Journal of Intelligent and Connected Vehicles
2024, 7(2): 129–137
https://doi.org/10.26599/JICV.2023.9210034

Research Article

Roadside cross-camera vehicle tracking combining visual and spatial-temporal information for a cloud control system

Bolin Gao[1], Zhuxin Li[2], Dong Zhang[3], Yanwei Liu[2,✉], Jiaxing Chen[1], Ziyuan Lv[2]

[1]*School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China*
[2]*School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China*
[3]*Department of Mechanical and Aerospace Engineering, Brunel University London, UB8 3PH, UK*

Received: November 23, 2023; Revised: January 17, 2024; Accepted: February 20, 2024

ABSTRACT: Roadside cameras play a crucial role in road traffic, serving as an indispensable part of integrated vehicle-road-cloud systems due to their extensive visibility and monitoring capabilities. Nevertheless, these cameras face challenges in continuously tracking targets across perception domains. To address the issue of tracking vehicles across nonoverlapping perception domains between cameras, we propose a cross-camera vehicle tracking method within a Vehicle–Road–Cloud system that integrates visual and spatiotemporal information. A Gaussian model with microlevel traffic features is trained using vehicle information obtained through online tracking. Finally, the association of vehicle targets is achieved through the Gaussian model combining time and visual feature information. The experimental results indicate that the proposed system demonstrates excellent performance.

KEYWORDS: integrated Vehicle–Road–Cloud, cross-camera, online tracking, intercamera association

1 Introduction

"Vehicle–Road–Cloud" collaboration is a current research hotspot and development trend in the field of intelligent transportation. It is a crucial competitive area in the new round of scientific and technological development and industry growth, holding significant importance for improving traffic safety, alleviating congestion, promoting energy conservation and emission reduction, and driving the development of upstream and downstream industries. Li et al. (2020) reported that roadside perception, as a vital component of collaborative perception in the "Vehicle–Road–Cloud" system, plays an indispensable role in enhancing perception accuracy and expanding the perception range. However, roadside perception is a challenging task that requires overcoming various difficulties and obstacles. Among the challenges faced, tracking moving targets using cameras in complex environments is a crucial aspect of roadside perception.

Currently, target tracking solutions using a single camera are relatively mature. However, cameras on real roads often operate independently, and there is no information exchange between cameras, leading to different identifications of the same target in different cameras. Cross-camera target tracking technology assigns consistent identities to the same target in different camera views, enabling continuous tracking across cameras.

2 Related studies

2.1 Vehicle detection

Liu et al. (2016) introduced the Single Shot MultiBox Detector (SSD), Girshick (2015) proposed the Fast Region with CNN features(R-CNN), and Ren et al. (2017) presented the Faster R-CNN. These models address the challenges associated with identifying and tracking vehicles. Fast R-CNN and Faster R-CNN are two-stage detectors that typically offer higher accuracy and flexibility but at the cost of being more time consuming. On the other hand, You Only Look Once (YOLO), proposed by Redmon et al. (2016), and SSDs are single-stage detectors. SSD uses a set of bounding boxes with different aspect ratios and sizes to predict object classes and their locations. YOLO effectively turns an object detection task into a classification task, achieving a balance between speed and accuracy, which enables real-time operation.

2.2 Single-camera tracking

In most single camera tracking (SCT) methods, the task is typically split into two main steps: First, a detection step is conducted, followed by an association step. During the association step, Bewley et al. (2016), Bochinski et al. (2017, 2018), and Ren et al. (2021) linked the detection of the same targets together based on a similarity measure.

The tracking process can be conducted either offline, for tasks such as traffic analysis, or in real time alongside camera or video input frames.

In offline methods, the model utilizes detection across the entire sequence of frames, followed by global optimizations, including graph-based and hierarchical methods. Standard offline methods often utilize a graph-based model, which can be further improved through techniques such as the minimum cost flow proposed by Wang et al. (2015) and subgraph decomposition proposed by Tang et al. (2015).

On the other hand, online methods follow the tracking-by-detection paradigm, utilizing only the current and previous frames to link detection results for each frame. The primary challenge in online methods is to associate features between tracked objects

✉ Corresponding author.
E-mail: ywliu@gdut.edu.cn

and detection results. To address this issue, processes such as the Kalman filter-based approach proposed by Bewley et al. (2016) can be employed. Cao et al. (2023) introduced Observation-Centric Simple Online and Realtime Tracking (OC-SORT), while Wojke et al. (2017) proposed DeepSORT for object association, leveraging pure motion tracking and deep visual features. Using deep OC-SORT, Maggiolino et al. (2023) proposed introducing visual appearance to OC-SORT, which adaptively integrates appearance matching into existing motion-based high-performance methods.

2.3 Cross-camera object tracking

Recently, the field of multiple camera view tracking (MCVT) proposed by Javed et al. (2003) has gained significant attention, driven by the increasing demands of city-scale traffic management. The core of MCVT lies in cross-camera tracking technology. Presently, methods employed in cross-camera tracking research can be divided into two main categories: image feature-based tracking and motion feature-based tracking. In terms of image features, shortly after Porikli proposed brightness transfer functions, Madden et al. (2007) suggested using the main color spectrum features of targets as the basis for cross-camera target matching. Additionally, Collins et al. (2000), Porikli and Divakaran (2003) and Velipasalar et al. (2008) focused on studying the robustness of UV chromaticity features as features for cross-camera target appearance. Regarding motion features, Javed et al. (2003) used the Parzen window method to estimate target motion parameters within the camera field of view, such as speed, entry and exit positions, and transition time intervals. Dick and Brooks (2004) described human target motion within and between different fields of view using random transformation matrices. Caspi and Irani (2000) and Stein (1999) proposed methods based on temporal consistency between cameras to establish spatiotemporal constraints.

