Research on Accompanying Vehicle Recognition Based on Big Data Technology

Jinsong Li[1, a]

[1]Liaoning Communication University, Shenyang, Liaoning, China

[a]lijinsong_lncm@sohu.com

Abstract

With the acceleration of urbanization, the transportation system is facing increasingly complex governance challenges, among which the identification and analysis of accompanying vehicle behavior has become a key issue. Traditional methods have some defects, such as low efficiency and limited coverage, and it is difficult to deal with massive data. In this study, an accompanying vehicle identification framework based on big data technology is constructed. By efficiently processing massive trajectory data at city level, fusing multi-source heterogeneous information and designing extensible algorithms, real-time accurate identification of accompanying behavior is realized. A hybrid model combining spatio-temporal feature engineering and graph neural network (GNN) is proposed to effectively describe the spatio-temporal similarity between vehicles and capture the complex dynamic adjoint relationship. The experimental results show that the framework has excellent performance in the accompanying vehicle identification task, which can provide technical support for public safety applications such as deck car tracking and gang crime warning, and also provide theoretical and practical reference for intelligent transportation system optimization and spatio-temporal data mining.

Keywords

Accompanying Vehicle Recognition; Big Data; Graph Neural Network.

1. Introduction

With the acceleration of urbanization and the rapid growth of motor vehicle ownership, the transportation system is facing increasingly complex governance challenges. Among them, the identification and analysis of the behavior of accompanying vehicles (that is, the phenomenon of multiple vehicles moving in a highly coordinated manner in time and space) has become a key issue in traffic management, public safety and other fields [1]. For example, licensed cars evade supervision by imitating legal vehicle tracks, criminal gangs use multi-vehicle cooperation to carry out theft or escape, and logistics enterprises optimize transportation efficiency through fleet cooperation [2]. The traditional identification method of accompanying vehicles mainly relies on manual inspection or simple matching based on rules, which has some defects such as low efficiency, limited coverage and difficulty in dealing with massive data. At the same time, the multi-source heterogeneous data generated in the transportation system has exploded, which provides a data foundation for deep mining with vehicle behavior, but also puts forward higher requirements for computing power and analysis methods.。

The rise of big data technology provides a new paradigm to solve the above problems. The combination of distributed storage, parallel computing framework and machine learning algorithm can effectively process PB-level trajectory data and extract hidden spatio-temporal correlation patterns from it [3]. For example, cluster analysis can identify frequently co-occurring vehicle groups, association rules mining can discover cooperative behavior rules among vehicles, and deep learning model can capture complex dynamic adjoint relations [4]. In addition, the big data platform supports real-time data stream processing, which makes it possible to respond to traffic incidents in seconds. However, the existing research still has limitations [5-6]. Most methods are only verified on small-scale data sets, and lack of adaptability analysis of city-level total trajectory data; The time/space resolution of GPS, bayonet, video and other data is quite different, and the fusion mechanism is not perfect [7]; Traffic flow is significantly affected by road conditions, weather and other factors, and the existing models are not robust to unsteady environment [8].

The purpose of this study is to build an accompanying vehicle identification framework based on big data technology, and realize real-time and accurate identification of accompanying behavior by efficiently processing massive trajectory data at city level, fusing multi-source heterogeneous information and designing extensible algorithms. A hybrid model combining spatio-temporal feature engineering and graph neural network (GNN) is proposed, which not only provides technical support for public safety applications such as deck car tracking and gang crime warning, but also provides theoretical and practical reference for intelligent transportation system optimization and spatio-temporal data mining.

2. Identification Method of Accompanying Vehicles Based on Big Data

2.1. System architecture design

The system architecture proposed in this study consists of four layers, and the overall architecture is shown in Figure 1 below. The data acquisition and preprocessing layer is responsible for obtaining vehicle trajectory data from multi-source sensors such as bayonet and GPS, and cleaning, compression, segmentation and spatio-temporal index construction; The distributed storage and computing layer uses HDFS to save massive data, and realizes efficient parallel computing based on Spark. Accompanying behavior recognition algorithm layer is the core, which integrates trajectory similarity calculation, spatio-temporal feature extraction and GNN modeling to complete accompanying behavior judgment. The application service layer provides services such as accompanying vehicle early warning, track tracing and statistical analysis to traffic management and police system through API, and realizes the closed loop of the whole process from data to decision support.

