Anomaly Detection on ITS Data via View Association

Junaidillah Fadlil
nedijf@gmail.com

Hsing-Kuo Pao
pao@mail.ntust.edu.tw

Yuh-Jye Lee
yuh-jye@mail.ntust.edu.tw

National Taiwan University of Science and Technology
No. 43, Sec. 4, Keelung Rd.
Taipei, Taiwan 106

ABSTRACT

We focus on detecting anomalous events in transportation systems. In transportation systems, other than normal road situation, anomalous events happen once in a while such as traffic accidents, ambulance car passing, harsh weather conditions, etc. Identifying the anomalous traffic events is essential because the events can lead to critical conditions where immediate investigation and recovery may be necessary. We propose an anomaly detection method for transportation systems where we create a police report automatically after detecting anomalies. Unlike the traditional police report, in this case, some quantitative analysis shall be done as well to provide experts with an advanced, precise and professional description of the anomalous event. For instance, we can provide the moment, the location as well as how severe the accident occurs in the upstream and downstream routes. We present an anomaly detection approach based on view association given multiple feature views on the transportation data if the views are more or less independent from each other. For each single view, anomalies are detected based on a manifold learning and hierarchical clustering procedures and anomalies from different views are associated and detected as anomalies with high confidence. We study two well-known ITS datasets which include the data from Mobile Century project and the PeMS dataset, and we evaluate the proposed method by comparing the automatically generated report and real report from police during the related period.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern analysis*; H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support (e.g., MIS)*

General Terms

Algorithms

ODD'13, August 11th, 2013, Chicago, IL, USA.

Keywords

Anomaly Detection, Intelligent Transportation System (ITS), Association, Trajectory, Manifold learning.

1. INTRODUCTION

We have entered an era where sensor devices are massively utilized to monitor the environment around us. Sensors can communicate between each other and can communicate with backend systems as well. Based on that, we can distributively collect data from regional sensor readings to profile regional patterns for further follow-up examinations. In transportation monitoring, not too long ago people in USA call 911 to report accidents in traffic when accidents occur. As time goes by, deployed sensors on roadside are now commonly used for traffic information collection. Given the traffic information, we can understand traffic status so that traffic police and drivers like us can take appropriate actions afterwards, so called the Intelligent Transportation System (ITS). In recent applications, smartphone devices have also been included as part of the monitoring system.

Since roads are covered by the sensor-based information system, many technologies have been applied. Examples include incident detection systems [13] where detection of incident can significantly reduce the number of unnecessary highway patrols; automatic plate number recognition [3] for the purpose of surveillance and traveling time estimation; traffic signal control system [14] to help us for traffic flow optimization[1]. Many of the above technology can also be combined together as an integrated system to built a more complex ITS. In metropolitan area, such ITS becomes necessary in all respects. In this work, we focus on incident detection based on an anomaly detection approach.

Given the data collected on an ITS, the purpose of this study is to detect anomalies within the traffic which may be due to incidents, and to estimate the influence of anomalous events in nearby incident area such as upstream and downstream routes of the incident location. That is, we intent to produce a report automatically that is similar to a police report which includes all necessary information about an incidents or an anomalous event; furthermore, we would like to add additional quantitative information to extend the police report. For instance, a police report may record information merely about an accident including the accident location, when the accident happened, and what kind of accident such as traffic collision with unknown reason, hit and

[1]Zhang et al. [22] reported that almost 40% of the population spends at least one hour on the road each day in USA.

run, traffic collision with injuries, and so on. However, the accident might affect the nearby area, as illustrated in Figure 1. As depicted in the illustration, traffic in upstream routes is likely to be more severely congested than that in the downstream routes. This kind of information, even it is useful for drivers and police, is usually not included in the police or ITS report.

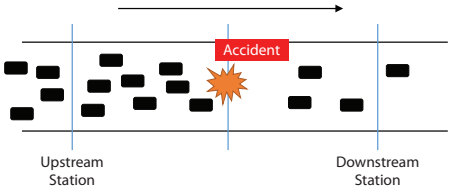


Figure 1: Illustration of traffic accident, and the status near the upstream and downstream routes.

We propose an anomaly detection method that can detect traffic anomalies by feature view association. Given a multi-viewed dataset, we assume that the information of different views are collected separately and there exists no *contextual anomalies* across different views[2] and the anomalies can be found within each single view. Based on a few anomaly detection results from different views, and to associate anomalies detected from those views, we can confirm the anomalies with high confidence.

