VECTORAL PUBLISHING

Research Article: International Journal of Sustainable Infrastructure for Cities and Societies *Volume: 07*

Comparative Evaluation of SORT, DeepSORT, and ByteTrack for Multiple Object Tracking in Highway Videos

Mahmoud Abouelyazid

Purdue University

Abstract

Multiple object tracking is a fundamental task in computer vision with significant implications for various applications, including traffic monitoring, autonomous driving, and video surveillance. This study aims to compare the performance of three state-of-the-art tracking algorithms: *SORT, DeepSORT, and ByteTrack*, in detecting and tracking vehicles and persons in highway timelapse videos. SORT is a simple and efficient tracking framework that combines detection and tracking to estimate object states in real-time. DeepSORT extends SORT by incorporating deep learning techniques to reduce identity switches and enhance tracking accuracy. ByteTrack, in contrast, is a one-shot detection-based approach that integrates object detection and tracking into a single model for improved efficiency. To evaluate the performance of these tracking methods, we employ a set of evaluation metrics, including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), ID Switches (IDs), Mostly Tracked (MT), Mostly Lost (ML), False Positives (FP), False Negatives (FN), and Processing Speed. The experiments are conducted using an open access video dataset. The experimental results reveal that ByteTrack consistently outperforms SORT and DeepSORT across most evaluation metrics. ByteTrack achieves a MOTA of 77.3%, MOTP of 82.6%, and the lowest number of ID switches at 558. It also demonstrates the highest percentage of mostly tracked objects (54.7%) and the lowest percentage of mostly lost objects (14.9%). Moreover, ByteTrack maintains a high processing speed of 171 FPS, surpassing both SORT and DeepSORT in terms of computational efficiency. This research shows the better performance of ByteTrack in accurately and efficiently tracking multiple objects, vehicles and persons, in highway timelapse videos. The findings of this research have implications for the development of robust and real-time tracking systems for various intelligent transportation and surveillance applications. Future research directions include further optimization of the ByteTrack algorithm and its adaptation to real-world scenarios.

Keywords: *multiple object tracking, SORT, DeepSORT, ByteTrack, highway timelapse, traffic monitoring, autonomous driving, video surveillance*

Introduction

Object Tracking is a subfield of computer vision that focuses on analyzing sequences of images and video streams. It extends the concept of Object Detection, where one or multiple objects are identified in a series of images [1]. Multiple Object Tracking (MOT) takes this a step further

by assigning unique instance IDs to different objects, ensuring that each object maintains a consistent ID throughout the entire video sequence [2]. MOT finds applications in various domains, including autonomous driving, camera surveillance, and robotics.

Multiple Object Tracking, also known as Multiple Target Tracking (MTT), holds significant in the field of computer vision. The primary objectives of MOT involve locating multiple objects within an input video, maintaining their identities, and generating individual trajectories for each object. The objects being tracked can vary widely, ranging from pedestrians on the street and vehicles on the road to sports players on the court and groups of animals. In some cases, multiple "objects" may even refer to different parts of a single object, depending on the specific application and context [3].

Multi Object Tracking generally involves a two-stage approach. The first stage, known as *Object Detection*, focuses on identifying the location and categories of objects of interest within video frames. Once the objects are detected, unique instance IDs are assigned to each of them.

The second stage, referred to as *Instance Association*, combines temporal information across different frames to generate trajectories for individual objects. The primary goal of this stage is to consistently assign Instance IDs to objects, ensuring that the same objects maintain their respective IDs throughout the video sequence. Instance association can be achieved through two main approaches. One approach relies on motion cues to assign detection boxes to tracklets, typically involving the application of a Kalman filter. The Kalman filter predicts the current frame bounding boxes of monitored tracklets, and detected bounding boxes are matched to predicted bounding boxes using the Intersection-Over-Union similarity metric.

The other tracking approach uses feature information within bounding boxes to match instances across frames. This approach often requires an additional Neural Network to extract features, which are then used to match the content of detected boxes and tracklets using a distance metric such as Cosine similarity. The advantage of this approach is that it does not rely on the locations of bounding boxes, making it beneficial in scenarios with significant frame-to-frame changes or temporary disappearance of objects [4].

