Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 2 No.3 (2025)

A Study on Lane Congestion Recognition Mechanism for Highways Based on Multi-Source Data

Linlin Li[1]*, Zihao Lü[2]

1.Commercial Management Branch of Yunnan Communications Investment Group Business Development Co., Ltd., 65000, China

2.International College, Kunming City University, 650106, China

Corresponding author: Linlin Li

Abstract: This paper mainly studies the recognition mechanism of traffic congestion on the highway based on multi-source data. To form an accurate and good means for recognizing lane congestion by putting together various data sources such as traffic flow, speed, density and video surveillance data. We propose the use of the combination of machine learning algorithms and traditional traffic theory for its data fusion model. Realworld highway data is used for experiments to prove this method. The results show that the proposed mechanism performs better than traditional single-source data-based approach w.r.t accuracy and robustness.

Keywords: Highway Congestion; Lane-Level Recognition; Multi-Source Data

Published: Jun 25, 2025

DOI: https://doi.org/10.62177/jaet.v2i3.476

1.Introduction

Highway transportation is critically important in modern society, facilitating the movement of people and goods. However, increasing vehicle accumulation on roadways creates significant traffic challenges. Traffic congestion resembles vehicles tightly packed in extended queues, degrading air quality with harmful pollutants. Effective traffic management techniques—such as dynamic traffic control and route guidance—require timely, accurate recognition of congestion on individual lanes.

Traditional congestion recognition methods often rely on single data sources, such as loop detectors, facing limitations in coverage breadth and accuracy. Equipment failures or extreme weather conditions can cause loop detector errors, resulting in data gaps. Conversely, multi-source data integrates information from various sensors—including inductive loop detectors, microwave sensors, video cameras, and GPS-equipped vehicles—providing a more comprehensive perspective. Merging these diverse data resources enables the creation of significantly more robust and accurate systems for lane-specific congestion recognition.

2.Related Work

Extensive research has focused on highway congestion recognition. Early approaches predominantly utilized single-source data. For instance, some studies employed loop-derived parameters such as flow, speed, and density to identify congestion based on empirically established thresholds. However, these methods depend solely on one data source, meaning their results may contain errors and are inherently limited to available sensor coverage[1].

With advancements in sensor technology and data fusion techniques, researchers increasingly consider multi-source data integration. Studies have applied fusion methods like Kalman filters and fuzzy logic to combine information from diverse sensors. Machine learning algorithms, including support vector machines and neural networks, have also been employed to classify congestion states using multi-source data. Nevertheless, more effective integration of heterogeneous data sources and a deeper understanding of the relationship between various traffic data types and congestion conditions remain essential research objectives[2].

3.Methodology

3.1 Multi-source Data Collection

In this research multiple data sources will be utilized for collecting the traffic information of highway These include:

• Inductive loop detectors: Installed on road side to measure number of passing cars, speed and occupancy(time period car occupies detector).

• Microwave sensors: Next to the highway, it detects the traffic flow and speed through emitting microwave and analyzing the returning microwave

• Video surveillance systems: The sensors come with cameras that could capture pictures in real time on the highway, from where information like the quantity of cars and how long the queue is will be got by computer vision.

• GPS-equipped vehicles: Individual vehicles real-time location and speed data can be given and this data can be aggregated to get lane-level traffic flow data.

Table 1 provides a summary of the characteristics of each data source.

Data Source	Sampling Frequency	Coverage	Advantages
Inductive Loop Detectors	1 Hz	Single lane at detector location	High accuracy for flow, speed, and occupancy
Microwave Sensors	0.5 Hz	Multiple lanes over a certain distance	Non-intrusive, wide coverage
Video Surveillance Systems	25 fps	Camera field of view	Visual information for vehicle density and queue length
GPS-equipped Vehicles	1 Hz	Entire highway network (depending on vehicle penetration)	Real-time individual vehicle data

3.2 Data Preprocessing

First, preprocess raw data of each source, clean and normalize it. Noise-removal techniques such as median filtering are applied to remove any anomaly and error of the signal[3]. For example, for a vehicle having GPS, if the speed suddenly changes a lot because the signal was lost, we use a median filter with a window of 5 to remove these sudden changes. Normalization is done by converting data to common scale with the help of Min-Max Normalization which is defined as:

x_{normalized} = (x - x_{min})/(x_{max} - x_{min})

where x denotes the original data value and x_{min} denotes the smallest value in the set of data values, and x_{max} indicates the largest value in the set of data values;