The evaluation is done mainly on the *Mobile Century* dataset which is a well-known ITS dataset for traffic analysis. Moreover, to apply the proposed method to a relatively large-scale dataset, and to further study the spatial and temporal relationship between data, we also test the proposed method on the *PeMS* data, another well-known ITS dataset. Details of the two datasets are shown in Section 3.1.

The benefit of the proposed anomaly detector can be summarized as follows:

1. Different from most of the one-class anomaly detection methods, the proposed method needs very few parameters or threshold tuning to decide how likely to be considered as normal patterns.
2. In principle, the proposed method needs no "clean" data for the training of normal patterns. In general, the search of clean data can be difficult, or as arbitrary as suggested by subjective domain experts.
3. The computation of the proposed method is efficient in the sense that the computation on each single feature view can be done separately. Therefore, we can easily extend the algorithm to a parallel version when multiple-process or parallel computation is available.

We expect that the proposed method can be applied to applications other than the traffic incident detection. We should also expect that the proposed method can be extended to solve the anomaly detection task on large-scale datasets.

Before we go on to introduce the proposed method, we discuss some previous works on anomaly detection and traffic analysis in Section 2; after that, we present the datasets that we use in this work in Section 3.1, and in Section 3.2, we describe the proposed method. The experiment result is shown in Section 4 and in Section 5, we conclude our presentation.

[2]There may exist contextual anomalies *within* a single feature view though.

2. RELATED WORK

Many studies have focused on traffic analysis and one major approach is to detect anomalous events from traffic patterns. For example, given taxi trajectories recorded from GPS, Chawla et al. [4] proposed a framework to infer the main reason why some anomalies appear in road traffic data. In their framework, they modeled the road structures as a directed graph then employed PCA algorithm to detect anomalies. Chen et al. [5] proposed *iBOAT* that can detect anomalous trajectories "on-the-fly". They extracted useful information from the behaviors of urban road users, and analyzed adverse or possibly malicious events such as a driver taking a questionable route.

In many anomaly detection schemes, an effective data representation can reveal the difference between normal and anomalous patterns. Thajchayapong et al. [20] monitored traffic anomalies using microscopic traffic variables such as *relative speed* and *inter-vehicle spacing*. By using Gaussian process to model the microscopic traffic variables, they can grab temporary changes in the traffic pattern and detect anomalies.

Several techniques have been proposed to detect anomalies in video surveillance. Fu et al. [8] proposed using a hierarchical clustering framework to classify vehicle motion trajectories in real traffic video. They showed that their proposed method performs better than the conventional fuzzy K-means clustering. Piciarelli et al. [15] clustered the trajectories in an online fashion, and modeled the trajectory data in a tree-like structure where some probability information is also included. Jiang et al. [10] proposed video event detection based on unsupervised clustering of object trajectories, which are modeled by Hidden Markov Model [16]. This study employ a dynamic hierarchical process for trajectories clustering to prevent model overfitting together with a 2-depth greedy search for efficient clustering.

Similar to our approach, Agovic et al. [1] investigated an anomalous cargo using a manifold embedding method for feature representation. Although, they focused on both linear and nonlinear methods, the paper results show that nonlinear methods outperform the linear methods. Also related to our research, Kind et al. [12] proposed feature-based anomaly detection that constructs histograms of different traffic features. In a survey paper for the outdoor surveillance task, Zhang et al. [23] compared six different similarity measures that are used for trajectory clustering. They showed that the hybrid PCA [11] and Euclidean distance combined method outperforms other methods; and PCA method is very sensitive to its parameters. Ringberg et al. [17] studied the sensitivity of PCA for the anomaly detection. They pointed several challenges in their work such as evaluating how sensitive the false positive rate to small differences of the dimensionality in normal space, and to the level of aggregation in traffic measurement. Danilo et al. [21] studied the estimation of cellular network to build road traffic system. Hence, based on the explorative analysis of real-time signaling data the result showed whether the traffic is normal or abnormal.