Recent research trends involve the integration of image feature information and motion feature information for collaborative tracking. Tran et al. (2022) proposed a method that combines spatial region partitioning and image features for vehicle association. Building on this approach, Yao et al. (2022) introduced a time decay mechanism, leading to improved tracking performance. However, most of the research on cross-camera tracking has been conducted offline, with relatively less emphasis on online tracking. For instance, Chen et al. (2021) utilized dual-appearance matrices to achieve cross-domain tracking of vehicles in overlapping regions.

3 Research gaps and objectives

Depending on the distribution of camera fields of view, cross-camera tracking research can be classified into overlapping and nonoverlapping scenarios. In nonoverlapping scenarios, current research methods are mainly divided into visual feature matching and spatiotemporal relationship-based association methods. Visual feature matching methods use neural networks to extract image features for target association. However, due to variations in matched angles and the impact of illumination on camera imaging, visual-based methods suffer from poor tracking performance. On the other hand, methods based on spatiotemporal information establish the topological structure between cameras according to the entrances and exits of the cameras, and trajectory similarity is calculated based on the transfer time of targets between cameras. However, these methods place high demands on the target detector, and issues such as detector false positives and false negatives can render the approach ineffective.

Therefore, this study proposes a cross-camera tracking method to address the shortcomings of both spatiotemporal-based and visual-based methods. The primary objectives of this research are as follows:

- A cross-camera online tracking method based on real road scenes is proposed.
- A spatial point matching-based filtering mechanism is proposed to enhance the utilization of spatial information.
- A Gaussian model with microlevel traffic patterns is established to integrate visual features and spatiotemporal information to compensate for their respective shortcomings.

4 Method

The cross-camera vehicle tracking (CCVT) system framework is detailed in this section. As depicted in Fig. 1, the system is divided into two key components: single-camera vehicle tracking and intercamera association. Within single-camera vehicle tracking, three modules are integrated—vehicle detection, vehicle visual feature extraction, and vehicle tracking. Concurrently, CCVT is primarily composed of an intercamera association. The main contribution of this article lies in the intercamera association module.

4.1 Vehicle detection and feature extraction

In single-camera vehicle tracking, the fundamental tasks involve detection and feature extraction, and there are well-established solutions for these tasks. We will follow established methods for generating detection boxes and reidentification (ReID) features in this study.

For the detection task, we employ YOLOv8x1 as our single-stage detector because of its finely tuned equilibrium between speed and accuracy. For the ReID feature extraction task, we use CSP-Darknet-53 as the backbone to obtain robust and discriminative visual feature representations for vehicles. This backbone is pretrained on the COCO dataset.

4.2 Single-camera vehicle tracking

To ensure better real-time performance and accuracy in single-camera tracking, this research adopts Deep OC-SORT as the tracker. The deep OC-SORT algorithm builds upon the OC-SORT algorithm, improving multiobject tracking performance by integrating a novel visual appearance method. Moreover, Deep OC-SORT secured first and second places in the MOT20 and MOT17 competitions, respectively. This also demonstrates the outstanding tracking performance of Deep OC-SORT.

4.3 Intercamera association

Yang et al. (2022) reviewed the intercamera association (ICA), which serves as the final yet crucial module in the CCVT. By leveraging the trajectories produced by the preceding modules, ICA links all tracklets with identical identities based on visual features and spatiotemporal information. It employs two consecutive cameras to match tracklets in accordance with road entry and exit points.

4.3.1 Spatial point matching-based filtering mechanism

To better utilize spatial information, this study proposes a spatial point matching-based filtering mechanism. The filtering

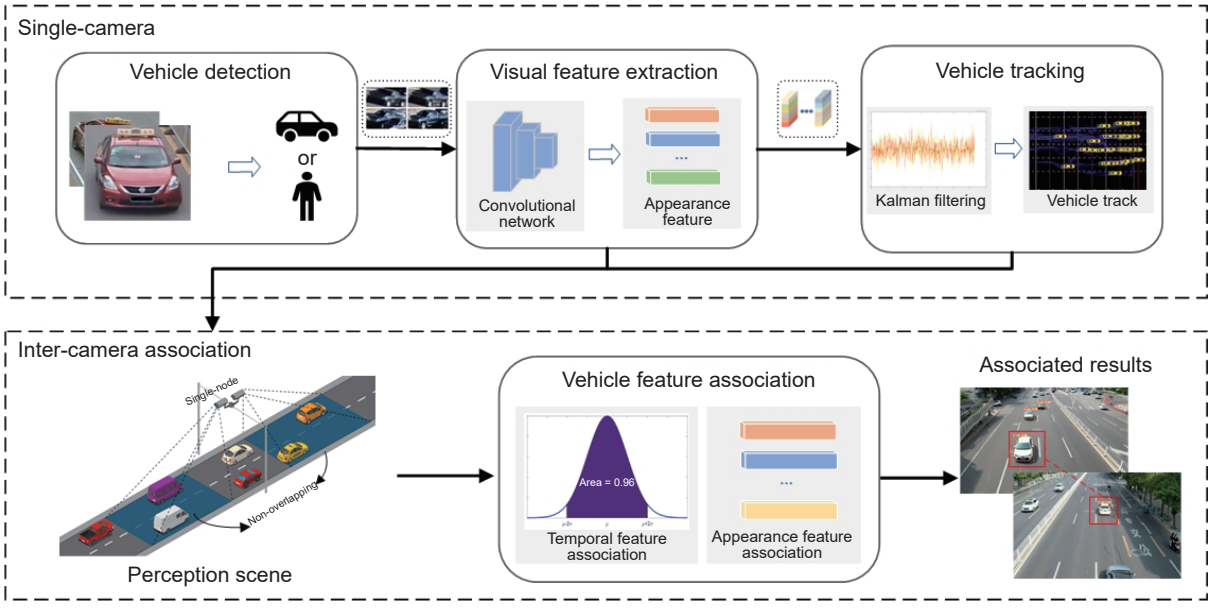

Fig. 1 Scheme of the proposed CCVT system.

