Spatial-Temporal Data Mining in Traffic Incident Detection

Ying Jin, Jing Dai, Chang-Tien Lu

Department of Computer Science, Virginia Polytechnic Institute and State University

{jiny, daij, ctlu}@vt.edu

Abstract — Real time traffic incident detection is critical for increasing safety and mobility on freeways. There have been incident detection approaches based on traffic behavior or mathematical models proposed for this task. However, earlier incident detection methods are limited in distinguishing recurrent and non-recurrent congestions. The complexity of current approaches makes them insufficient to handle the real time task. In this paper, a new approach for detecting incidents is proposed. Different from traditional traffic incident detection methods, both spatial and temporal information are considered to find the potential incidents. Meanwhile, adaptive learning ability and short detection response time are achieved in the new method. To analyze the high dimensional traffic data, Mahalanobis distance is applied to discover potential incidents according to the traffic pattern. Lifeline style detection and visualization is utilized to provide intuitive user interface. Methodology analysis and preliminary evaluation have been performed to validate the detection effectiveness on the integrated traffic visualization system.

Keywords — incident detection, spatial-temporal data mining, visualization

1 Introduction

Studies on transportation congestions have shown that, freeway incidents cause approximately 60 percent of all urban freeway delays in the United States [1]. As such, accurate and fast traffic incident detection is critical for minimizing traffic delays and increasing safety. There are two major usages of automatic incident detection in a traffic management system. The first is to signal the dispatch of emergency crews for medical support, obstruction removal, and general safety maintenance; the second is to provide useful information to the routing control system to maintain and optimize system-wide performance. As traffic data streams arrive in varies of speed with large amount, quick and reliable automatic detection of traffic incidents becomes a key issue in managing freeway systems.

The traffic incident detection problem can be viewed as recognizing incident patterns from observed data series obtained from loop detectors. A number of incident detection algorithms have been developed over the past three decades [4, 6, 11, 12]. The major disadvantages of earlier algorithms are their unreliability in differentiating between recurrent and non-recurrent congestion events resulting in a high false alarm rate. In recent years, computational intelligence approaches including neural-computing, evolutionary computing, wavelet analysis, and fuzzy logic have been employed to solve the complex and mathematically intractable incident detection problems [2, 3, 5, 8, 13]. However, most of them are based on single station pattern, i.e., geospatial neighborhood relationships between stations do not involve in these pattern detection methods.

Data received from detectors on the freeway are not only temporal related but also have spatial features in nature. For example, detectors on a highway (shown in Figure 1) can be treated as points on a straight line. To handle these data, a new incident detecting method is proposed in this paper to provide a real-time spatial-temporal pattern mining approach, specifically for the traffic data. Mahalanobis distance is applied to consider covariance of the detector stations along the freeway, and along the time line. In addition, a graphic interface is designed to display the real-time incident alarm level and collect feedbacks for adaptive revision of the system. An incrementally learning method is proposed as well to keep the historical traffic model up-to-date. Proposed incident detection is implemented based on the existing AITVS (Advanced Interactive Traffic Visualization System)[9] to perform real-time incident detection on Interstate 66 (I-66 shown in Figure 1). The major contributions of this incident detection approach are listed as below: 1) Consider both spatial and temporal information in detection, 2) Correlations of traffic data between stations and between time slots are counted, 3) Real-time lifeline style visualization is provided for effective navigation, and 4) The approach has been validated using traffic data on I-66.

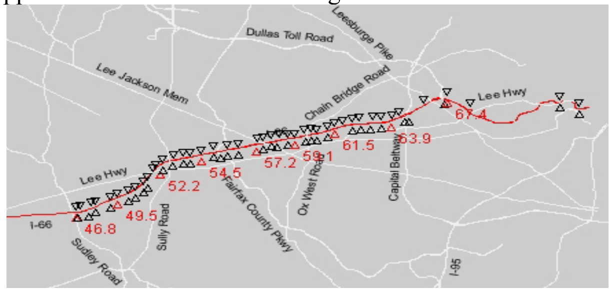


Figure 1. Locations of all detectors/stations on I-66.