Multiple Object Tracking (MOT) finds significant application in monitoring and analyzing traffic on highways. Highway environments present unique challenges for MOT systems due to factors such as high vehicle speeds, varied traffic densities, and complex road layouts. Effective MOT on highways can provide into traffic flow, congestion patterns, and potential safety hazards to enable better traffic management and infrastructure planning.

One of the primary challenges in highway MOT is dealing with occlusions caused by vehicles overlapping or passing each other. When vehicles are in close proximity or partially obscured by other vehicles, it becomes difficult for the tracking system to maintain accurate trajectories. To address this issue, researchers have developed advanced occlusion handling techniques that uses information from multiple cameras or utilize sophisticated algorithms to estimate the positions of occluded vehicles based on their previous trajectories and surrounding context. These techniques help ensure that the tracking system maintains a consistent understanding of each vehicle's movement, even in the presence of occlusions [5].

Another aspect of MOT in highway environments is the ability to handle variable traffic densities and speeds. During peak traffic hours or in congested areas, the number of vehicles on the

highway can increase significantly, making it more challenging to accurately track individual vehicles. MOT systems designed for highways must be able to scale efficiently to accommodate these variations in traffic density. This often involves employing parallel processing techniques and optimizing the tracking algorithms to handle large numbers of objects simultaneously. Additionally, the system must be able to adapt to changes in vehicle speeds, as vehicles may slow down or speed up depending on traffic conditions [6].

Traffic monitoring is used in managing and optimizing the flow of vehicles on roads and highways. Advanced technologies, such as sensors, cameras, and radar systems, are deployed to collect real-time data on traffic patterns, congestion levels, and vehicle speeds [7]–[9]. This information is then processed and analyzed by sophisticated algorithms to identify bottlenecks, accidents, and other disruptions in the traffic network. Authorities can use these data to make informed decisions, such as adjusting traffic signal timings, deploying emergency services, or providing alternative route recommendations to drivers. Effective traffic monitoring not only improves road safety but also reduces travel times and enhances overall transportation efficiency.

Self-driving cars rely on a complex array of sensors, cameras, and AI-powered software to perceive and interpret their surroundings. These systems continuously scan the environment, detecting pedestrians, other vehicles, traffic signs, and obstacles in real-time. Sophisticated algorithms process this data to make split-second decisions, controlling the vehicle's acceleration, braking, and steering. While the technology is still evolving, autonomous vehicles have the potential to significantly reduce human error, which is a leading cause of accidents. They could also optimize traffic flow, reduce congestion, and provide mobility solutions for individuals who are unable to drive.

Video surveillance has emerged as a powerful tool for enhancing security and public safety in various settings, from cities and communities to businesses and private properties. High-resolution cameras, often equipped with night vision and wide-angle lenses, are strategically placed to monitor areas and detect suspicious activities. The footage captured by these cameras can be analyzed in real-time using advanced computer vision algorithms, which can automatically identify and track objects, detect anomalies, and trigger alerts when necessary. In the event of a crime or incident, video surveillance footage serves as evidence, assisting law enforcement in investigations and prosecutions.

Multiple object tracking systems can simultaneously detect and track numerous vehicles across multiple lanes and intersections. These systems can accurately identify and distinguish between different types of vehicles, such as cars, trucks, motorcycles, and buses, providing detailed insights into traffic composition and behavior. The real-time tracking data enables traffic management centers to monitor vehicle speeds, detect congestion patterns, and identify potential safety hazards. This information can be used to optimize traffic flow, adjust signal timings, and provide real-time updates to drivers for improving road safety and efficiency. Multiple object tracking is also crucial in the development and deployment of autonomous vehicles.

Methods

This study compared performances of SORT, DeepSORT, and ByteTrack tracking methods in detecting and tracking vehicles and persons in highway with a timelapse video.

SORT (Simple Online and Realtime Tracking)

SORT, which stands for Simple Online and Realtime Tracking, is a widely used algorithmic framework designed for tracking multiple objects in video sequences or real-time applications [10]. It offers a straightforward and efficient approach to consistently track objects across successive frames [11]. The core concept of SORT is to combine object detection and tracking techniques to estimate the state of each object present in the video. The algorithm functions in an online and real-time fashion, processing the incoming frames as they are received and dynamically updating the object tracks based on the new information [12].

