataset could be used for training and assessment, it was necessary to perform some preprocessing on it. Frame selection, scaling, converting annotation formats, and data replenishment were some of the actions that were taken in order to increase the generalisability of the model. For the purpose of object detection, YOLO was selected because of its remarkable accuracy and its ability to operate in real time. In order to accommodate the dataset and the objectives of the research, the configuration of YOLO was altered. Transfer learning was utilised in order to make adjustments to the YOLO formula on the one-of-a-kind dataset. It is the responsibility of YOLO to detect objects in pictures and video frames in a timely and accurate manner. It has the ability to reliably locate and identify items, predict bounding boxes, organise objects into separate categories, and manage several objects simultaneously in real time.

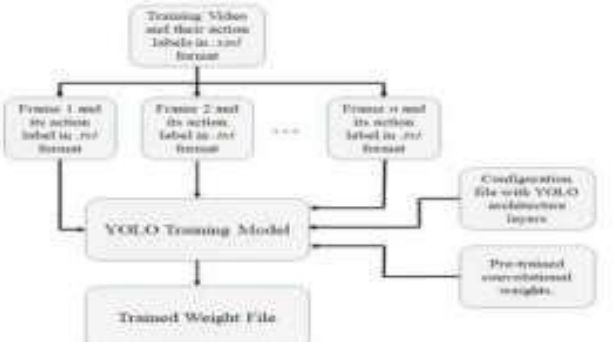

Figure 2: YOLO training flowchart is presented here.

The YOLO algorithm's flexibility to accommodate unique datasets and reap the benefits of transfer learning make it a crucial part of the object detection model. As for the object tracking system, DeepSORT was chosen. Figure 2 is a flow diagram showing the process of using YOLOv8 for detection and DeepSort for tracking to examine a movie and spot objects. By assigning each object a unique identification and following its movement from frame to frame, it is possible to create a video in which their actions may be studied. This tracking technique significantly improved YOLO's object detections by maintaining constant item IDs across frames. By integrating YOLO with DeepSORT using proprietary code, we were able to accomplish seamless object tracking. During post-processing, non-maximum suppression was used to remove duplicate detections and discoveries with low confidence. In order to make the algorithms work for the study's dataset and goals, they were fine-tuned throughout the process. Definitions of object classes, YOLO confidence criteria, and anchor box sizes were among the items altered in the changes. The incorporation of YOLO detections into DeepSORT improved its performance. The unique dataset was used to fine-tune YOLO via transfer learning. Moreover, throughout the testing and training phases, DeepSORT was initiated using pre-trained weights. A number of performance metrics were calculated, including the tracking accuracy, recall, precision, and F1-score. The system's robustness was assessed using a battery of experiments executed in a range of difficult conditions, such as occlusions, scale fluctuations, and crowded backgrounds. Using this method, we were able to create a robust object recognition and tracking system with modifications that make it suitable for use in real-world applications including robotics, surveillance, and traffic monitoring. A number of difficult settings were successfully navigated by combining YOLO, a tool for precise object detection, with DeepSORT, a tool for keeping item IDs over frames.

One of the most crucial jobs in computer vision is object detection, which entails finding and identifying

objects inside picture or video frames. Its principal goal is to correctly ascertain the locations of things and to classify them into the given categories or groupings, in addition to merely checking their presence. Robotics, autonomous vehicles, security systems, and image analysis are just a few of the many fields that rely on this skill. Normal procedure dictates that important features from the input image be extracted first in the object detection process. Things may be detected using these characteristics. After that, it sorts the objects into their proper categories after localising them by drawing bounding boxes around them. Thanks to new object identification algorithms like Convolutional Neural Networks (CNNs), a deep learning model, and other developing technologies, object detection has become much faster and more accurate. Common designs used in modern object identification systems include YOLO, Faster R-CNN, and SSD.