mechanism focuses on road surveillance cameras that are arranged with a pair of reverse-direction cameras. This structure enables adjacent cameras to have small blind zones and symmetrical fields of view. The small blind zone between cameras makes it challenging for vehicles to undergo significant transfers within the blind zone. The spatial topological relationships between adjacent cameras are established based on the symmetrical field of view. This means that when a vehicle departs from a certain pixel point in the field of view of the front camera, it will appear at the corresponding pixel point in the field of view of the rear camera. However, considering the tendency of lane changes in the blind zone, the pixel coordinates of the target vehicle when leaving from one camera's field and entering from another camera's field of view may not strictly satisfy the point-to-point matching relationship. Therefore, the spatial searching range needs to be enlarged. Expanding from a single pixel point to an entire region, the size of which is determined by the blind zone between the two cameras, as illustrated in Fig. 2. Under the influence of this filtering mechanism, whenever a vehicle appears in the rear-view camera, we can identify the area where the vehicle disappears in the front-view camera based on the spatial topological relationship. This helps narrow the range for vehicle association.

4.3.2 *Temporal-based vehicle association*

In terms of temporal relationships, this study observes real road scenes and summarizes the time transfer patterns of vehicles between adjacent roadside cameras. The functionality of associating target vehicles is achieved based on this pattern. In real road scenarios, the limited speed of vehicles, constrained by both small blind zones between adjacent cameras and urban road traffic regulations, ensures that the vehicle entering the camera's blind zone first is also the first to exit. Therefore, when a camera detects a new target vehicle, the study associates it with the vehicles leaving adjacent cameras based on the time transfer pattern between cameras.

While the time transfer pattern of vehicles between cameras aligns with most scenarios on real traffic roads, there is still a possibility of a target vehicle being overtaken by other vehicles in the blind zone, leading to association errors. Therefore, the method of target association based on time transfer patterns still has certain limitations.

4.3.3 *Visual feature-based vehicle association*

In addition to the target association method based on time transfer patterns, this study also incorporates the conventional approach of visual feature matching for target association. The visual feature matching method comprises two key components: feature extraction and nearest-neighbor matching. For efficient feature extraction while maintaining real-time performance, this study utilizes the same feature extraction network, CSP-DarkNet, as YOLOv8.

The nearest-neighbor matching component involves assessing the similarity between image feature vectors extracted through the CSP-DarkNet network via cosine distance calculations. The cosine similarity algorithm measures the dissimilarity between two entities by evaluating the cosine value of the angle between the vectors in a vector space, as represented by Eq. (1):

$$\text{Similarity} = \cos\theta = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\ \|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n}\boldsymbol{A}_i\boldsymbol{B}_i}{\sqrt{\sum_{i=1}^{n}\boldsymbol{A}_i^2}\sqrt{\sum_{i=1}^{n}\boldsymbol{B}_i^2}} \tag{1}$$

where $\boldsymbol{A}$ is the visual feature vector of the newly detected target by the rear-view camera, $\boldsymbol{B}$ is the visual feature vector of vehicles after the zone-based target candidate filter. The closer the cosine distance is to 0, the more similar the two feature vectors are.

4.3.4 *Online learning-based Gaussian model*

In response to the limitations of both the target association models based on time transfer patterns and visual feature matching, this study proposes a Gaussian model-based approach utilizing online learning to complement the strengths and address the weaknesses of each individual model. Due to various factors such as traffic signals, speed limits, and regional development affecting real roads, each road embodies specific traffic patterns. This study statistically conducts a dataset of German motorways, and it shows that the speed of vehicles traveling on highways follows a Gaussian distribution, as shown in Fig. 3. The dataset used for the statistical analysis was the recently released HighD dataset from the Institute for Automotive Engineering at RWTH Aachen University in Germany, with a sample size of 1,044,634. This suggests that the transit time of vehicles between the blind zones of adjacent cameras should also follow a Gaussian distribution.