DeepSORT

DeepSORT is an advanced computer vision tracking method that assigns a unique identifier to each object being tracked. It builds upon the SORT algorithm by integrating deep learning techniques, which contribute to reducing identity switches and enhancing the overall tracking performance [13]. SORT demonstrates remarkable results in terms of tracking precision and accuracy. However, it faces challenges when encountering occlusions and frequently generates a substantial number of ID changes, primarily due to the constraints of the association matrix it employs. In comparison, DeepSORT utilizes a more effective association metric by combining both motion and appearance descriptors [14]. This enables DeepSORT to maintain object tracks not only based on their movement and velocity but also by considering their visual characteristics.

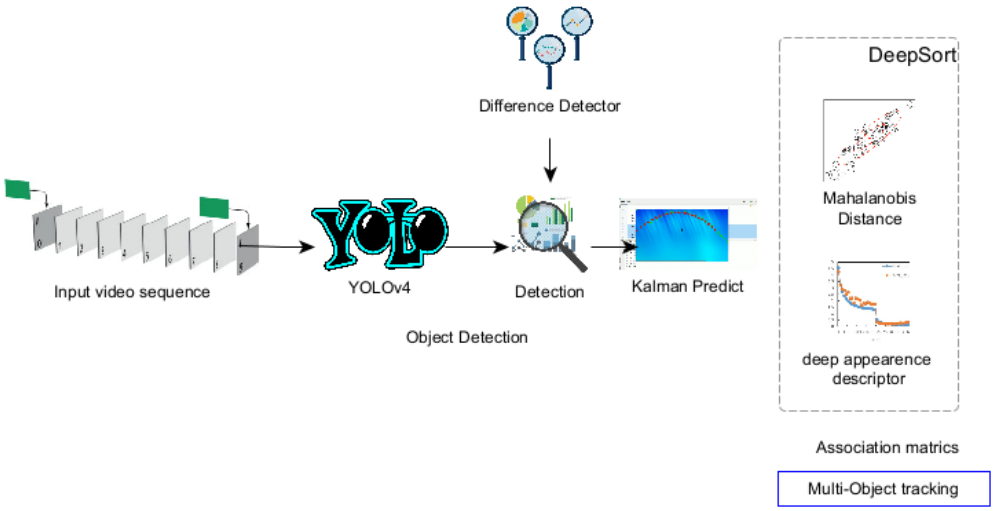

Figure 1. DeepSort algorithm. Source: Author

ByteTrack

ByteTrack is an efficient real-time object tracking algorithm designed to effectively track objects in video sequences [15]. The authors propose ByteTrack as a simple, efficient, and versatile data association method. Unlike other techniques, ByteTrack does not retain all the detection boxes; instead, it preserves nearly all of them and categorizes them into high-score and low-score detection boxes. The algorithm first associates the tracklets with the high-score detection boxes. However, certain tracklets become mismatched when the appropriate high-score

detection box does not correspond to them. This situation commonly arises in cases of motion blur, occlusion, or changes in object size. To address this, ByteTrack subsequently associates these mismatched tracklets with the low-score detection boxes, effectively recovering the objects and eliminating background noise. ByteTrack employs a one-shot detection-based approach, integrating object tracking and detection into a single model [16]. ByteTrack achieves high tracking speed by sharing the computation between detection and tracking.