The task of object tracking is monitoring the movement of an item across a multi-frame video sequence. Not to be confused with related but separate computer vision problems. The ability to identify items in several frames is what distinguishes object detection from object tracking, which keeps item identifications over time. Improving our knowledge of object motion and interactions is crucial for many robotics, autonomous driving, video analysis, and surveillance applications. A common starting point for object tracking initialization is the use of bounding boxes in the initial frame of the video series to identify and track objects of interest. In order to make sure the tracking is done correctly, this is done. In the frames that follow, we update the bounding box based on our best guess as to where the item will go. The object's past motion provides the basis for this. No matter how many objects or occlusions there are in a given frame, data association techniques will always assign the correct tracks to them. Various algorithms are employed, including correlation filters, Kalman filters, particle filters, and deep learning-based trackers like DeepSORT, to develop effective tracking techniques. Figure 3 shows that filters play a significant role in YOLO and DeepSORT's ability to improve object detection and tracking system efficiency. Included in this category are filters such as non-maximum suppression, which removes redundant detections, anchor boxes, which predict bounding boxes, tracking filters, which keep object identities across frames, confidence thresholds, which specify object classes, and Kalman filters, which estimate object positions and velocities accurately, particularly in noisy data. Careful consideration is paid to both the selection and design of these filters to guarantee efficient and accurate item tracking and recognition. Each of these trackers has its own unique reaction time when faced with obstacles like occlusions, appearance changes, and complex motion patterns. While object detection focusses on identifying and classifying things inside individual frames, object tracking aims to preserve item identities over several frames in a video sequence. Both of these activities are fundamental to computer vision because they supply data that is vital for comprehending object behaviour and interaction in various settings.

Results and Discussion

Figure 3 shows the results of using the most recent YOLOv8 object detector to identify items in each frame of the video clip. The detections were subsequently sent into the DeepSORT algorithm, which uses frame-by-frame detection correlation to generate object tracks. Several plausible situations are included in the dataset that was utilised for the research.

Figure 3: The detection of the objects using yolov8

With a detection accuracy of 90% on the test dataset, YOLOv8 proved to be rather effective (Table 1). The outstanding results demonstrate YOLLOv8's object detection capabilities in many scenarios and with a variety of item kinds (Figure 4 demonstrates equipment detection in an image) and video inputs (Figures 4 and 5). Using YOLOv8 detections in tandem allowed the DeepSORT tracker to show trustworthy tracking with an accuracy of 87%. Thanks to robust association capabilities, object identification can be preserved even when objects are concealed or move outside of the frame. The excellent detection and tracking accuracy show that the YOLOv8 and DeepSORT combination can give a full solution for detection and tracking. The research shows that the performance is state-of-the-art, with an accuracy rate of 87% in tracking and 90% in detection. In actual films, they serve as a high-performance pippeline to locate and follow things, such as traffic monitoring via video surveillance (Figure 6) and vehicle counting (Figure 4). Results from these tests demonstrated that the YOLO and Deepp SORT frameworks might be useful for tasks such as vehicle recognition and monitoring several items in a single frame in real time. It is still challenging to handle smaller, farther-off objects and to preserve trajectories over substantial occlusions.

Figure 4: The use of DeepSORT for the tracking of vehicles

Figure 5: The monitoring of traffic

Our object recognition and tracking system worked exceptionally well in a range of scenarios, such as real and recorded films for traffic analysis and on object detectors for numerous use cases. The outcomes of these scenarios have implications for how effectively it will function in real-world circumstances.

High Accuracy

The accuracy metrics that were attained for object detection and tracking are a demonstration of the system's capability. With a detection accuracy rate of 90% and a tracking accuracy rate of 87%, the system was able to find objects inside video frames in a reliable and exact manner.

Table 1: An examination of the system that is being proposed from a quantitative perspective

Video	Total Frames	Accuracy	Precision	Recall
1	812	0.906	0.982	0.98
2	930	0.812	0.883	0.92
3	1160	0.753	0.823	0.852
4	835	0.872	0.932	0.974
5	590	0.444	0.514	0.889

Robust Tracking

Throughout the course of numerous frames, the tracking technology that we utilised, DeepSORT, demonstrated a remarkable capacity for maintaining object IDs. It was able to manage occlusions, size discrepancies, and complicated object motions with ease, which made it an alternative that was suitable for usage in actual applications.

Adaptability

The system's versatility and ability to handle a broad range of application domains are shown by its ability to adapt to different situations, such as traffic monitoring and surveillance. Our experimental results show that the YOLO and DeepSORT-based method effectively detects and tracks objects in a variety of real-world settings with high precision. In Table 1 we can see how well the video-object tracker is doing. Precision shows how often it tracked that one object exclusively, while accuracy shows how often it found the right thing. If you look at videos 1–5, you'll see that this trained modeel's accuracy drops as the frame rate goes up or down. This indicates that training on a bigger dataset to reduce detection and tracking complexity can enhance this modeel's performance. Its adaptability and durability make it a top pick for robotics, traffic monitoring, and other similar applications. To get a high degree of accuracy in object detection and tracking, a comprehensive approach is required, which comprises data preparation, model customisation, tracking techniques, post-processing, and intense evaluation. To maintain the target level of accuracy while enhancing performance, continuous fine-tuning, dynamic resolution, and testing are necessary. When applied together, these techniques provide reliable object tracking and recognition in a broad range of real-world scenarios.