The rest of this paper is organized as follows. The related work is discussed in Section 2; the proposed method is explicated in Section 3; important parameters and detection effectiveness are discussed in Section 4; implementation and demonstrations are illustrated in Section 5; finally, we conclude our work and present future work in Section 6.

2 Related Work

A number of incident detection algorithms have been developed over the past three decades [4, 6, 11, 12] and some of them have been deployed in urban freeway systems in selected areas. One of the earliest and most popular algorithms is the California Algorithm [11]. This algorithm is based on the logical assumption that a traffic incident increases the traffic occupancy upstream of the incident and decreases the traffic occupancy downstream of the incident significantly. However, these earlier algorithms are not reliable in differentiating between recurrent and non-recurrent congestion events. Persaud *et al*. [12] propose a single station algorithm known as McMaster algorithm, where congestion is detected using traffic volume, occupancy and speed data from a single station and the parameters are related using catastrophe theory. The Minnesota Algorithm [4] attempts to minimize false alarms and missed incidents, by filtering out the effects of high frequency random fluctuations in traffic flow using averaging occupancy measurements over contiguous short-term intervals. In order to provide more accurate detection, computational intelligence approaches have been employed in recent years. To reduce the dimension of the input space without any significant loss of information, a wavelet-based feature extraction approach has been proposed [2, 13], which is computationally efficient and provides greater resolution control over Fourier analysis. The fuzzy-wavelet radial basis function neural network (RBFNN) freeway incident detection model [3], is proposed as a single station pattern-based algorithm. Recently, Karim and Adeli [8] present a two-stage single-station freeway incident detection model based on advanced wavelet analysis and pattern recognition techniques. An energy representation of the traffic pattern in the wavelet domain is found to best characterize incident and non-incident traffic conditions. However, as all of them, except California Algorithm, are based on single station pattern, the spatial relationships between stations have not been taken into account for pattern detection.

Since data collected from freeway detectors have an inherently temporal and spatial context, the time and space components must be taken into consideration in the mining process in order to accurately interpret the collected data. Recently, spatial-temporal pattern mining has attracted many research efforts [7, 10]. We propose a new incident detection method in this paper based on spatial-temporal data mining techniques. Comparisons between the proposed approach and previous methods have been summarized according to three measures: considering multiple stations, distinguishing recurrent and non-recurrent events, and incremental learning, as shown in Table 1.

Table 1. Comparison between Incident Detection Methods.

Incident Detection Method	Multiple Stations	Distinguish Recurrent & Non-recurrent Events	Increment - learning
California	√		
McMaster		√	
Minnesota		√	
Wavelet-neural		√	
RBFNN		√	
Proposed Method	√	√	√

3 Proposed Approach

The proposed approach includes spatial-temporal data mining and visualization components. To detect incidents, the system will generate spatial-temporal traffic models for each day-of-week based on speed, volume, or occupancy, and then identify the outliers based on comparing real time traffic data with the historical models. The traffic model is defined as the typical day-of-week traffic data with incremental learning ability. The outlier will be visualized to illustrate the alarm level, position, and time. The detection process is illustrated in Figure 2. This figure shows how the traffic incident on Wednesday is detected. In each chart, the *Y*-axis represents the mileposts, *X*-axis denotes time intervals, and colors represent occupancy values, where each row represents the occupancy over a whole day at one particular station. At first, all the traffic data (occupancy in this example) of historical Wednesdays are collected from the traffic archive. For each station, a vector of daily occupancy that describes average value of each time slot is calculated from the historical Wednesdays' occupancy. Thus the occupancy model of Wednesday is generated for all stations. Then at each time the occupancy values are received from the stations for a certain time slot, say, five minutes, a comparison is made between the vector of current occupancy from all stations and the occupancy vector of all stations from the model at the same time slot. If the current vector varies greater than a predetermined threshold to the vector from Wednesday occupancy model, an alarm will be triggered for reporting potential incident at that time slot.