3.3 Feature Extraction

tems related to those are found from processed information. Regarding data about traffic flow collected by inductive loop detectors, microwave sensors et al., information on things like average speed, standard deviation of speed, the rate of traffic flow, and vehicular density is picked out. From looking at the video data, we can determine things like how many vehicles are in each lane, what the average distance is between vehicles, and if there are any vehicles that have stopped by checking to see if their position stays the same for a certain amount of time, say 5 seconds. For the GPS equipped vehicle data, we calculate the metrics like -avg_speed_of_vehicle_in_lane, percent_with_speed_below_given_speed for speed less than 30km/h. feature extracttion of every data sources is listed as table 2.

Table 2: Extracted Features from Multi-source Data

Data Source	Extracted Features
Inductive Loop Detectors and Microwave Sensors	Mean speed, standard deviation of speed, traffic flow rate, vehicle occupancy
Video Surveillance Systems	Vehicle density, queue length, number of stopped vehicles
GPS-equipped Vehicles	Average vehicle speed, percentage of slow-moving vehicles

3.4 Data Fusion Framework

A hierarchical data fusion framework consisting of three layers is suggested: the data preprocessing layer, the feature extraction layer, and the decision-making layer[4]. As demonstrated in Figure 1.

At the data preprocessing level, as mentioned above, we clean the raw information and make it uniform. In the feature extraction layer, the corresponding feature is extracted as mentioned[5]. the decision making layer adopts the bayesian network and support vector machine method Bayesian Networks is applied for modeling the probability relationships between the extracted features and congestion state and then the SVM is used for classification based on the result of BN[6].

The Bayesian Network is built by first determining the conditional probability distributions between the features and the congestion states based on the historical data. Next we train SVM by the outcome probabilities from Bayesian Network to classify the traffic congestion into 3 levels, i.e., free-flow, slow-flow and congestion.

3.5 Congestion Recognition Mechanism

Congestion states are classified into free flow,slow flow,congestion It’s based on a set of traffic parameters and threshold, and it’s determined by history data. It is clear from table 3 threshold values at varying degree of congestion with respect to vehicle speed. Similarly by density and flow.

Table 3: Thresholds for Different Congestion Levels

Congestion Level	Vehicle Speed (km/h)	Vehicle Density (vehicles/km/lane)	Traffic Flow Rate (vehicles/h/lane)
Free Flow	v>60	k<20	q>1500
Slow Flow	30\leq v\leq60	20\leq k\leq40	800\leq q\leq1500
Congestion	v<30	k>40	q<800

4.Experiments

4.1 Experimental Data and Preprocessing

The experimental authentication of the suggested congestion recognition scheme is performed on real traffic data gathered from a 10 - kilometer stretch of a city highway for a thirty - day term comprising both rush hour and off - rush hour periods[7]. This collection involves data taken from many different kinds of sensors - there are 10 inductive loop detectors delivering 1Hz worth of traffic flow numbers, vehicle speed, plus how occupied those lanes are; 5 microwave devices give 0.5Hz information on combined multilane speed and amount within 500meter areas as measured via Doppler effects; 3 HD video cameras take pictures every 25framespersecond where vehicles are seen, then analyzed by computers to tell about stuff like car crowds and lines using vision technology; lastly, privacy protected movement details from 2,000 special vehicles with on board GPS trackers are also used, offering 1Hz location and velocity figures, permitting us to figure out each line's traffic condition all along the whole area[8].

Preprocessing process made for quality and consistency. Speed anomaly due to signal interference or equipment noise is handled by median filtering a 5-sample window, Missing sensor data within 5 minute intervals is interpolated using linear regression on the adjacent time points[9]. Perform min-max normalization on all features to scale them into the range so as to provide uniform input for machine learning models. For GPS data, lane-level speed is aggregated via an average of all speed values over every 10 second window to mitigate the sparsity of GPS sampling and for the video based features a median

smooth is applied so as to dampen short term variations from camera frames or temporary occlusions[10].

4.2 Dataset partition and evaluation indicators

The dataset is divided into 70% training set and 30% test set. The temporal order is kept to maintain sequencing dependency in traffic data. This way of breaking it down makes sure that the models will get trained using regular traffic patterns and then be tested with uncommon heavy traffic situations and weird disruptions in the flow of traffic. four evaluation metric will be used in order to fully evaluate the classification accuracy,precision,recall,F1 score.