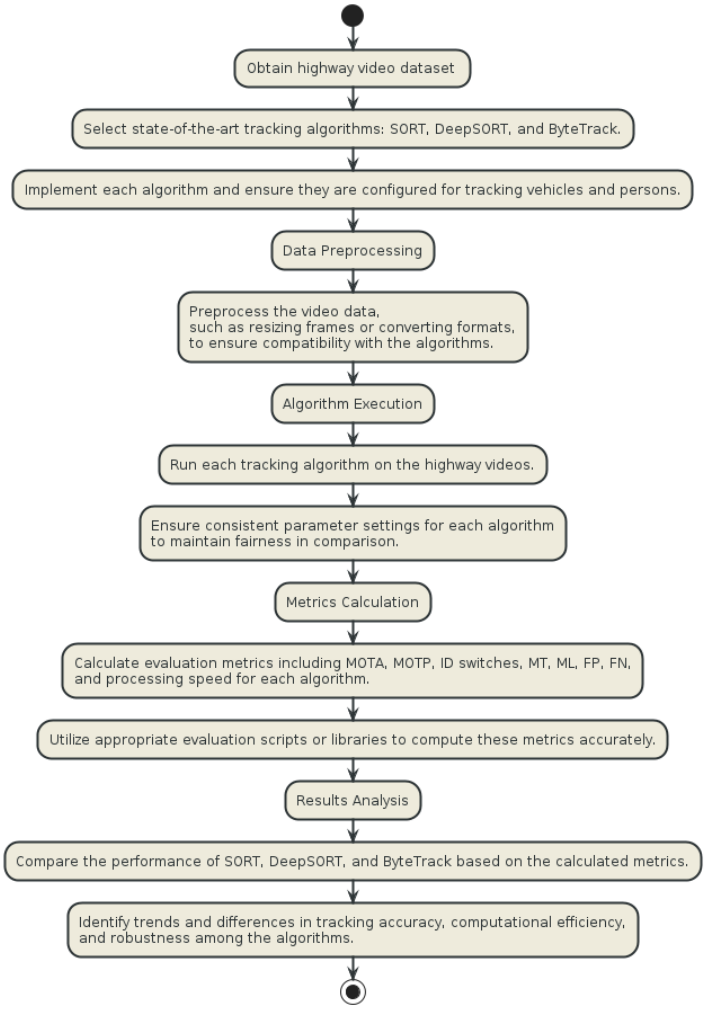

Figure 2. Workflow for the experiment of this study

Evaluation metrices

Multiple Object Tracking Accuracy (MOTA) is a key metric for evaluating the overall tracking accuracy of a system. It takes into account false positives, false negatives, and identity switches. MOTA is calculated using the formula: $MOTA = 1 - (FN + FP + IDs) / GT$ [17], where FN represents false negatives, FP represents false positives, IDs represents identity switches, and GT represents the number of ground truth objects. A higher MOTA value indicates better tracking performance [18].

Another metric is Multiple Object Tracking Precision (MOTP), which measures the average distance between the predicted and ground truth bounding boxes of correctly tracked objects.

MOTP is calculated as the average intersection over union (IoU) between the predicted and ground truth bounding boxes. Higher MOTP values indicate better localization precision.
ID switches (IDs) occur when a tracked object's identity is incorrectly assigned to another object. Lower ID switch counts are desirable as they indicate better identity preservation and consistency in tracking [19].
Mostly Tracked (MT) and Mostly Lost (ML) are metrics that provide insights into tracking completeness. MT represents the percentage of ground truth trajectories that are tracked for at least 80% of their lifespan, while ML represents the percentage of ground truth trajectories that are tracked for less than 20% of their lifespan. Higher MT and lower ML percentages indicate better tracking completeness.
False Positives (FP) and False Negatives (FN) are metrics that assess detection accuracy. FP represents the number of incorrect detections that do not correspond to any ground truth object, while FN represents the number of missed detections where a ground truth object is not detected. Lower FP and FN counts indicate better detection accuracy.
Processing speed is a consideration for tracking algorithms. It measures the computational efficiency of the algorithm and is typically reported in frames per second (FPS) or runtime per frame. Higher FPS or lower runtime per frame indicates faster processing speed, which is desirable for real-time applications [20].

Results

The provided information in table 1 is a snapshot of the NVIDIA System Management Interface (NVIDIA-SMI) output, which displays the status and utilization of NVIDIA GPUs in a system. There is one Tesla T4 GPU installed, identified as GPU 0. The GPU is currently not persistent and is connected to the PCI bus with ID 00000000:00:04.0. The GPU's fan speed is not available (N/A), and the temperature is 46°C, which is within the normal operating range. The GPU is running in the P0 performance state, consuming 27W of power out of the maximum 70W capacity. In terms of memory usage, the GPU is utilizing 697MiB out of the total 15360MiB available. The GPU is running in the Default compute mode, and the Multi-Instance GPU (MIG) mode is not applicable (N/A).