Considering the correlation of the detector stations along the freeway and the correlation along the time line in mining process, Mahalanobis distance is used to measure the difference among traffic data. Mahalanobis distance is superior to Euclidean distance because it accounts for ranges of acceptability and compensates for dependencies between variables. Mahalanobis distance *D* between sample data *H*

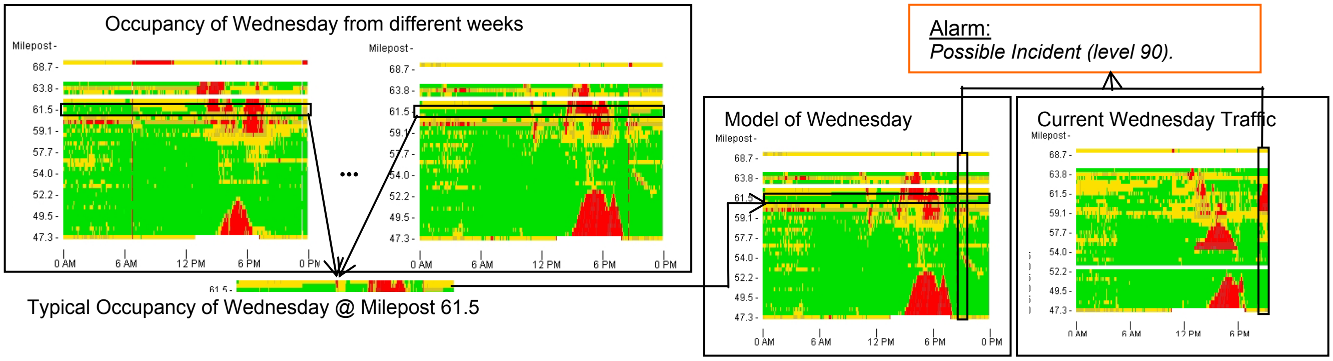


Figure 2. Incident Detection Process

and data model $\boldsymbol{\mu}$ is defined as $(H\text{-}\boldsymbol{\mu})^T\Sigma^{-1}(H\text{-}\boldsymbol{\mu})$, where Σ is the variance-covariance matrix of H and $\boldsymbol{\mu}$. For instance, when current traffic is compared with traffic model (the right part in Figure 2), H is the vector of occupancy of all stations on current time interval, and $\boldsymbol{\mu}$ is the corresponding column in model with same time interval. An important character of Mahalabobis distance is that when H follows multivariate normal distribution, D will follow Chi^2 distribution. This character helps to identify the probability of occurrence of a data point, which will be used to determine the outlierness.

Users are allowed to provide feedbacks to validate the correctness of outlier (incident) detection. The feedback helps to refine the model and improve the accuracy of detection.

The incident detection process can be divided into four steps as listed below. The implementation details will be provided in Section 4.

Step 1: Data Clean & Preparation

In this task, the raw data retrieved from traffic detectors will be cleaned and organized for the mining stage. As the data received from loop detectors contains noise and missing values caused by malfunction of the detector or transmission problems, a data cleaning must be performed to identify and remove these data to assure the data quality. On the other hand, the expected traffic daily models are different from weekdays to weekends. Public holidays also have unique traffic patterns. Therefore, a categorization is needed to separate the traffic data into different weekdays and weekends. To summarize this task, three subtasks are listed as follows:

1. Scan the database and identify the abnormal records from malfunction detectors.
2. Label the records of public holidays.
3. Categorize data of different weekdays and weekends by building separate data views.