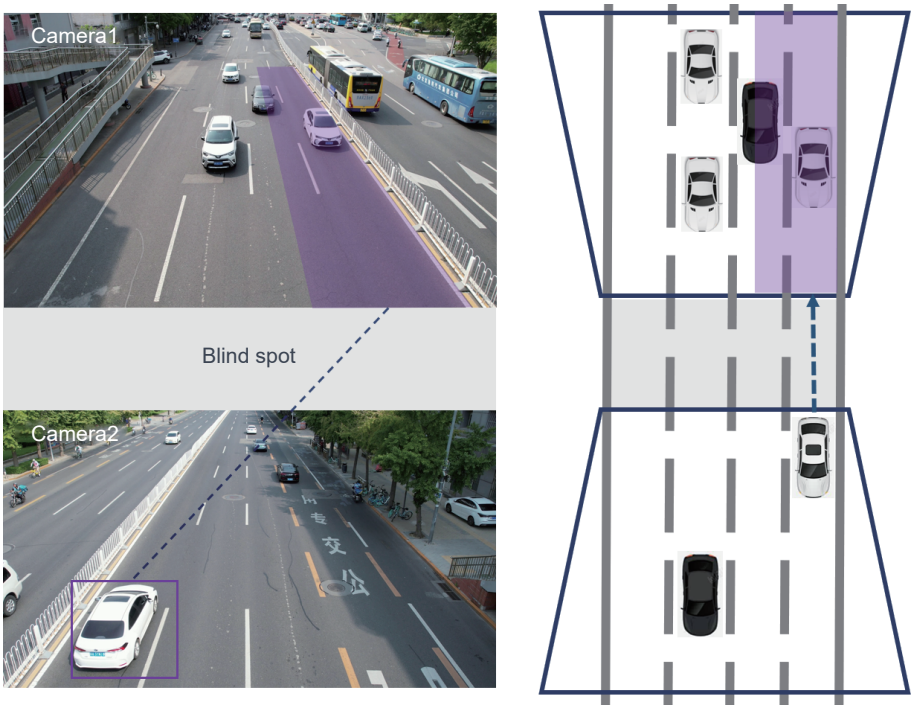

Fig. 2 Spatial point matching-based filtering mechanism.

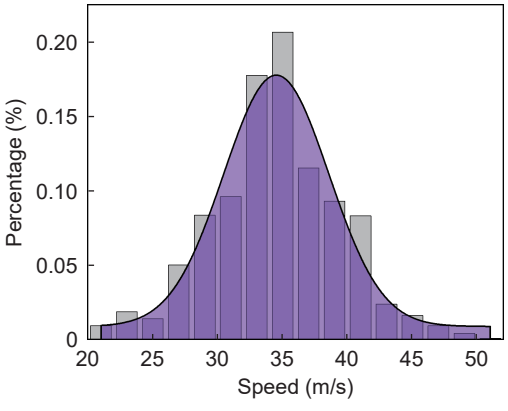

Fig. 3 Statistics of vehicle travel speed on German highways.

Hence, real-time transfer data of vehicles between adjacent roadside cameras enables the establishment of a Gaussian model reflecting microlevel traffic patterns. However, the time transfer model will fail in the following two situations. The first is when the minimum time difference between gallery samples leaving the field of view of one camera and newly detected query samples from the adjacent camera exceeds the 95% confidence interval. The second is when the transfer time of associated vehicles in the blind zone deviates significantly from the average passage time.

In contrast, the target association model based on visual feature matching directly searches for vehicles in the gallery with image features most similar to the query sample, effectively avoiding issues associated with the target association model based on time transfer patterns. Therefore, in certain situations, a model based on visual feature matching can be employed to compensate for the limitations of models based on time transfer patterns. When the time difference adheres to the 95% confidence interval of the Gaussian model, the target can be associated using the model based on time transfer patterns. After completing the association, the transfer time of target vehicles between adjacent cameras is continually updated to maintain the Gaussian model reflecting the microlevel traffic patterns on the road, as illustrated in Fig. 4.

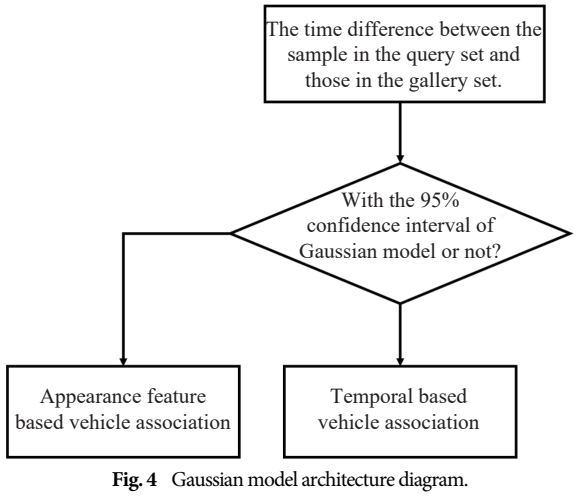

Fig. 4 Gaussian model architecture diagram.

The overall association accuracy of the target association model based on time transfer patterns surpasses that of the model based on visual feature matching. Therefore, when the sample size of the Gaussian model is inadequate, it is necessary to use the target association model based on time transfer patterns to accomplish target association and sample collection. The acquired transfer times of vehicles between adjacent cameras are then utilized to continually update the Gaussian model, facilitating the summarization of microlevel traffic patterns on the road for online learning purposes.

5 Experimental

5.1 Datasets

In this study, real road datasets were utilized to validate the accuracy of the cross-camera tracking algorithm. The dataset used for the experiment is based on real urban road scenarios, with the selected road being Fuxing Road in Haidian District, Beijing, as illustrated in Fig. 5. The chosen route is a bidirectional eight-lane

road. Considering the limited lateral field of view of the cameras, simultaneously monitoring both directions on a bidirectional road would decrease the quality of the captured vehicle images. This, in turn, could impact the performance of the tracking algorithm. Therefore, in this experiment, video collection focused on the traffic flow of a single-direction four-lane segment of the road.

The parameters for recording the road traffic videos were set at 30 frame/s, with a resolution of 1,920 × 1,080 pixels. The total duration of the road traffic videos was 1 h, including 40 min of daytime recording from 10:00 to 11:00 and 20 min of nighttime recording from 19:00 to 21:00.

5.2 Programming environment

The programming environment for the experiments in this research is shown in Table 1. The average computation time for tracking per frame is 32 ms.