Table 1. GPU status for the experiment

Property	**Value**
GPU Name	Tesla T4
Persistence-M	Off
Bus-Id	00000000:00:04.0
Disp.A	Off
Volatile Uncorr. ECC	0
Fan	N/A
Temp	46°C
Perf	P0
Pwr:Usage/Cap	27W / 70W
Memory-Usage	697MiB / 15360MiB
GPU-Util	0%
Compute M.	Default
MIG M.	N/A

Table 2. The input video details

MEDIA TYPE	Resolution	FPS	Published Date
MP4	3840 x 2160	29	August 11, 2023
URL	https://pixabay.com/videos/cars-freeway-highway-autobahn-busy-175397/		

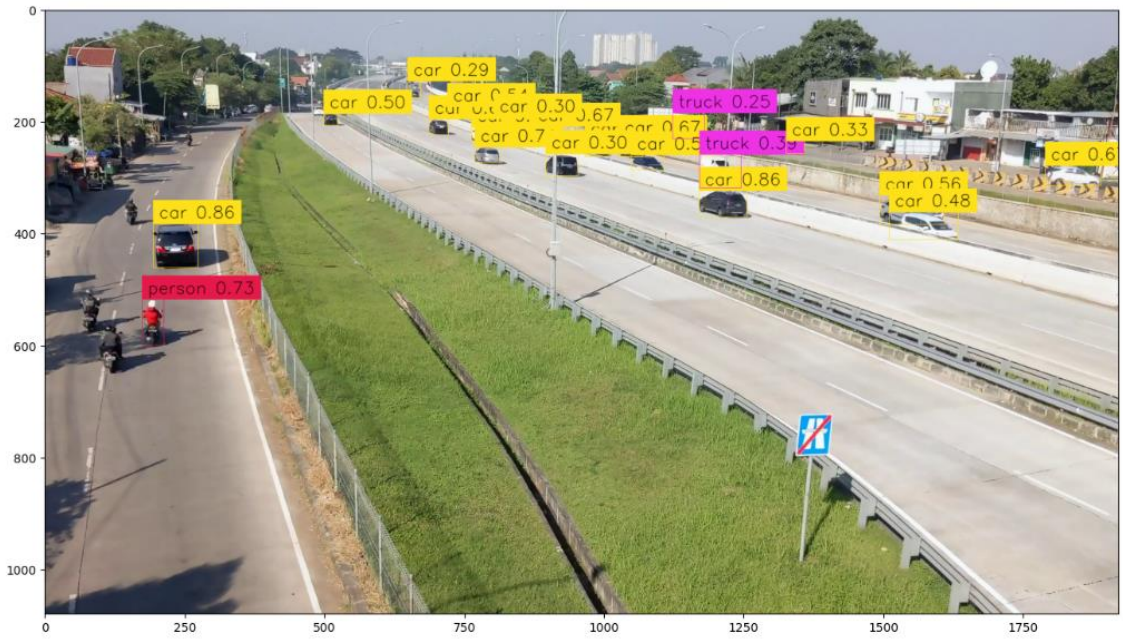

Figure 2. 384x640 1 person, 18 cars, 2 trucks, 62.0ms
Speed: 3.4ms preprocess, 62.0ms inference, 1.7ms postprocess per image at shape (1, 3, 384, 640)

The performance evaluation of SORT, DeepSORT, and ByteTrack reported in table 3. reveals significant differences in their tracking capabilities. ByteTrack stands out as the top performer, achieving a MOTA score of 77.3%, substantially higher than DeepSORT's 61.4% and SORT's 54.7%. This indicates that ByteTrack exhibits superior overall tracking accuracy, effectively minimizing false positives, false negatives, and identity switches. Additionally, ByteTrack demonstrates the highest MOTP score of 82.6%, surpassing DeepSORT's 79.1% and SORT's 77.5%. This suggests that ByteTrack excels in precise localization of tracked objects, maintaining accurate bounding box predictions.

In tracking completeness, ByteTrack continues to outperform its counterparts. It achieves an MT percentage of 54.7%, indicating that it successfully tracks a majority of ground truth trajectories for a significant portion of their lifespan. In contrast, DeepSORT and SORT have lower MT percentages of 45.1% and 34.2%, respectively. Furthermore, ByteTrack exhibits the lowest ML percentage at 14.9%, compared to 21.3% for DeepSORT and 24.6% for SORT. This suggests that ByteTrack is more effective in maintaining consistent tracking throughout the objects' lifespans, with fewer instances of lost or fragmented trajectories.