To speak of methodology, the proposed method is also inspired by the well-known *co-training* algorithm developed by Blum and Mitchell [2] for semi-supervised learning. The

co-training algorithm splits data attributes into several subsets, given the assumption that the attribute subsets are conditionally independent with the known label information. Each subset plays a view and is *sufficient* to learn a classifier; therefore, it can use the prediction from one view to help other views to learn the label information of unlabeled data. The multi-view approach proposed in this work is similar to the co-training method in the sense that we also use information from different views to decide anomalies.

3. DATASETS AND PROPOSED METHOD

In this section, we explain the proposed anomaly detection method in detail. In order to build intuition on the method, we describe the datasets that are used in this study before the method and then we can illustrate ideas through concrete examples.

3.1 Datasets

We evaluate the proposed method through two datasets: the first dataset is the one from *Mobile Century* project [9], a traffic dataset that was collected on February 8, 2008 along a 10-mile stretch of I-880 highway near Union City, California, USA, for over eight hours (10:00 AM - 18:00 PM). In this work, we focus on two parts of the dataset, the *GPS* individual trajectories and the loop detector *PeMS* data. The second dataset is the *PeMS* dataset from California Department of Transportation (Caltrans) website[3]. The Caltrans website has been keeping records starting from 1993. In order to differentiate between *PeMS* data from *Mobile Century* project and *PeMS* data from California DOT, we call the first *PeMS Century PeMS* and and the latter *PeMS* data *Caltrans PeMS*.

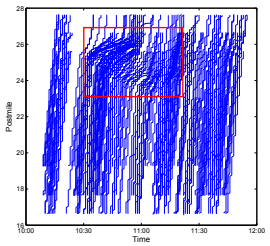


Figure 2: Plot of individual trajectories produced by *GPS*; the red square indicates the space and time information of an accident. In the *Caltrans PeMS* website, the accident report recorded the accident for only one station (postmile 26.641); however, as shown in this plot, the accident propagates and gives effect to the upstream and (a little to) the downstream stations.

Figure 2 shows the *GPS* data that indicates where and when the accident happened. The *Century PeMS* data were recorded for every 30 seconds while *GPS* data were recorded for every 3 seconds. The *GPS* data contains information about latitude and longitude. We shall use the information to associate with the location information of *PeMS* stations. Moreover, we just sample the trajectories that start from 10:00 AM and end at 11:45 AM as our data in this study.

[3]http://pems.dot.ca.gov/

The data provided by the *Caltrans PeMS* website are aggregated for every 5 minutes. In the *Caltrans PeMS* experiment, we select data close to postmile 26.641 (station 400165), starting from 10:00 AM until 18:00 PM.

Ground Truth.

The ground truth was obtained from *Caltrans PeMS* website which records incidents on several California freeways including accidents, traffic hazards, congestions, traffic breakdowns, and so on. On the *Mobile Century* data for the first experiment, we have an accident reported on postmile 26.641, occurred from 10:34 AM, February 8, 2008, with a duration of 34 minutes. In the second experiment, the *Caltrans PeMS* reported an accident on the same postmile (26.641), which occurred at 1:00 PM, December 14, 2007, with a duration of 38 minutes.

3.2 Method

In this subsection, we explain the proposed anomaly detection method in full details. As shown in Figure 3, the main purpose of our work is to create a report for traffic incidents given different *views* (features) of data. The proposed method consists of four steps. The first step is to extract useful features from the data. Second, we utilize a manifold embedding method called Isomap [19] for data representation. After that, in the third step, we cluster the projected points using a hierarchical clustering method [18]. We detect anomalies based on the hierarchical clustering result and that is done for each single feature view. Overall, we may have anomalies that are detected based on several individual views. In the end, the last step is to automatically create a report based on the different views' hierarchical clustering and anomaly detection result obtained from the previous step. To produce the final report, we associate anomalies that are detected from different views and make the final call of anomalies if they belong to the anomalous group in many different views. By detecting anomalies from different views, we believe that we can have high confidence on making the final decision which may include dispatching a police patrol to the accident location for further investigation and accident recovery.