5.3 Evaluation metric

IDF1 is a commonly used performance evaluation metric in multiobject tracking. It quantifies the ratio of correctly identified and tracked targets by the tracker to the total number of targets identified and tracked by the tracker. IDP and IDR, representing precision and recall, respectively, are two crucial metrics utilized to assess the performance of a classification model. IDs refer to the identities for cross-camera tracking.

$$\begin{cases} \mathrm{IDF1} = \dfrac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \\ \mathrm{IDP} = \dfrac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFP}} \\ \mathrm{IDR} = \dfrac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFN}} \end{cases} \tag{2}$$

where IDTP indicates the number of times the system successfully associated or correctly tracked a target across different camera frames different camera frames. IDFP indicates the number of times the system incorrectly associated a nontarget object or another target with the trajectory of the target. IDFN indicates the number of times the system fails to correctly associate the target, resulting in the loss of continuity for the target across frames from different cameras. Therefore, the higher IDF1 is, the better the performance of cross-camera vehicle tracking.

5.4 Experimental results

5.4.1 *Comprehensive experiment*

Based on the method described in Section 4, we conducted a quantitative analysis of the cross-domain tracking algorithm on the above datasets. The overall tracking results are shown in Table 2.

Fig. 6 illustrates the tracking performance of the cross-camera vehicle tracking algorithm. Cars with ID 8 and ID 6, along with an Sport Utility Vehicle(SUV) with ID 5, exit the field of view of camera 1 and reappear from camera 2 after some time. The cross-camera tracking algorithm correctly associates the targets in these instances.

As indicated in Table 2, in the daytime scenario, both the single-camera detection and tracking module and the cross-camera tracking module demonstrate stability, showing accurate target detection with fewer ID jumps. The overall performance of cross-camera tracking is commendable. However, during the nighttime, the stability of the single-camera detection and tracking module significantly decreases. The frequency of ID jumps is greater, resulting in a substantial increase in IDFN. In camera 1, the occurrence of multiple ID jumps interferes with the normal

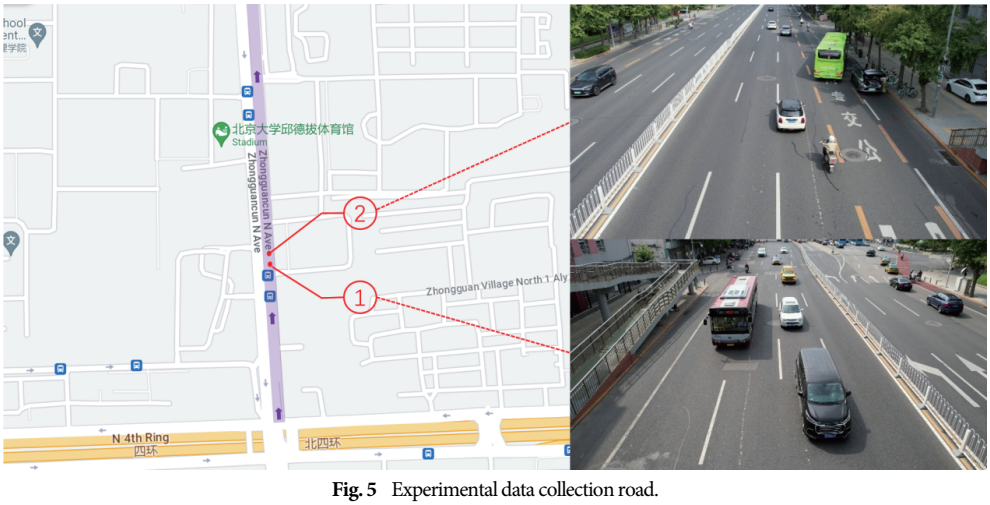

Fig. 5 Experimental data collection road.

Table 1 Programming environment

Environment	Configuration parameter	Version/model
Hardware environment	GPU model	NVIDIA GeForce RTX 3060
	GPU memory	12 GB
	CPU model	12th i7-12700F
	RAM	32 GB
Software environment	Operating system	Ubuntu 18.04
	Programming language	Python 3.7
	Parallel computing architecture	CUDA 11.7
	Deep neural network library	cuDNN 8.3

Table 2 Overall tracking results

Operating condition	IDTP	IDFP	IDFN	IDF1
No. of targets in daytime	842	50	32	**0.953**
No. of targets in nighttime	228	25	48	0.862
No. of targets in daytime and nighttime	1,070	75	80	0.932

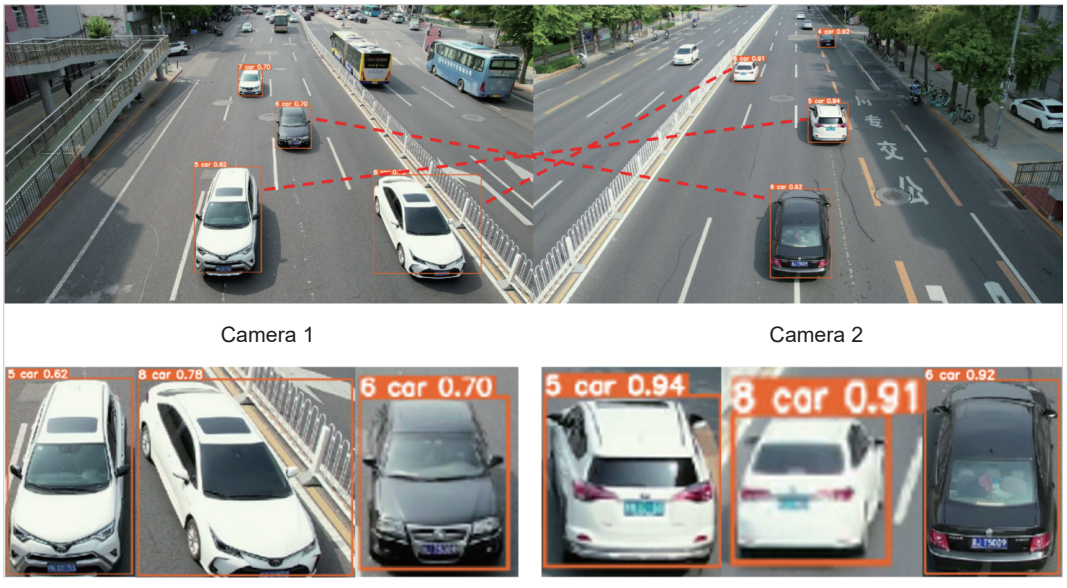

Fig. 6 Algorithm performance demonstration.

association of targets by the cross-camera tracking algorithm, leading to a reduction in the accuracy of cross-camera tracking.