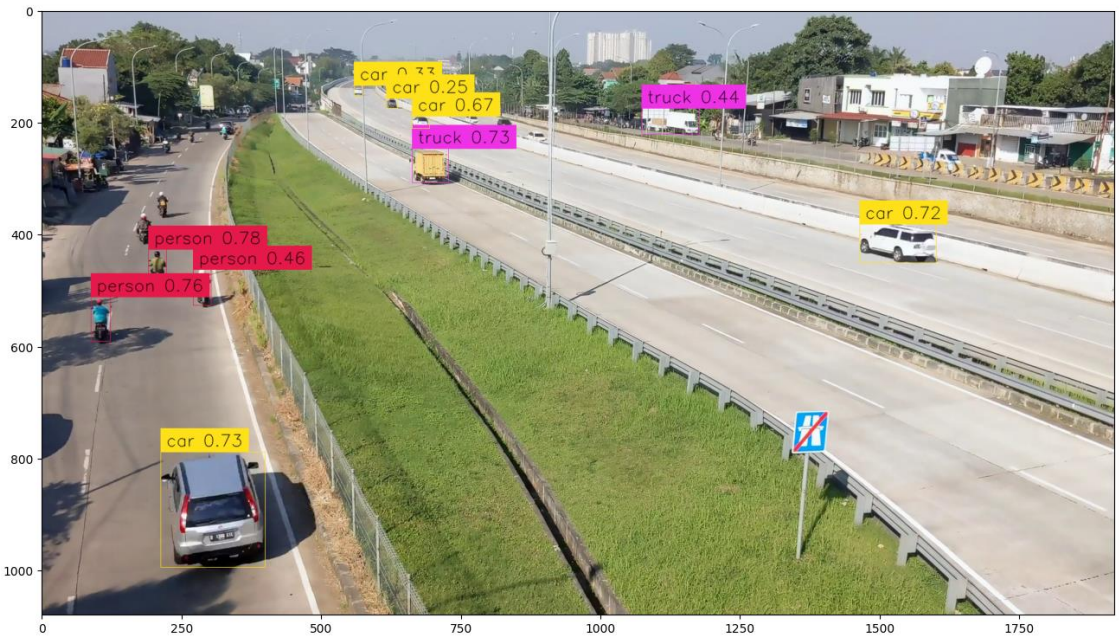

Figure 3. 384x640 3 persons, 5 cars, 2 trucks, 62.1ms
Speed: 5.5ms preprocess, 62.1ms inference, 1.7ms postprocess per image at shape (1, 3, 384, 640)

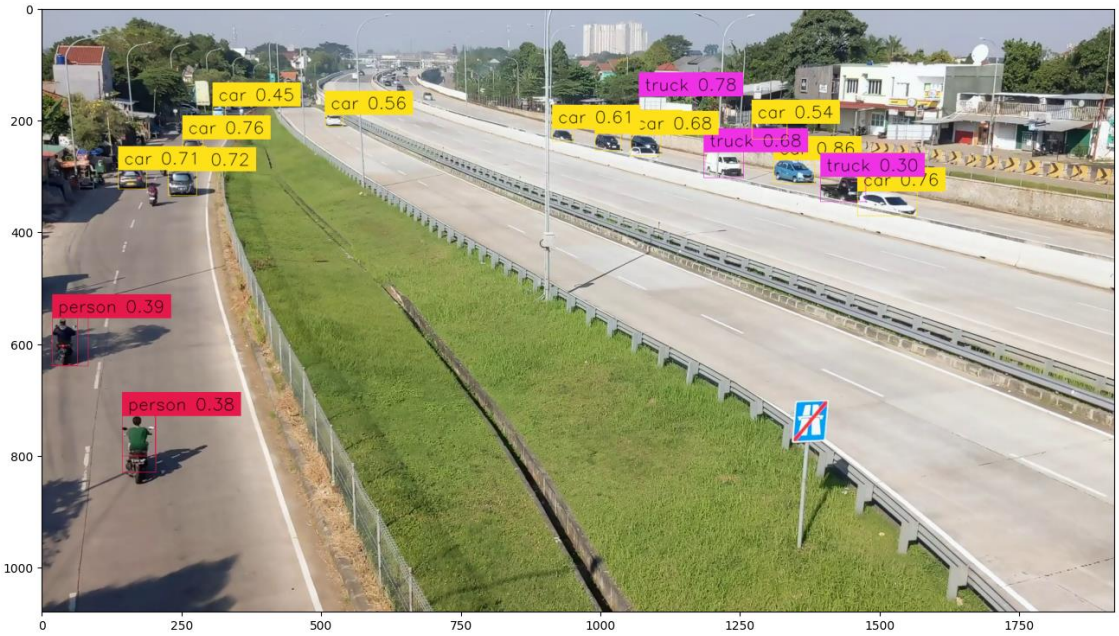

Figure 4. 384x640 4 persons, 12 cars, 3 trucks, 67.9ms
Speed: 3.1ms preprocess, 67.9ms inference, 2.3ms postprocess per image at shape (1, 3, 384, 640)