Step 2: Traffic Model Generation

In the model generation task, traffic data on different weekdays will be analyzed to construct traffic models respectively. While generating the model, some non-recursive incidents in the cleaned data sets, which appear to be outliers, must be removed to refine the traffic models. Temporary models will be calculated as the average of historical traffic data. For example, average speed value at 12 PM on historical Wednesdays of station $\boldsymbol{x}$ will be used as the speed at 12 PM on Wednesday of station $\boldsymbol{x}$ in the model. Having the temporary model, Mahalanobis distance will be calculated between the station-daily values in model and in historical data samples. When the distance is greater than a predetermined threshold ***d1***, the corresponding historical station-daily sample will be removed from the model. After eliminating all the outliers, the final traffic models will be generated as the average of the remaining historical data. Three steps are listed to summarize this process.

1. Calculate the **mean** value for each daily-station traffic data view for different weekdays and weekends.
2. Determine the incidents in historical data using **Mahalanobis** distance to.
3. Compute the **mean** value after removing the outliers for each daily-station traffic data view as the final models.

Step 3: Detecting Incidents

This task will be executed in real-time to discover potential incidents based on the traffic model. The traffic data, which is updated every five minutes, is collected from the loop detectors and cleaned in runtime. Detection will be performed by calculating the Mahalanobis distance between real-time data and the corresponding time slot in the traffic model. If the distance calculated is greater than threshold ***d2***, a possible incident will be identified and a certain level of alarm will be reported to the traffic operator. In case that the possible incidents have been detected in consecutive time slots, which indicates the high potentiality of real incident, the alarm level will increase to a certain value ***AL(t),*** which is a function of number of consecutive time slots ***t***. When no possible incidents are detected, the alarm level will decrease until it reaches the safe level. The following steps summarize the incident detection task.

1. Calculate the **Mahalanobis** distance with the corresponding vector in model, using the traffic data for all stations on one time interval as a vector.
2. If the distance is larger than ***d2***, which can be determined by the assumed distribution of Mahalanobis distance, there could be a possible incident occurs at that time point.
3. If consecutive possible incident occurs in ***t*** time slots, the alarm level increases to ***AL(t)***.

Step 4: Incremental Learning Model

The system requires the ability to dynamically learn from incoming traffic data to be adaptable to environment change, such as road construction and region development. The continuously coming traffic data without true incidents will be used to refine the model. After detecting the incident, user's feedbacks on detection accuracy will be collected. If the possible incident is verified as a true incident, the traffic data collected during the incident period will not be used in the model. By merging the new traffic data into the original model, the traffic model will be updated. The formula to calculate each new daily-station traffic pattern new_DS is defined as new_DS = (1-***f***)*old_DS + ***f****new_data, where **fading factor *f*** is in (0, 1). The value of ***f*** determines the learning rate of the traffic model.

To perform detection in real-time, the variance-covariance matrices for each time slot should be calculated beforehand, because it is the most time-consuming process and would delay the response of detection if being calculated on demand. Therefore, when generating the updated traffic model, the variance-covariance matrices should be calculated as well. Both the matrices and traffic model will be stored for further incident detection. These steps are summarized as follows.

1. Calculate the mean value of the existed model and the new data with a **fading factor *f*** (0<***f***<1) as the new model.
2. Compute the matrices for each time point in order to calculate the **Mahalonobis distance** in next week.
3. Store the model and matrixes.

4 Methodology Analysis

In the proposed incident detection approach, the value of the parameters will greatly impact the effectiveness of detection. Specifically, these parameters are, distance thresholds ***d1*** and ***d2***, alarm level function ***AL(t)***, and fading factor ***f***. Their values as well as their impacts for incident detection will be discussed in this section.

There are two distance thresholds defined in the detection process. Distance threshold ***d1*** is used to identify the outliers when initially generating the traffic model; ***d2*** is used to determine the possible incidents in real-time detection. As described in Section 3, Mahalanobis distance follows Chi^2 distribution, ***d1*** and ***d2*** can be assigned using the probability of incident occurrence. Assuming there would be 5% of the days in which incidents would occur in certain station, ***d1*** can be assigned with the value which has the density of 95% Chi^2 distributed variable with the degree of freedom 288 (number of time slots). For distance threshold ***d2***, in case that there will be 2% of the time slot in a certain weekday or weekend in which incidents would occur, ***d2*** can be defined as the value for range of 98% Chi^2 distributed variable with freedom degree as number of stations on the direction.