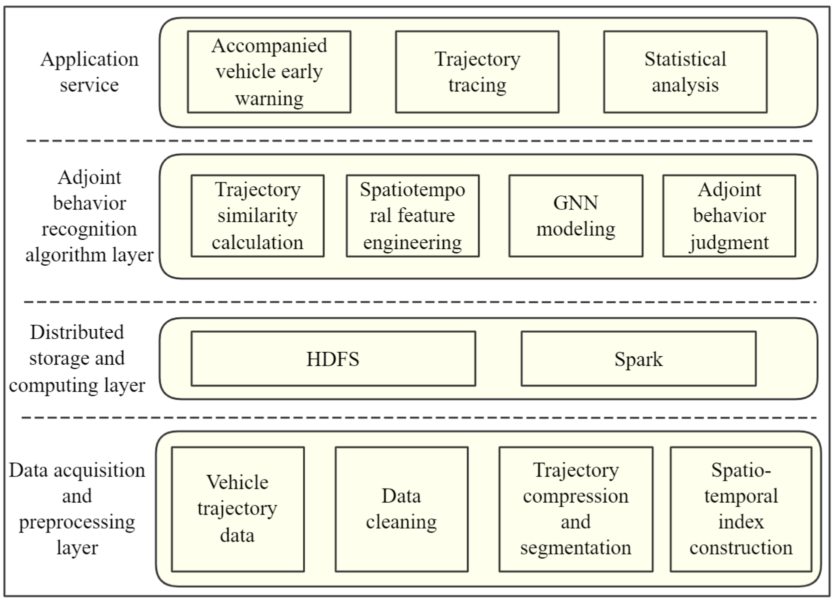

Figure 1. The structure of accompanying car identification system based on big data

2.2. Core algorithm design

The key of accompanying vehicle identification is to quantify the spatio-temporal similarity between vehicle tracks [9]. Define the similarity of two trajectories T_i, T_j in the time window Δt as:

$$S(T_i, T_j) = \frac{1}{N} \sum_{k=1}^{N} \exp\left(-\frac{d_k(T_i, T_j)^2}{2\sigma^2} \right) I\left(\left| t_{i,k} - t_{j,k} \right| < \tau \right) \tag{1}$$

Where $d_k(T_i, T_j)$ is the Euclidean distance of the trajectory point k; σ is a distance scaling parameter, which controls the decay rate of similarity; $I(\cdot)$ is indicator function, when the difference between the two tracks at the time stamp $t_{i,k}, t_{j,k}$ is less than the threshold τ, it is 1, otherwise it is 0; N is the number of trajectory points. This formula combines spatial proximity and time synchronization, and can effectively describe short-time adjoint patterns.

The vehicles are regarded as nodes, and the trajectory similarity between vehicles is regarded as edge weight, so as to construct a spatio-temporal trajectory diagram $G = (V, E)$, where V is the vehicle set; E is an edge set, and the weight is defined by $S(T_i, T_j)$.

Graph convolution network (GCN) is used to aggregate neighbor information and learn the embedded representation of vehicle nodes [10]. The features of node v_i at the $l+1$ layer are updated as follows:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W^{(l)} h_j^{(l)} \right) \tag{2}$$

Where $N(i)$ is the neighbor set of node i; α_{ij} is the normalized edge weight; $W^{(l)}$ is a trainable parameter matrix; σ is the activation function. Through multi-layer graph propagation, the model can capture the high-order adjoint relationship between vehicles.