Accuracy refers to the total percentage of correct predictions across all three congestion conditions (free flow, slow flow, congestion) as a general measure of model accuracy.

Precision is focused on the truth worthiness of our positive predictions, the correct amount of our recognized Congestion instances amongst all those predicted as congestions - preventing the system from raising traffic alerts inappropriately.

Recall indicates the model's capability to identify genuine congestion, which counts as the number of accurately marked congested examples over all true congested examples in the database and it is very important to prevent bottleneck unnoticeable.

The F1-score is a balanced harmonic mean measure of precision and recall. This makes it quite useful if your dataset is imbalanced with fewer congestion events (minority class) compared to free-flow.

4.3 Comparison Methods for Benchmarking

Three comparisons are made to examine the usefulness of the framework for handling multi-source data: these include two different single-source comparisons and one multi-source comparison.

Loop Only: This is a base line which is using empiric thresholds based on the traffic flow theory to differentiate congestion states. Specifically, vehicle speed threshold set as 60km / h and 30km / h as velocity threshold, density threshold set to be 20 and 40 vehicles / km / lane as congestion threshold. It is fairly simple and very common, but it cannot be used where detectors aren't available; also they are very vulnerable to equipment failure.

Single Source (only video sensor) Using visual feature which is extracted by camera picture. Infer the congestion status using rule based logic. Key indicators are vehicles/km·lane and queue length, which is defined as the distance of stopped cars following one another. Congestion occurs when density surpasses 40 cars/km/lane, but when queue is longer than 100 meters, it will suffer from poor performance due to low-light or complex traffic occlusion situation.

Multi-source (Kalman filtering): As a traditional data fusion representative, it models traffic parameters such as speed and density as state variables of a linear state-space system to fuse loop detector and video sensor data. KalmanFilter gives optimal estimates by cutting down the errors in between what these different sensors detect about something moving along, which is helpful for this way of joining up those changing pictures across time when tracking something in traffic.

4.4 Implementation details and model config

All models are written in python, scikit learn is used as the machine learning portion and opencv is used to extract features from the videos being analyzed. Proposed framework BN is built with domain knowledge to set up cause-and-effect relationship between 12 car traffic details such as loop detector's average speed, video's vehicle count, GPS's slower vehicle proportion and there are 3 traffic congested state For continuous features, conditional probability distributions are modeled as Normal distributions, with estimates for mean and variance calculated via maximum likelihood estimation based on training data

the support vector machine(SVM) component employs the radial basis function(RBF) kernel to handle non linear decision boundaries, something needed due to the multi source feature interaction. hyperparameter optimazation, particularily for the kernel width (γ) and regulization parameter (C), is carried out via 5 - fold cross - validation in the training set and the aim of hyperparameter tuning is to maximize F1 to balance the classes. Computational experiments are performed on a workstation with intel i7 - 10700k cpu and 32gb ram for reproducibility and practical scalability for real - time applications.

This experimental setting will lead to a good exploration on the effectiveness of this proposed method in the integration of different sources, which is better than the single-source baseline, and has clear advantages over existing traditional multi-source fusion. By systematically tending to data quality, evaluative thoroughness and computational feasibility, the

experiments set up solid groundwork for testing the proposed congestion recognition method in real traffic environments.

5.Results and Discussion

5.1 Experimental Results

Table 4: Performance Comparison of Different Methods

Method	Accuracy	Precision	Recall	F1-score
Single-source (loop detectors only)	0.75	0.72	0.78	0.75
Single-source (video only)	0.70	0.68	0.72	0.70
Multi-source (Kalman filtering)	0.82	0.80	0.84	0.82
Proposed method	0.88	0.87	0.89	0.88

From Table 4, it can be seen that the proposed method gets the highest accuracy, precision, recall, and F1-score among all the methods. The combination of many data sources has an obvious improvement over single source in recognition performance. The single source methods are less effective because of the little information given For example, using a loop detector only method cannot detect congestion in places without a loop detector, and also the loop detector can be abnormal, and the video only method in the low-light environment will find it difficult, and if there is an obstacle, the view will be occluded and so on. The multi-source (Kalman filtering) method, although it is better than the single-source methods, yet has a lower performance when compared to the proposed method. It is because the combination of Bayesian network and SVM is better than the Kalman filtering method based on linearity for the reason that it can more effectively characterize the complex correlation among the features and the congestion states, including the nonlinear correlations.