3.2.1 Feature Extraction

In this study we use three views for anomaly detection to create incident report in *Mobile Century* data. The first view is based on the *flow* information and the second view is based on the *speed* information, which are obtained from *Century PeMS* and *GPS* data respectively. The third view is based on the *duration* information, derived from *GPS* data. We use two views for anomaly detection in *Caltrans PeMS* data. The first view is *flow* and the second view is *speed*. Feature extraction on each view is defined as follows:

- **Flow.** The *flow* data are obtained by temporal sensors. We use an appropriate time window size w to extract the features. Let $Q = \{x_1, \ldots, x_T\}$ to be a T-length time series of *flow*, we discuss many derived

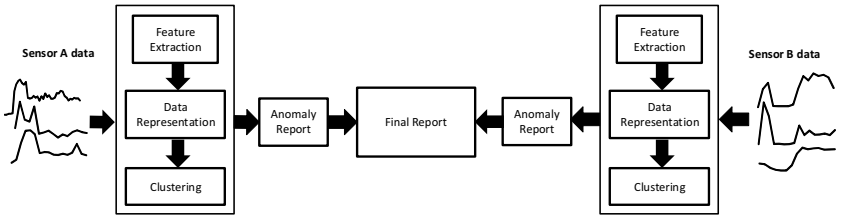


Figure 3: The proposed anomaly detection method. The final report is created based on the anomalies detected from different views, such as sensor readings of different locations, different types of sensors, different measurements, etc.

features as follows.

$$
\begin{aligned}
Q &= \bigcup_{i=1}^{N} q_i \\
q_i &: \{x_1^i, x_2^i, \ldots, x_{k+1}^i, \ldots, x_{|q_i|}^i\} = \{x_j, x_{j+1}, \ldots\} \\
&\quad \text{for some } j, \text{ data in } i\text{-th window} \\
|q_i| &: \text{the number of data in } q_i \\
N &: \text{the number of windows} \\
w &: \text{size of each window, } \frac{T}{w} = N\,.
\end{aligned}
$$

Moreover, we define mean of q_i as:

$$m_{flow}^i = \frac{\sum_{x_j \in q_i} x_j}{|q_i|}\,, \tag{1}$$

also, standard deviation of q_i as:

$$s_{flow}^i = \sqrt{\frac{1}{|q_i|-1}\sum_{x_j \in q_i}^{|q_i|}(x_j - m_{flow}^i)^2}\,, \tag{2}$$

and skewness of q_i as:

$$g_{flow}^i = \frac{\sum_{x_j \in q_i}^{|q_i|}(x_j - m_{flow}^i)^3}{(|q_i|-1){s_i}^3}\,. \tag{3}$$

On the other hand, we also compute the difference between the (j-1)-th and the j-th *flow* values for each window q_i. We define L_{flow}^i as:

$$L_{flow}^i = (\ell_1^i, \ldots, \ell_{|q_i|-1}^i) \tag{4}$$

where $\ell_k^i = x_{k+1}^i - x_k^i, k = 1, \ldots, |q_i| - 1$. The features extracted from L_{flow}^i is mean $m_{\triangle flow}^i$, standard deviation $s_{\triangle flow}^i$ and skewness $g_{\triangle flow}^i$. Note that the $|q_i|$ will be the same for each window.

- **Speed.** The *speed* data are obtained from spatio-temporal sensors. First we associate the locations of *GPS* data with the *PeMS* station locations. After that, we collect another set of features that are related to *speed* information. We collect a *speed* time series $V = (v_1, \ldots, v_K)$ with length K. From the data associated with each station, we extract six features:

$$
\begin{aligned}
V &= \bigcup_{i=1}^{M} p_i \\
p_i &: \text{set of speed data in station } i \\
|p_i| &: \text{number of data in station } i \\
M &: \text{number of stations}
\end{aligned}
$$

Table 1: Summary of feature extraction

View	Data Source	Feature
Flow	*Century PeMS*	1. mean of flow 2. std. of flow 3. skewness of flow 4. mean of △flow 5. std. of △flow 6. skewness of of △flow
Speed	*Century GPS* (trajectory)	1. mean of speed 2. std. of speed 3. skewness of speed 4. mean of △speed 5. std. of △speed 6. skewness of of △speed
Duration	*Century GPS* (trajectory)	1. mean of duration 2. std. of duration 3. skewness of duration 4. total duration
Flow	*Caltrans PeMS*	similar to *Century PeMS*
Speed	*Caltrans PeMS*	similar to *Century GPS*

the mean of p_i as m_{speed}^i, the standard deviation of p_i as s_{speed}^i and the skewness of p_i as g_{speed}^i. Similar to the *flow* view we also compute mean, standard deviation and skewness of △speed between the (i-1)-th and the i-th *speed* data. Note that data size for each Station can be different, unlike the case when we collect *flow* features where we have identical number of data in each window.