Alarm level function ***AL(t)*** generates a lifeline style representation to the belief degree of incident. When the first time a possible incident is detected, a low initial alarm level ***AL(1)*** will be assigned. Once the incident is detected in *t* consecutive time slots, the alarm level will increase to ***AL(t)***, until it reaches a limitation (***AL*** is ranged from 0 to 100). In the contrary, when no more possible incident is detected, the alarm level should decrease, until it reaches a limitation, which is a safe level. To summarize the discussion, definition of ***AL(t)*** is given as:

- ***AL(t)*** = Min(***AL(t-1)*** + *k*, 100), when incident detected in time slot *t*;
- ***AL(t)*** = Max(***AL(t-1)*** – *k*, 0), when no incident detected in time slot *t*.

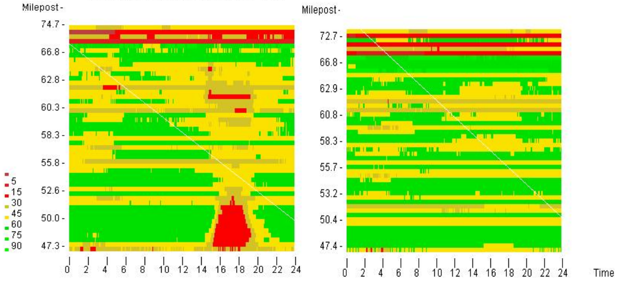
Therefore, ***AL(t)*** will vary from 0 to 100, and higher value indicates more potentiality to be an incident. Constant *k* can control the increasing and decreasing rate of ***AL(t)***.

Fading factor ***f*** is used to determine the learning rate of the traffic model. The value of ***f*** should be consistent to the traffic environment. If the traffic environment is changing rapidly due to road constructions, routing policies, and weather conditions, ***f*** should be assigned a relatively high value to reflect on a short term impact. Otherwise, ***f*** can be small to make the model relatively stable for a long term model. Using fading factor ***f***, learning rate can be conveniently configured. For example, if we are going to mainly consider the traffic data of the most recent 10 weeks, we can define ***f*** to make the data of 11th recent week contribute less than a certain small portion in the traffic model.

5 Implementation & Case Study

We implemented the spatial-temporal incident detection component in AITVS [9]. Based on this system, a spatial-temporal data mining component, user feedback function and the corresponding visualization are implemented for incident detection task. Data fusing and preparing is performed in this service to clean and fuse the data before store them into traffic database. A data processing module, which contains spatial data modeling and spatial-temporal mining functions, is used to organize the traffic data, and to discover inherent patterns from traffic archive.

The historical data is collected from Virginia Department of Transportation (VDOT) every five minutes. In the implementation, speed is used for incident detection, because it is an appropriate measurement to indicate the congestion. The daily traffic models for Thursday on West Bound and Sunday on East Bound are illustrated as examples in Figure 3, where *X*-axis represents the time, and *Y*-axis represents the milepost of stations. The data shown in these two traffic models look smooth, except several malfunctioned stations, as the outliers have been eliminated. These malfunctioned stations are still reflected in the traffic model because they have never functioned according to the historical data. In addition, the speed patterns reflect the traffic situation properly. As we can find from the Figure 3, there are recurrent congestion from 3pm to 7pm on Thursday's west bound, which is mainly reflected as a red triangular region at the bottom of the chart, while the traffic on Sunday looks just smooth all the day.