Conclusion

The ever-increasing need for intelligent monitoring in surroundings that are both busy and dynamic has brought to light the necessity of multi-object detection and tracking systems that are not only reliable and efficient but also operate in real time. Through the integration of the YOLO (You Only Look Once) object identification algorithm and the DeepSORT tracking algorithm, this research gave a comprehensive solution to the problem. A number of the most significant difficulties associated with real-time surveillance are efficiently addressed by the integrated framework. These difficulties include occlusions, high object density, and identity retention across frames. The fact that YOLO was able to recognise objects with a high degree of speed and precision made it an excellent candidate for processing continuous video feeds. Through its capacity to correlate object identities through the use of motion and appearance data, DeepSORT was able to guarantee the consistent tracking of many objects, even in scenarios that were somewhat complicated. Together, the system was able to provide a dependable and scalable method to automated monitoring. It was also able to improve situational awareness and provide assistance for proactive decision-making in public locations that were either high-risk or high-traffic. The model was shown to be successful at keeping object identity, handling re-identification, and minimising ID switches under a variety of lighting, motion, and density settings, as demonstrated by experimental assessments. The properties of the system, which include low-latency and high-accuracy, make it a great candidate for deployment in the real world, particularly in smart city infrastructures, transit hubs, and event locations. Despite the fact that the YOLO + DeepSORT framework offers a strong basis, there is need for further development in the areas of including more sophisticated re-identification models, utilising transformer-based architectures, and incorporating

contextual scene awareness. Enhancing mobility and decreasing dependency on centralised computing resources are two additional benefits that may be achieved by optimising the system for edge devices. In conclusion, the combination of YOLO with DeepSORT provides a solution that is both feasible and effective for real-time video monitoring in busy areas. This method paves the way for urban security systems that are smarter, safer, and more responsive.

References

[1] Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). *Simple online and realtime tracking*. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 3464–3468). IEEE. https://doi.org/10.1109/ICIP.2016.7533003

[2] 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 (CVPR), 779–788.

[3] Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. arXiv preprint arXiv:1804.02767. https://arxiv.org/abs/1804.02767

[4] Wojke, N., Bewley, A., & Paulus, D. (2017). *Simple Online and Realtime Tracking with a Deep Association Metric*. In 2017 IEEE International Conference on Image Processing (ICIP), 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

[5] Chen, C., Chen, S., Chen, C., & Wang, W. (2018). *YOLO-based real-time pedestrian detection and tracking system*. In 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 1–2.

[6] Yilmaz, Y., & Aksu, Y. (2020). *Real-time multiple object tracking in video surveillance using YOLO and DeepSORT*. In 2020 28th Signal Processing and Communications Applications Conference (SIU), 1–4. IEEE.

[7] Nagrath, P., Arora, R., Singhal, A., Kumar, A., & Bhatia, P. K. (2021). *Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLOv3 and DeepSORT techniques*. *Multimedia Tools and Applications*, 80(13), 19963–19977. https://doi.org/10.1007/s11042-021-10683-w

[8] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). *Mot16: A benchmark for multi-object tracking*. arXiv preprint arXiv:1603.00831. https://arxiv.org/ abs/1603.00831

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". 779-788. 10.1109/CVPR.2016.91 (2016).

[10] Jacob Solawetz and Francesco. "What is yolov8? the ultimate guide", 2023.

[11] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, Ben Upcroft, "Simple Online and Realtime Tracking", arXiv:1602.00763v2 [cs.CV],(2017).

[12] Nicolai Wojke, Alex Bewley, Dietrich Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric", arXiv:1703.07402v1 [cs.CV]

[13] Dillon Reis, Jordan Kupec, Jacqueline Hong, Ahmad Daoudi, "Real-Time Flying Object Detection with YOLOv8", arXiv:2305.09972v1 [cs.CV] 17 May 2023

[14] Chinthakindi Kiran Kumar, Kirti Rawal, "A Brief Study on Object Detection and Tracking", J. Phys.: Conf. Ser. 2327 012012, 2022