5.4.2 *Comparative experiment*

To validate the effectiveness of our proposed algorithm, we conducted comparative experiments on five distinct target association methods: visual feature-based target association, time-based target association, visual feature-based target association under spatial constraints, time-based target association under spatial constraints, and our proposed target association method, which integrates both temporal and spatial information with visual features. These experiments were conducted using the YOLOv8 detector for object detection and the Deep OC-SORT tracker for tracking as the baseline.

The experimental results are presented in Table 3 and Figs. 7–9. With the Zone-based Target Candidates Filter, a method proposed in this study, both temporal information-based target association and visual feature-based target association methods achieved significant improvements in tracking performance. Even compared with the reidentification algorithm proposed by He et al. (2019), the method introduced in this study demonstrates commendable performance.

In terms of tracking accuracy, the temporal information-based target association method outperformed the visual feature-based method, demonstrating superior performance in both daytime and nighttime tracking. Regarding tracking stability, the experiments revealed that in scenarios with dense road traffic or when false detections occurred in a single-camera detector, as shown in Fig. 10, the temporal information-based target association method exhibited continuous association errors for trailing vehicles. In contrast, the visual feature-based target association method, due to the characteristics of its working principles, was less affected by the introduction of falsely detected targets, resulting in higher tracking stability.

Table 3 Comparative experiment

Type of object association algorithms	Operating condition	IDTP	IDFN	IDFP
Visual feature	No. of targets in daytime	523	36	369
	No. of targets in nighttime	96	42	157
Temporal information	No. of targets in daytime	612	33	280
	No. of targets in nighttime	152	43	101
Visual features under spatial constraints	No. of targets in daytime	774	35	118
	No. of targets in nighttime	143	42	110
Temporal information under spatial constraints	No. of targets in daytime	782	35	110
	No. of targets in nighttime	209	46	44
Combining visual and spatial-temporal information (this study)	No. of targets in daytime	842	32	50
	No. of targets in nighttime	228	48	25
Visual Features and Spatial-Temporal (VFST)	No. of targets in daytime	729	96	103
	No. of targets in nighttime	188	63	50

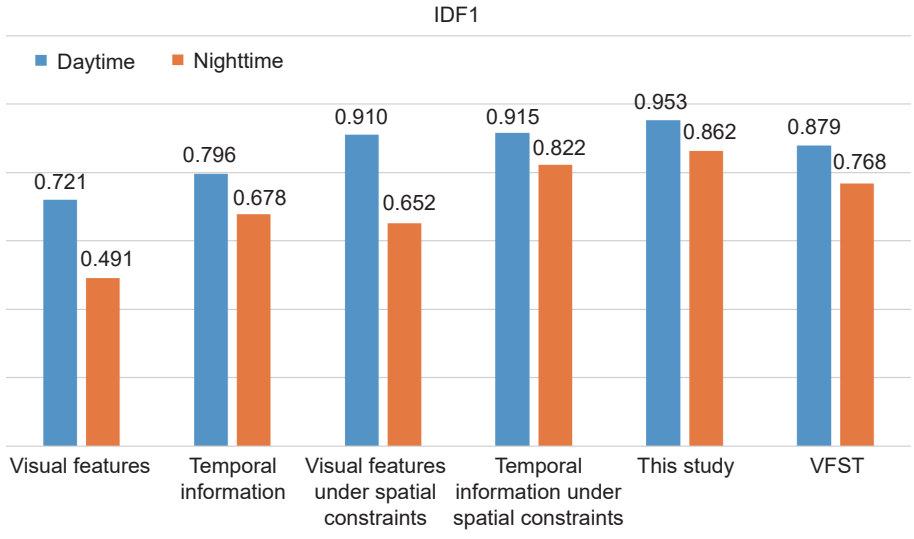

Fig. 7 IDF1 under different operating conditions for different algorithms.

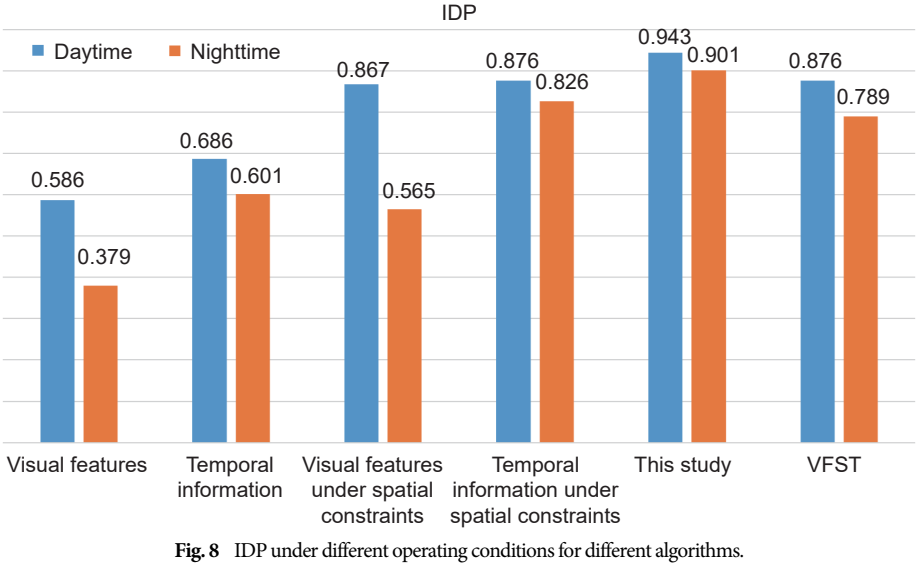

Fig. 8 IDP under different operating conditions for different algorithms.