5.2 Case Study

In order to further validate the effect of the proposal, a time period during which a traffic accident happened causing severe congestion is selected for a case study. It can be seen from Fig.2 that the traffic parameters, different data sources, and congestion classification are presented with our suggested approach.

When the accident occurs around 10:00am, the vehicle speed detected by the loop detector and the Gps equipped vehicles dropped drastically, the vehicle amount in the video increased sharply, and the traffic flow rate decreased. The method proposed correctly predicts the congestion state at this time, the single source (loop detectors only) method takes several minutes longer to detect the congestion because there aren’t many detectors in the area, the single source (video only) method mis-classifies the first stage of the congestion as it gets temporarily occluded in the camera view.

5.3 Discussion of Limitations and Future Work

although the proposed method is successful and promising but it has a limitation too: Currently, the studies used historical data to determine the threshold values for congestion levels, and there may need to be different threshold values for different highway segments and traffic conditions. In addition to this it’s performance for video based feature extractions can have a negative impact due to bad weather condition such as raining or fog and poor quality of our image would emerge.

future direction will be as follows:

1. Develop adaptive threshld adjustments algorithims that can make autmatic adjustmnets according to thte current traffic dats.
2. Making the video-based feature extractors' resistance to poor weather better with better picture techniques.
3. Deep learning algorithms like CNN’s to do full video traffic congestion recognition from raw video.
4. Large scale testing of highway networks that were different from each other to test the proposed method and see if it was widely applicable.

6.Conclusion

Conclusion: This paper puts forward a new highway lane congestion recognition technology by means of multi-source data. A Bayesian network plus SVM data fusing framework is proposed, it effectively combines all kinds of traffic data sources and enhances congestion recognition accuracy and robustness compared with traditional single-source means of transport. From

the experimental results we can see the efficiency of this method is proven, and the result of recognizing has been improved. The studies shows that using more data sources will give you a better identification of congestion as well as giving a good base for future works of traffic systems. And continue improving the above-mentioned mechanism, addressing its limitations, I believe this method will be able to be popularly used in managing real highway traffic congestion, which would ultimately result in faster traffic flow and better traveling experience.

Funding

no

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Fu L ,Xu Y ,Gao S .A multi-agent deep distribution approximation strategy optimization algorithm with multi-threaded parallel computing mechanism suitable for large-scale and complex urban road networks[J].Engineering Applications of Artificial Intelligence,2025,155110999-110999.

[2] Haldar S ,Mondal A ,Maity R , et al.ANN based traffic congestion analysis applied to parking recommendation system for electric three-wheelers[J].Engineering Research Express,2025,7(2):025242-025242.

[3] Chen S ,Zhao H ,Dou H .Short-Term Traffic Flow Prediction and Congestion Mitigation Using Generalized Regression Neural Network and Low-Rank Matrix Recovery in Urban Road Networks[J].Journal of Circuits, Systems and Computers,2025,(prepublish):

[4] Selvan C ,Kumar S R ,Joseph T I S , et al.Traffic Prediction Using GPS Based Cloud Data Through RNN-LSTM-CNN Models: Addressing Road Congestion, Safety, and Sustainability in Smart Cities[J].SN Computer Science,2025,6(2):159-159.

[5] Anciaes P ,Cheng Y ,Watkins J S .Policy measures to reduce road congestion: What worked?[J].Journal of Transport & Health,2025,41101984-101984.

[6] Karafyllis I ,Theodosis D ,Papageorgiou M , et al.From road congestion to vehicle-control enabled artificial traffic fluids[J].Annual Reviews in Control,2025,59100989-100989.

[7] Rammutla J .Building pathways to digitally transformed, sustainable and safer roads[J].Civil Engineering: Magazine of the South African Institution of Civil Engineering,2024,32(11):22-23.

[8] Samad H ,Wunas S,Jinca Y M, et al.Reliability of Road and Rail Transportation in the Distribution of Goods Commodities on the Makassar-Parepare Route[J].Journal Européen des Systèmes Automatisés,2024,57(5):

[9] Güven G .Analytical Explanation for the Effects of Working from Home on Optimal Environmental Road Pricing[J]. Transportation Research Record,2024,2678(10):1252-1272.

[10] Xiaohan Z ,Shaopeng Z ,Ao L , et al.Review on road congestion pricing: a long-term land use effect perspective[J]. Proceedings of the Institution of Civil Engineers - Transport,2024,1-32.