- **Duration**. In our study we compute duration between the (i-1)-th data and the i-th data for each station that has been passed by vehicle with GPS-enabled smartphones. The way to extract the feature is similar to the *speed* view, but we only extract four features: mean, standard deviation, skewness and total duration to pass a station.

Table 1 gives a summary of the complete feature set. Since we have the data from different features, we normalize them into the range of $[0, 1]$.

3.2.2 Data Representation

Given the features computed in previous subsubsection, first, we compute Euclidean distances between each pair of data points. After that, we utilize a manifold learning method called Isomap for data representation. The Isomap method consists of three steps:

- Construct a neighborhood graph based on k-nearest neighbor (kNN) information.
- Create the shortest path between each pair of data

points. This can be done by, for instance, Dijkstra's shortest path algorithm.
- Apply Multidimensional Scaling (MDS) [6] to find low-dimensional embeddings for data points.

In our study data representation plays an important role. We assume that the space, so-called the intrinsic space is better than the original space to show the relationship between data points. Some detection techniques are indeed more effective if worked on low-dimensional intrinsic space [7]. Moreover, a low-dimensional representation is often desirable for experts to visualize data relationship. To speak of the efficiency issue, working on low-dimensional space usually has low computing complexity.

3.2.3 Data Clustering

Given the projected data in low-dimensional space, we use hierarchical clustering method to cluster the projected data into groups and split the data into two clusters. In our opinion, the normal and anomalous patterns should show difference on the clustering result and therefore be separated into different clusters; hopefully, one for the normal cluster and the other for the anomalous cluster. In real life, the number of anomalous events is usually much smaller than the number of normal events. Hence, to further improve the clustering result we can apply the so-called 90-10 rule. This rule means the normal group should include at least 90% of the whole data points and the anomalous group may include only at most 10% of the whole set. We can confirm the clustering result by the rule to guarantee it is a detection of anomalous behavior rather than a classification of data points into two or more types. Figure 4(b), 5(b) and 6(b) illustrate how clustering has been done and they also show the anomalous events have fewer points compared to the normal points.

3.2.4 Final Anomaly Report

The data will be labeled based on the result of data clustering, for each single view. Afterward, associating the results from two or more views by computing their intersection, will generate final result automatically. That means if all views detect some anomalies occurred in the same location and at the same time, it will be considered as the final predicted anomalies. The final anomaly report contains information about time and location when and where the anomalies happened; also some influence of the anomalies will be included as well.

4. EXPERIMENTS

We divide the experiments into two parts. First, we use the *Mobile Century data* to evaluate the proposed method. We utilize four combined views to detect an anomalous event that happened on Feb. 8, 2008, and the combined views include: ($flow$, $speed$), ($speed$, $duration$), ($flow$, $duration$), and ($flow$, $speed$, $duration$). Second, we study the performance of the proposed method on the *Caltrans PeMS* dataset. In this study we discuss the performance of the proposed method on detecting an accident that happened on Dec. 14, 2007 and we use the only applicable views, namely $flow$ and $speed$ views for the detection. In addition, we show some preliminary online learning result that use previous days to train a model and then test on a later day when the accident happened.

4.1 Experimental Settings

In this series of experiments we set the number of neighbors to be five for kNN which is used in Isomap to build the neighborhood graph, and we set the window size as five minutes and 30 minutes for the *Mobile Century* data and the *Caltrans PeMS* data respectively. In data clustering we use Euclidean distance to compute pairwise distance between points, and use the shortest distance as the distance between clusters to create a cluster tree. Table 2 shows a summary of our experimental settings.

4.2 Experiment Results

4.2.1 Evaluation via the Mobile Century Data

In this section we show the evaluation result given the *Mobile Century* data. We show how we detect anomalies given different view combinations. Some low-dimensional data representation shows that the normal and anomalous patterns are well separated from each other and the proposed method is effective for the detection. Table 2 shows the *PeMS* stations of interest, where the traffic near the stations may be affected by the accident. Note that the dataset is a one-day dataset.