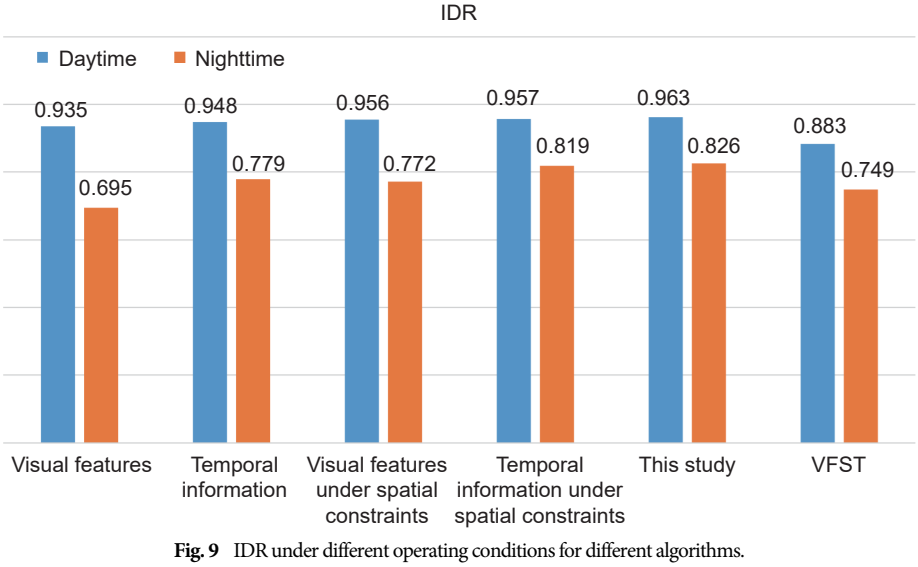

Fig. 9 IDR under different operating conditions for different algorithms.

This study combines visual feature-based target association methods with temporal information-based target association methods and proposes a novel cross-camera target association method that seamlessly integrates visual features and spatiotemporal information. Even in scenarios with dense road traffic or false detections in a single-camera detector, the proposed

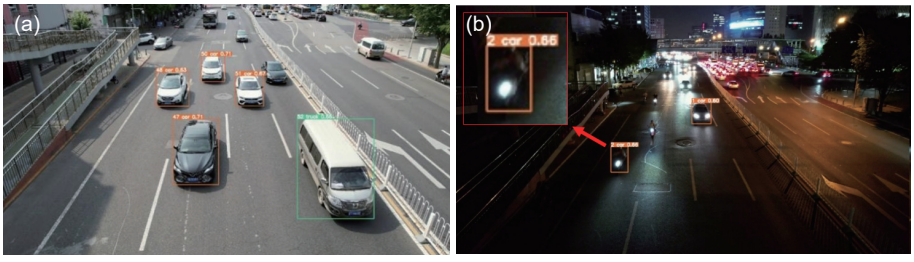

Fig. 10 Special working conditions: (a) heavy traffic flow; (b) false positive.

method maintains stable tracking performance, thereby significantly enhancing the overall accuracy of cross-domain tracking algorithms.

6 Conclusions

To address these challenges and enhance tracking performance in cross-camera vehicle tracking, this study proposes advanced solutions and strategies. First, it addresses issues in both visual-based vehicle association and spatiotemporal-based vehicle association in cross-camera tracking, compensating for their shortcomings by combining visual features and spatiotemporal information. Next, by utilizing the number of vehicles transferred between cameras, a Gaussian model is established and updated to facilitate the summarization of microlevel traffic patterns on roads for online learning. Finally, the proposed method achieves an IDF1 score of 0.932 in real-road scenarios.

Replication and data sharing

The python program code within this research can be made accessible upon request via email to the corresponding author.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (52172389), Natural Science Foundation of Guangdong Province (2022A1515012080), and Tsinghua–Toyota Joint Research Institute Interdisciplinary Program.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

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), 3464–3468.

Bochinski, E., Eiselein, V., Sikora, T., 2017. High-speed tracking-by-detection without using image information. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–6.

Bochinski, E., Senst, T., Sikora, T., 2018. Extending IOU based multi-object tracking by visual information. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–6.

Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K., 2023. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9686–9696.

Caspi, Y., Irani, M., 2000. A step towards sequence-to-sequence alignment. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), 682–689.

Chen, M., Wu, Y. M., Gao, B. L., Zheng, K. Y., 2021. Vehicle target oriented bidirectional matching handover method for multi camera on roadside. Automot Eng, 43, 1435–1441.

Collins, R. T., Lipton, A. J., Kanade, T., 2000. Introduction to the special section on video surveillance. IEEE Trans Pattern Anal Mach Intell, 22, 745–746.

Dick, A. R., Brooks, M. J., 2004. A stochastic approach to tracking objects across multiple cameras. In: Proceedings of the 17th Australian joint conference on Advances in Artificial Intelligence, 160–170.