(a) Thursday-West Bound (b) Sunday-East Bound

Figure 3. Daily Traffic Speed Models

Known historical incidents are used to validate the effectiveness of the proposed incident detection approach. In the implementation we assume outliers are 5% of all data samples. In Figure 4 and Figure 5, incident detection results on a normal Thursday and a Sunday which contains incident (5/1/2005) are illustrated. In each figure, the Alarm Level (AL) is visualized as the top horizontal color bar. Legend for the alarm level is shown on the right side of the figure. **Tolerance**: In a series of tests, the system reports low alarm levels on some noise, i.e., light and non-recurrent congestions, and passes the recurrent congestions. Comparing Figure 3 (a) and Figure 4, the sample traffic data on Thursday are quite similar to the model on west bound. The alarm levels in Figure 4 are usually low, even zero in most time, although there are recurrent congestions, missing data (blank vertical strips) and malfunctioned stations (blank horizontal strips). The results in this figure show that the incident detection can properly handle missing data and noise, as well as to distinguish non-recurrent congestions from recurrent congestions. **Effectiveness**: The system reports top alarm levels (100%) to all the seven known real incidents. In situations other than incidents in our experimental cases, the alarm level never exceeds 80%, while is kept from 0% to 30% in most of the time. In Figure 5, there is a noticeable red bar in the top horizontal line, indicating a high alarm level lasting for a long time. The position of the red bar in the figure is right after the congestion occurs at one station, and before it expands to multiple adjacent stations. This result shows the detection approach can effectively identify the incident using the daily traffic model. **Quick response**: In the experiments, no additional response time for detection and visualization is required, except the time to load a static web page in browser, i.e., usually less than 3 seconds (depends on the network). The major reason is that there is a background program doing the real-time detection and drawing the charts all the time. When the detection results are requested from the web, a corresponding HTML page will display the recently completed charts. Therefore, the response time only depends on the network connection. Furthermore, the detecting program usually finishes detecting for one time slot and draws the charts in approximately one second. Performing the tests on known incidents and normal weekdays, our approach shows its effectiveness, efficiency, and the ability to handle real time and noise data in real traffic management system.

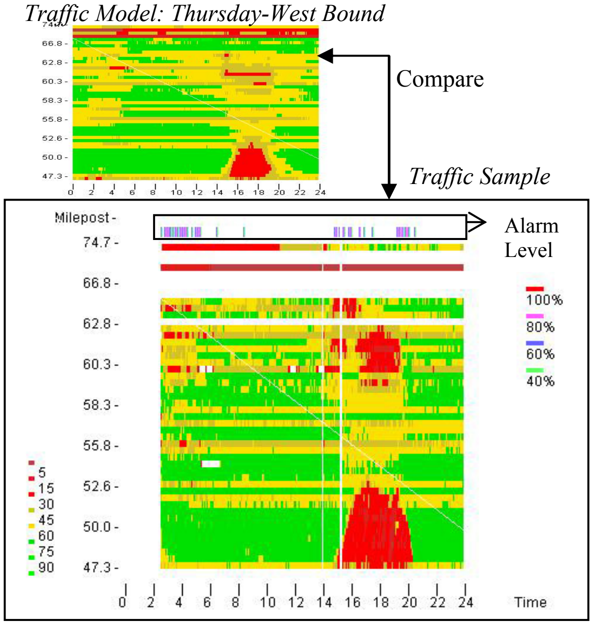


Figure 4. Sample Thursday West Bound (no incident)

6 Conclusion & Future Work

In this paper, we propose a new method to identify incident in real time. It is based on spatial-temporal data view and applies Mahalanobis distance to consider the correlation of traffic data from neighboring stations and consecutive time slots. A lifeline-style alarm level for incidents is implemented to support effective data navigation. Our approach utilizes user feedback to support learning ability. Moreover, fast response time is achieved by using active detection strategy. A set of tests have been conducted in real system to validate the effectiveness and efficiency of this approach. Future efforts will be needed to refine the parameters in this approach. For incremental learning, re-calculating variance-covariance matrices costs extensive system resources. Incrementally update or approximate computing techniques can be applied to improve the computational efficiency of the proposed method. This approach can also be applied to other applications which consider both temporal and spatial features, such as disease control and weather monitoring.