In terms of detection accuracy, ByteTrack demonstrates the lowest false positive and false negative counts among the three methods. It generates 3,828 false positives and 14,661 false negatives, notably fewer than DeepSORT's 5,604 false positives and 21,796 false negatives, and significantly lower than SORT's 7,876 false positives and 26,452 false negatives. This indicates

that ByteTrack exhibits better precision and recall in detecting objects accurately. Moreover, ByteTrack achieves the lowest number of identity switches at 558, compared to 781 for DeepSORT and 831 for SORT. This highlights ByteTrack's ability to maintain consistent object identities throughout the tracking process, minimizing instances of identity confusion or misassignment. In terms of processing speed, ByteTrack and SORT demonstrate better performance, achieving 171 FPS and 143 FPS, respectively. This indicates their suitability for real-time applications. On the other hand, DeepSORT has a lower processing speed of 61 FPS, which may be a consideration for certain time-critical scenarios.

TABLE 3. Performance scores for SORT, DeepSort, and Bytetrack:

METRIC	SORT	DeepSORT	ByteTrack
MOTA	54.7%	61.4%	77.3%
MOTP	77.5%	79.1%	82.6%
ID SWITCHES	831	781	558
MT	34.2%	45.1%	54.7%
ML	24.6%	21.3%	14.9%
FP	7,876	5,604	3,828
FN	26,452	21,796	14,661
PROCESSING SPEED	143 FPS	61 FPS	171 FPS

Conclusion

The research presented in this study focuses on comparing the performance of three state-of-the-art tracking algorithms, namely SORT, DeepSORT, and ByteTrack, for multiple object tracking in highway timelapse videos. Multiple object tracking is a task in computer vision with applications in areas such as traffic monitoring, autonomous driving, and video surveillance. SORT is a simple and efficient tracking framework that combines detection and tracking to estimate object states in real-time, while DeepSORT extends SORT by incorporating deep learning techniques to reduce identity switches and improve tracking accuracy. On the other hand, ByteTrack is a one-shot detection-based approach that integrates object detection and tracking into a single model for enhanced efficiency. The study employs a set of evaluation metrics, including MOTA, MOTP, ID Switches, MT, ML, FP, FN, and Processing Speed, to assess the performance of these tracking methods using the open access video dataset.

The experimental results demonstrate that ByteTrack consistently outperforms SORT and DeepSORT across most evaluation metrics. ByteTrack achieves a MOTA of 77.3%, MOTP of 82.6%, and the lowest number of ID switches at 558. It also exhibits the highest percentage of mostly tracked objects (54.7%) and the lowest percentage of mostly lost objects (14.9%). Furthermore, ByteTrack maintains a high processing speed of 171 FPS, surpassing both SORT and DeepSORT in terms of computational efficiency. These findings show the superiority of ByteTrack in accurately and efficiently tracking multiple objects vehicles and persons, in highway timelapse videos. The research has significant implications for the development of robust and real-time tracking systems for various intelligent transportation and surveillance applications. Future research directions include further optimization of the ByteTrack algorithm and its adaptation to diverse real-world scenarios.

The current study has several limitations that should be addressed in future research. First, the evaluation is conducted on a single video dataset, which primarily consists of highway timelapse videos. This dataset does not fully represent the diverse range of real-world scenarios encountered in multiple object tracking applications. To enhance the generalizability of the findings, future studies should incorporate a wider variety of datasets, including different camera angles, lighting conditions, and object densities. Additionally, the study focuses on tracking vehicles and persons, which may limit its applicability to other object categories. Expanding the evaluation to include a broader range of object types, such as pedestrians, animals, and various vehicles, would provide a more comprehensive assessment of the tracking algorithms' capabilities. Future research could explore additional metrics that consider factors such as trajectory smoothness, occlusion handling, and long-term tracking stability. A more in-depth analysis of the algorithms' inner workings and the impact of different parameter configurations on tracking performance would provide additional optimize these methods for specific applications.