Girshick, R., 2015. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), 1440–1448.

He, Z., Lei, Y., Bai, S., Wu, W., 2019. Multi-camera vehicle tracking with powerful visual features and spatial-temporal cue. In: 2019 CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Javed, O., Rasheed, Z., Shafique, K., Shah, M., 2003. Tracking across multiple cameras with disjoint views. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, 952–957.

Li, K. Q., Chang, X. Y., Li, J. W., Xu, Q., Gao, B. L., Pan, J. A., 2020. Cloud control system for intelligent and connected vehicles and its application. Automot Eng, 42, 1595–1605.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al., 2016. SSD: single shot multi-box detector. In: European Conference on Computer Vision, 21–37.

Madden, C., Cheng, E. D., Piccardi, M., 2007. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Mach Vis Appl, 18, 233–247.

Maggiolino, G., Ahmad, A., Cao, J., Kitani, K., 2023. Deep OC-sort: Multi-pedestrian tracking by adaptive identification. https://doi.org/10.48550/arXiv.2302.11813

Porikli, F., Divakaran, A., 2003. Multi-camera calibration, object tracking and query generation. In: 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), 1–653.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 39, 1137–1149.

Ren, P., Lu, K., Yang, Y., Yang, Y., Sun, G., Wang, W., et al., 2021. Multi-camera vehicle tracking system based on spatial-temporal filtering. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4213–4219.

Stein, G. P., 1999. Tracking from multiple view points: Self-calibration of space and time. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 521–527.

Tang, S., Andres, B., Andriluka, M., Schiele, B., 2015. Subgraph decomposition for multi-target tracking. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5033–5041.

Tran, D. N. N., Pham, L. H., Jeon, H. J., Nguyen, H. H., Jeon, H. M., Tran, T. H. P., et al., 2022. A robust traffic-aware city-scale multi-camera vehicle tracking of vehicles. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 3150–3159.

Velipasalar, S., Schlessman, J., Chen, C. Y., Wolf, W., Singh, J., 2008. A scalable clustered camera system for multiple object tracking. EURASIP J Image Video Process, 2008, 542808.

Wang, X., Turetken, E., Fleuret, F., Fua, P., 2015. Tracking interacting objects using intertwined flows. IEEE Trans Pattern Anal Mach Intell, 38, 2312–2326.

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.

Yao, H., Duan, Z., Xie, Z., Chen, J., Wu, X., Xu, D., et al., 2022. City-scale multi-camera vehicle tracking based on space-time-appearance features. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 3310–3318.

Yang, X., Ye, J., Lu, J., Gong, C., Jiang, M., Lin, X., et al., 2022. Box-grained reranking matching for multi-camera multi-target tracking. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 3096–3106.

Bolin Gao received the B.S. and M.S. degrees in Vehicle Engineering from Jilin University, Changchun, China, in 2007 and 2009, respectively, and the Ph.D. degree in Vehicle Engineering from Tongji University, Shanghai, China, in 2013. He is now an Associate Research Professor at the School of Vehicle and Mobility, Tsinghua University. His research interests include the theoretical research and engineering application of the dynamic design and control of Intelligent and Connected Vehicles, especially about collaborative perception and tracking method in cloud control system, intelligent predictive cruise control system on commercial trucks with cloud control mode, as well as the test and evaluation of intelligent vehicle driving system.

Zhuxin Li received the B.Eng. degree in Vehicle Engineering from FoShan University, China, in 2021. Currently, he is a M.Eng. candidate in School of Mechanical and Electrical Engineering, Guangdong University of Technology, China. Since 2022, he has engaged in research work in the School of Vehicle and Mobility, Tsinghua University, China. His research interests include cross camera vehicle tracking and cloud control system.

Dong Zhang is currently a Lecturer in Department of Mechanical and Aerospace Engineering, Brunel University London, UK. He joined Brunel in 2021 after having spent one year at the Nanyang Technological University, Singapore as Research Fellow in School of Mechanical and Aerospace Engineering, and School of Electrical and Electronic Engineering. He received the M.Sc. degree in 2015 from Jilin University, Changchun, China and Ph.D. degree in 2019 from University of Lincoln, Lincoln, UK. He joined Brunel and founded the Intelligent Driving and Transportation Research Group in September 2021. He has published more than 30 articles in international journals as well as numerous conference articles, and about 10 granted patents, mostly in the area of intelligent driving and transportation control and active safety systems for road vehicles.

Yanwei Liu received the B.S. degree in Vehicle Engineering from Jilin University, Changchun, China, in 2007, and the Ph.D. degree in Vehicle Engineering from the South China University of Technology, Guangzhou, China, in 2012. He is currently an Associate Professor with the Guangdong University of Technology. His current research interests include control and optimization of electric vehicles and intelligent connected vehicles.

Jiaxing Chen received an M.S. degree in Electrical and Computer Engineering from the University of Illinois, Chicago, USA, in 2021 and worked as an Algorithm engineer at the National Innovation Center of Intelligent and Connected Vehicles, Beijing, China from 2021 to 2022. He is currently pursuing a Ph.D. degree at the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include Computer Vision with Deep Learning and perception of Intelligent and Connected Vehicles and Transportation.

Ziyuan Lv received the B.S. degree in Engineering from Jiangxi Agricultural University, Jiangxi, China, in 2021. He is currently pursuing the M.S. degree in mechanical engineering with the Guangdong University of Technology. Since 2022, he has engaged in research work in the School of Vehicle and Mobility, Tsinghua University, China. His research interest includes roadside sensor deployment and cloud control system.