Based on the learned node embedding, the adjoint probability between vehicles is calculated:

$$P(y_{ij} = 1) = sigmoid\left(MLP\left(h_i \| h_j \right) \right) \tag{3}$$

Where $y_{ij} = 1$ indicates that the vehicle i, j is an accompanying vehicle; $\|$ stands for vector splicing; MLP is a multi-layer perceptron classifier. Finally, combined with the threshold value, the accompanying vehicle pairs are filtered and the real-time warning is generated.

3. Experimental Design and Result Analysis

In order to verify the effectiveness and superiority of the accompanying car recognition framework based on big data technology and GNN (hereinafter referred to as GNN-ACR framework), a complete experimental scheme was designed, and the results were analyzed from multiple dimensions.

The experimental environment is built on a distributed computing cluster, including one Master node and four Worker nodes. Each node is equipped with Intel Xeon Gold 6248R CPU, 128GB RAM and NVIDIA Tesla V100 GPU. The software environment includes Apache Spark 3.3.1, Python 3.8, Pytorch 1.13.1 and other tools. The data set adopts the real vehicle trajectory data of a megacity in southern China for two weeks in 2022, which is derived from the road bayonet monitoring system. After anonymization, it contains 100 million records of more than 1 million vehicles, and the data format mainly includes vehicle ID, time stamp, bayonet number and latitude and longitude. The data are divided into training set (the first 12 days), verification set (the 13th day) and test set (the 14th day) in chronological order, which are used for model training, parameter tuning and performance evaluation respectively.

In order to verify the effectiveness of the GNN-ACR framework, this study selected three representative baseline methods for comparison: Rule-Based method, which judges the adjoint relationship by setting the rules of continuous passing through the same bayonet in the time window; Trajectory similarity-based method (DTW-LR), using dynamic time warping (DTW) to measure trajectory similarity and combining with logistic regression (LR) to classify; And the deep learning method (LSTM-Siamese), which uses the twin-tower LSTM network to extract trajectory features and judge the accompanying behavior through similarity calculation. By comparing with these traditional and cutting-edge methods, the performance advantages of the proposed model in the accompanying vehicle identification task are comprehensively evaluated.

On the test set, the performance pairs of each method are shown in Table 1 below. It can be seen that the traditional Rule-Based method has the lowest precision and recall rate, because it is difficult to cope with complex road network and variable speed trajectory, and there are many false positives and false negatives. The performance of DTW-LR method is significantly improved, which shows that the trajectory similarity measurement is effective, but the linear model is difficult to capture the nonlinear relationship. The LSTM-Siamese method performs well, and the F1-Score reaches 0.840, which proves the powerful representation ability of the deep learning model. The GNN-ACR framework proposed in this study is the best in all indicators, especially the recall rate (0.854) and AUC(0.925). This shows that GNN can find hidden and cooperative accompanying behaviors more effectively by explicitly modeling the interaction between vehicles (graph structure), reduce the number of missed reports, and the overall comprehensive performance is the best.

Table 1. Performance comparison of different identification methods for accompanying vehicles

Method	Precision	Recall	F1-Score	AUC
Rule-Based	0.721	0.523	0.606	0.685
DTW-LR	0.835	0.704	0.764	0.812
LSTM-Siamese	0.881	0.802	**0.840**	0.872
GNN-ACR (our)	0.893	0.854	0.873	0.925

The performance of each method in processing 24-hour data stream (about 8 million records) is tested (as shown in Table 2 below). DTW-LR algorithm has high computational complexity, and it is almost impossible to process city-level data in a single computer environment, so it is not practical. This research framework is based on Spark distributed optimization, with the shortest total time (12.1 minutes) and the highest throughput (11.01 thousand pieces/second), which fully embodies the great advantages of big data technology in dealing with massive trajectory data and can meet the business needs of quasi-real-time identification.