Table 3 shows all the anomaly detection results from each of the following views: (a) *flow* from *Century PeMS*, (b) *speed* from *Century GPS*, and (c) *duration* from *Century GPS*, on six stations. The results indicate that the anomalous events can be detected based on each single view, just may not be perfect, still, we can observe how an accident affects the traffic along the road. As shown previously in Figures 1 and 2 that both of the upstream and downstream traffic can be affected by accidents if there is any. In Figure 2, the red square indicates the propagation of traffic flow from the accident. The y-axis indicates that the traffic close to the nearby postmiles or locations may also become affected. We can observe this kind of information in our detection results; however, this information is usually not recorded in the police patrol report which we consider as ground truth in this study. We can further improve the results if more than one view are considered simultaneously.

Table 4 shows view association results[4] based on different combined views: (a) *flow* and *speed*, (b) *flow* and *duration*,(c) *speed* and *duration*, and (d) *flow*, *speed* and *duration* all together. Overall, the result of anomaly detection by associating *flow* and *speed* views gives the best result if compared to that of all other combinations. On station 400165 where the accident happened, the result from *flow* and *speed* view association reports that the anomalous event starts from 10:35 AM and ends at 10:50 AM, which is very close to the ground truth where it indicates an accident from 10:34 AM to 11:08 AM. In addition, the table shows that the results from all view associations share high correlation to each other.

We also observe that the anomalous and normal patterns are indeed different if represented in low-dimensional space. In Figures 4, 5, and 6 the results are from *flow*, *speed* and *duration* view respectively. The figures show that the anomalous patterns are well separated from the normal ones. The labeling information is just used for visualization and not

[4]Note that the results produced by the *GPS* data and *PeMS* may have different resolutions. Hence, we have to round the data into five minute-based one before computing their association.

Table 2: Summary of experimental settings: $k_{Iso} = 5$ for kNN in Isomap, and the intrinsic dimensionality is set to 2. Data are collected for each 5 or 30 mins to compute the mean, standard deviation, etc., on each station, for *Mobile Century* data or *Caltrans* data respectively.

	Dataset	View	Data Interval	Station ID (Postmile)
Exp. 1	*Century PeMS*	flow	5 mins, 1 station	400488 (24.007) 401561 (24.477) 400611 (24.917) 400284 (25.767) 400041 (26.027) 400165 (26.641)
	Century GPS	speed	1 station	
	Century GPS	duration	1 station	
Exp. 2	*Caltrans PeMS*	flow, speed	30 mins, 1 station	400165 (26.641)

Table 3: The reports produced from different single views, namely *flow*, *speed* and *duration*.

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:50	2/8/2008 10:00	2/8/2008 10:00	2/8/2008 10:30	2/8/2008 10:25	2/8/2008 10:25
2/8/2008 10:55	2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:35	2/8/2008 10:30	2/8/2008 10:30
2/8/2008 11:00	2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:35	2/8/2008 10:35
	2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:40
	2/8/2008 11:00	2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:45
		2/8/2008 11:00	2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:50
				2/8/2008 10:55	

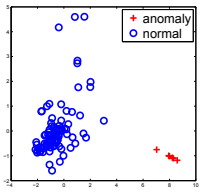
(a) The report produced by *flow* view (*Century PeMS* stations)

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:37	2/8/2008 10:43	2/8/2008 10:43	2/8/2008 10:38
2/8/2008 10:47	2/8/2008 10:41	2/8/2008 10:41	2/8/2008 10:46	2/8/2008 10:51	2/8/2008 10:39
2/8/2008 10:49	2/8/2008 10:42	2/8/2008 10:43	2/8/2008 10:48	2/8/2008 10:54	2/8/2008 10:44
2/8/2008 10:50	2/8/2008 10:46	2/8/2008 10:44	2/8/2008 10:53	2/8/2008 10:56	2/8/2008 10:44
2/8/2008 10:51	2/8/2008 10:49	2/8/2008 10:45	2/8/2008 10:55	2/8/2008 10:58	2/8/2008 10:50
	2/8/2008 10:51	2/8/2008 10:49	2/8/2008 10:56	2/8/2008 10:59	2/8/2008 10:55
	2/8/2008 10:53	2/8/2008 10:58	2/8/2008 10:58	2/8/2008 11:06	2/8/2008 11:00
					2/8/2008 11:01
					2/8/2008 11:07

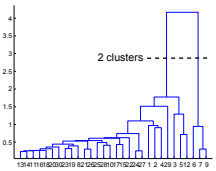
(b) The report produced by *speed* view (GPS data)