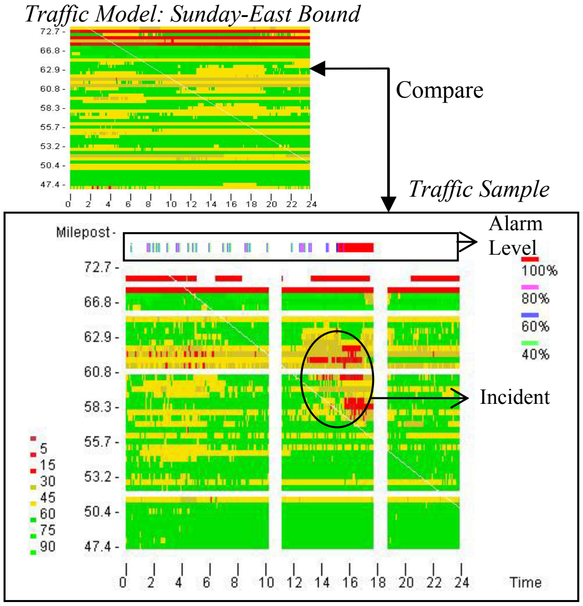


Figure 5. Sunday 5/1/2005 East Bound (traffic incident)

References

[1] "Freeway Incident Management Handbook," *Federal Highway Administration*, *http://www.its.dot.gov/jpodocs/rept_mis/@9201!.pdf*, 2000.

[2] H. Adeli and S. L. Hung, "Machine Learning –Neural Networks, Genetic Algorithms, and Fuzzy System," in *Distributed Computer-Aided Engineering*. New York, 1995.

[3] H. Adeli and A. Karim, "Fuzzy-Wavelet RBFNN Model for Freeway Incident Detection," *Journal of Transportation Engineering*, vol. 126, pp. 464-471, 2000.

[4] A. Chassiakos and Y. Stephanedes, "Smoothing algorithms for incident detection," *Transportation Research Record* pp. 8-16, 1993.

[5] R. Cheu and S. G. Ritchie, "Automated detection of lane-blocking freeway incidents using artificial neural networks," *Transportation Research-Part C: Emerging Technologies*, vol. 3, pp. 371-388, 1995.

[6] A. R. Cook and D. E. Cleveland, "Detection of freeway capacity-reducing incidents by traffic stream measurements," *Transportation Research Record*, vol. 495, pp. 1-11, 1974.

[7] V. Iyengar, "On detecting space-time clusters," *In Proceedings of ACM Conf. on Knowledge Discovery in Data Mining*, pp. 587-592, 2004.

[8] A. Karim and H. Adeli, "Incident detection algorithm using wavelet energy representation of traffic patterns," *Journal of Transportation Engineering*, vol. 128, pp. 232-242, 2002.

[9] C. T. Lu, A. P. Boedihardjo, and J. Zheng, "AITVS: Advanced Interactive Traffic Visualization System," *To appear in the Proceedings of IEEE International Conference on Data Engineering*, Apr 3-8, 2006.

[10] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel, "Detection of Emerging Space-Time Clusters," *In Proceedings of ACM Conf. on Knowledge Discovery in Data Mining*, pp. 218-227, 2005.

[11] H. J. Payne and S. C. Tignor, "Freeway incident detection algorithms based on decision tree with states," *Transportation Research Record*, vol. 682, pp. 378-382, 1978.

[12] B. N. Persaud, F. L. Hall, and L. M. Hall, "Congestion identification aspects of the McMaster incident detection algorithm," *Transportation Research Record*, vol. 1287, pp. 167-75, 1990.

[13] M. Wu and H. Adeli, "Wavelet-neural network model for automatic traffic incident detection," *Mathematical & Computational Applications*, vol. 6, pp. 85-96, 2001.