Table 2. Comparison of treatment efficiency

Method	Total elapsed time (minutes)	Average throughput (1000 bars/sec)	Average delay (ms)
Rule-Based	18.5	7.20	139
DTW-LR (single machine)	245.3	0.54	1850
GNN-ACR (Spark)	12.1	11.01	91

The influence of the critical parameter-adjoint decision time threshold τ on the model performance is studied. As can be seen from Figure 2 below, when τ is too small (< 60s), the required time is too synchronous, and many reasonable accompanying behaviors will be missed, resulting in a decrease in the recall rate; When τ is too large (> 300s), a large number of irrelevant vehicles will be introduced, which will lead to the decline of precision. This model achieves the best equilibrium point when τ =120s.

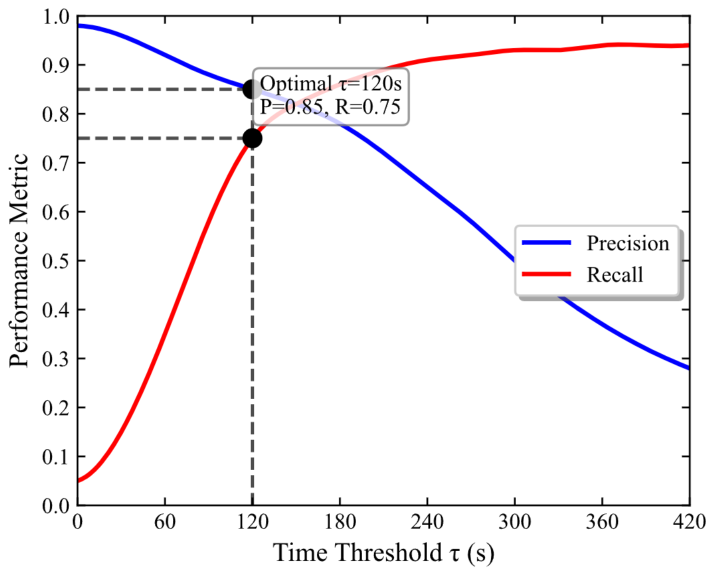

Figure 2. Influence of different time threshold τ on the performance of GNN-ACR model

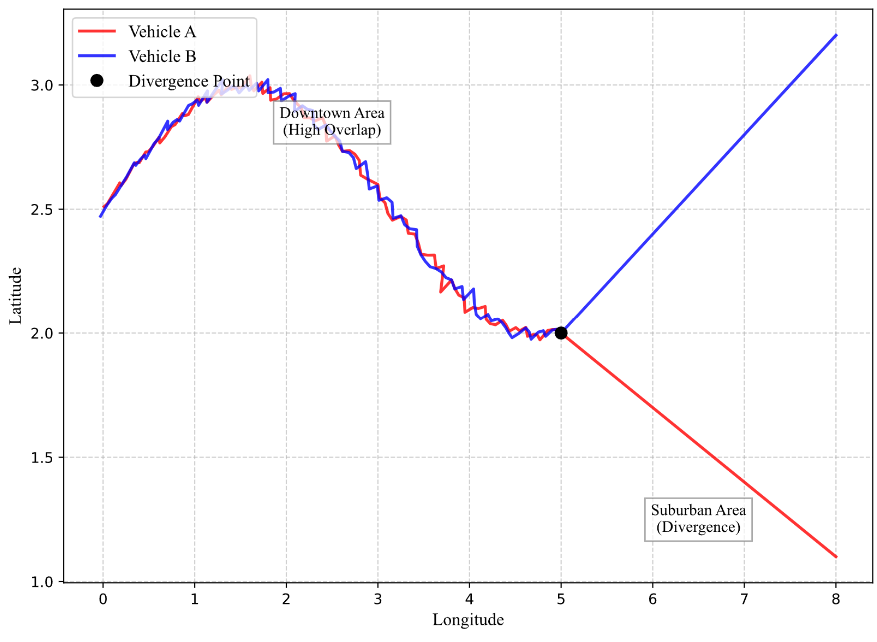

Figure 3. Schematic diagram of case trajectory of deck car

Using this framework, a pair of vehicles (A and B) which are highly suspected to be deck cars are successfully identified from the test data. As shown in Figure 3 above, the tracks of the two cars are highly overlapped and synchronized in the downtown area (in line with the characteristics of deck cars), but they parted ways after a certain section in the suburbs, which proves their illegal behavior. The case was handed over to the traffic control department and verified, which proved the value of this framework in practical application.