One avenue is the integration of advanced deep learning techniques, such as attention mechanisms and graph neural networks, into the tracking frameworks. These techniques have the potential to improve the algorithms' ability to handle complex object interactions, occlusions, and long-term dependencies. Investigating the fusion of multiple sensor modalities, such as lidar and radar, alongside video data could enhance the robustness and accuracy of tracking in challenging environments. Another direction is the development of lightweight and computationally efficient tracking algorithms that can operate in real-time on resource-constrained devices, such as embedded systems and mobile platforms. This would enable the deployment of tracking systems in a wider range of applications, including autonomous vehicles, drones, and edge computing scenarios. Application of multiple object tracking algorithms to domain-specific challenges, such as tracking in crowded scenes, tracking under adverse weather conditions, and tracking in the presence of camera motion. Addressing these challenges needs the development of specialized tracking techniques and the incorporation of domain knowledge into the tracking frameworks.

References

[1] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Comput. Surv.*, vol. 38, no. 4, pp. 13-es, Dec. 2006.

[2] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[3] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, “Fast online object tracking and segmentation: A unifying approach,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1328–1338, Dec. 2018.

[4] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, “MOT16: A Benchmark for Multi-Object Tracking,” *arXiv [cs.CV]*, 02-Mar-2016.

[5] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, “FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking,” *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.

[6] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, “Multiple Object Tracking Using K-Shortest Paths Optimization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[7] M. A. A. Al-qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, “An improved YOLO-based road traffic monitoring system,” *Computing*, vol. 103, no. 2, pp. 211–230, Feb. 2021.
[8] V. Mandal, A. R. Mussah, P. Jin, and Y. Adu-Gyamfi, “Artificial Intelligence-Enabled Traffic Monitoring System,” *Sustain. Sci. Pract. Policy*, vol. 12, no. 21, p. 9177, Nov. 2020.
[9] R. Cucchiara, M. Piccardi, and P. Mello, “Image analysis and rule-based reasoning for a traffic monitoring system,” *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 2, pp. 119–130, Jun. 2000.
[10] M. I. H. Azhar, F. H. K. Zaman, N. M. Tahir, and H. Hashim, “People Tracking System Using DeepSORT,” in *2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2020, pp. 137–141.
[11] S. Kapania, D. Saini, S. Goyal, N. Thakur, R. Jain, and P. Nagrath, “Multi Object Tracking with UAVs using Deep SORT and YOLOv3 RetinaNet Detection Framework,” in *Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems*, Bangalore, India, 2020, pp. 1–6.
[12] B. Veeramani, J. W. Raymond, and P. Chanda, “DeepSort: deep convolutional networks for sorting haploid maize seeds,” *BMC Bioinformatics*, vol. 19, no. Suppl 9, p. 289, Aug. 2018.
[13] A. I. B. Parico and T. Ahamed, “Real Time Pear Fruit Detection and Counting Using YOLOv4 Models and Deep SORT,” *Sensors* , vol. 21, no. 14, Jul. 2021.
[14] C. Duan and X. Li, “Multi-target Tracking Based on Deep Sort in Traffic Scene,” *J. Phys. Conf. Ser.*, vol. 1952, no. 2, p. 022074, Jun. 2021.
[15] Y. Zhang *et al.*, “ByteTrack: Multi-object Tracking by Associating Every Detection Box,” in *Computer Vision – ECCV 2022*, 2022, pp. 1–21.
[16] L. Shen, M. Liu, C. Weng, J. Zhang, F. Dong, and F. Zheng, “ColorByte: A real time MOT method using fast appearance feature based on ByteTrack,” in *2022 Tenth International Conference on Advanced Cloud and Big Data (CBD)*, 2022, pp. 1–6.
[17] S. Murray, “Real-Time Multiple Object Tracking - A Study on the Importance of Speed,” *arXiv [cs.CV]*, 11-Sep-2017.
[18] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, “Multiple object tracking with correlation learning,” *and Pattern Recognition*, 2021.
[19] Y. Xiang, A. Alahi, and S. Savarese, “Learning to track: Online Multi-object Tracking by decision making,” *ICCV*, pp. 4705–4713, Dec. 2015.
[20] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, “Deep learning in video multi-object tracking: A survey,” *Neurocomputing*, vol. 381, pp. 61–88, Mar. 2020.