4. Conclusion

Based on big data technology, the accompanying vehicle identification framework aims to effectively process massive trajectory data at city level, fuse multi-source heterogeneous information and design extensible algorithms to realize real-time and accurate identification of accompanying behaviors. By constructing the spatio-temporal trajectory diagram and using GCN to aggregate neighbor information, combined with the multi-layer perceptron classifier, the model can capture the high-order adjoint relationship between vehicles, thus improving the recognition accuracy and efficiency. The experimental results show that compared with the traditional methods, the proposed GNN-ACR framework achieves the best performance in precision, recall, F1-Score and AUC, especially in recall (0.854) and AUC(0.925). In addition, the framework shows significant advantages in processing large-scale data streams, with the shortest total time consumption (12.1 minutes) and the highest throughput (11.01 kilobytes per second), which fully reflects the great potential of big data technology in processing massive trajectory data. By adjusting the adjoint decision time threshold, the model can maintain the best performance balance in different scenarios. Practical application cases verify the effectiveness of the framework in identifying illegal activities such as deck cars, and provide strong technical support for traffic management, public safety and other fields. This study not only provides theoretical and practical reference for intelligent transportation system optimization and spatio-temporal data mining, but also lays a solid foundation for future related research.

References

[1] Xin Li,Guochen Liu,Kang Song & Yanlong Zhao.(2025).A Bilinear Parameter Identification Algorithm for Vehicle Mass Estimation.Journal of Systems Science and Complexity,38(5),1833-1852.

[2] Qian Pan,Hongwei Han & Dong Qiao.(2025).Rapid identification of time-varying aerodynamic parameters for aeroassisted vehicle via asynchronous iterative filtering.Advances in Space Research,76(4),2207-2220.

[3] Xiaobin Fan,Mingxin Chen,Shuaiwei Zhu & Shuwen He.(2025).Design of ABS sliding mode control system for in-wheel motor vehicle based on road identification.Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering,239(7),2741-2752.

[4] Sadaqat Hussain,Syed M. Hussain,Yu Xin & Zuo Cai Wang.(2025).Vehicle load identification based on bridge response using deep convolutional neural network.Journal of Asian Architecture and Building Engineering,24(3),1235-1252.

[5] Jianliang Zhang,Yuyao Cheng,Jian Zhang & Zhishen Wu.(2025).A spatiotemporal distribution identification method of vehicle weights on bridges by integrating traffic video and toll station data.Journal of Intelligent Transportation Systems,29(3),287-304.

[6] Quan Yuan,Rujun Yan,Ashton Yu Xuan Tan,Qing Xu & Jianqiang Wang.(2025).Technology roadmap of risk identification and collision avoidance decision-making in autonomous vehicles for domestic animals.International Journal of Crashworthiness,30(3),294-305.

[7] Su Hyun Park,Jeong Hyun Sohn,Hyeong Gi Min & Yong Jae Kim.(2024).Identification of operational limitations of an autonomous vehicle based on dynamic simulation.Journal of Mechanical Science and Technology,38(12),6519-6528.

[8] Xuesong Wang, Shikun Liu, Junyi Zhang & Daiheng Ni.(2024).Real-Time Risk Identification and Prediction for the Target Lane's Following Vehicle during Lane Change.Transportation Research Record, 2678(12), 1785-1798.

[9] WanyuWei, XinshaFu, SiqiMa, YaqiaoZhu & NingLu. (2024). Reducing overfitting in vehicle recognition by decorrelated sparse representation regularisation.IET Computer Vision, 18(8), 1351-1361.

[10] Soumyajit Gayen,Sourajit Maity,Pawan Kumar Singh & Ram Sarkar.(2024).SimSANet: a simple sequential attention-aided deep neural network for vehicle make and model recognition.Neural Computing and Applications, 37(1),1-21.