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:45	2/8/2008 10:39	2/8/2008 10:37	2/8/2008 10:29	2/8/2008 10:38	2/8/2008 10:50
2/8/2008 10:47	2/8/2008 10:40	2/8/2008 10:40	2/8/2008 10:30	2/8/2008 10:43	2/8/2008 10:55
2/8/2008 10:50	2/8/2008 10:41	2/8/2008 10:41	2/8/2008 10:33	2/8/2008 10:45	
	2/8/2008 10:42	2/8/2008 10:43	2/8/2008 10:35	2/8/2008 10:47	
	2/8/2008 10:44	2/8/2008 10:44	2/8/2008 10:40	2/8/2008 10:51	
	2/8/2008 10:46	2/8/2008 10:45	2/8/2008 10:43		
	2/8/2008 10:48	2/8/2008 10:48	2/8/2008 10:46		
	2/8/2008 10:50	2/8/2008 10:51	2/8/2008 10:50		
	2/8/2008 10:55	2/8/2008 10:58	2/8/2008 10:55		

(c) The report produced by *duration* view (GPS data)

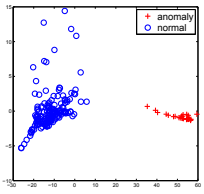


(a) *Flow* view (PeMS)

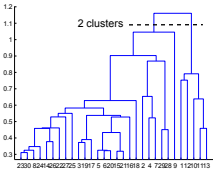


(b) Dendrogram of *flow* view

Figure 4: *Flow* view results from PeMS Station 400165 where the accident happened (a) the Isomap data plot of *flow* view (b) dendrogram of hierarchical clustering of *flow* view.



(a) *Speed* view (GPS)



(b) Dendrogram *speed* view

Figure 5: *Speed* view results from GPS phones (a) the Isomap data plots of *speed* view (b) dendrogram of hierarchical clustering of *speed* view.

used in our experiments. The anomaly detection is based on the hierarchical clustering result shown on the right-hand side of Figures 4-6. To speak of the detection on spatial domain, Figure 7 shows that the *flow* view detection result from different *PeMS* stations which were affected by accident. We adopt the 90-10 rule which was discussed in Sub-

Table 4: The reports produced by all combinations of view association.

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:40	2/8/2008 10:40	2/8/2008 10:35
2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:45	2/8/2008 10:45	2/8/2008 10:40
	2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:50	2/8/2008 10:50	2/8/2008 10:45
			2/8/2008 10:55	2/8/2008 10:55	2/8/2008 10:50

(a) The report produced by associating the *flow* and *speed* view

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:45	2/8/2008 10:30	2/8/2008 10:35	2/8/2008 10:50
	2/8/2008 10:50	2/8/2008 10:50	2/8/2008 10:35	2/8/2008 10:40	
	2/8/2008 10:55	2/8/2008 10:55	2/8/2008 10:40	2/8/2008 10:45	
			2/8/2008 10:45		
			2/8/2008 10:50		
			2/8/2008 10:55		

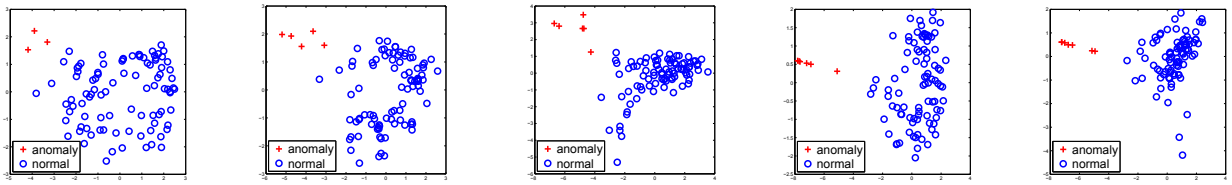
(b) Final report produced by associating the *flow* and *duration* view

400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:37	2/8/2008 10:43	2/8/2008 10:43	2/8/2008 10:50
2/8/2008 10:47	2/8/2008 10:41	2/8/2008 10:41	2/8/2008 10:46	2/8/2008 10:51	2/8/2008 10:55
2/8/2008 10:50	2/8/2008 10:42	2/8/2008 10:43	2/8/2008 10:55		
	2/8/2008 10:46	2/8/2008 10:45			
	2/8/2008 10:50	2/8/2008 10:58			

(c) Final report produced by associating *speed* and *duration* view

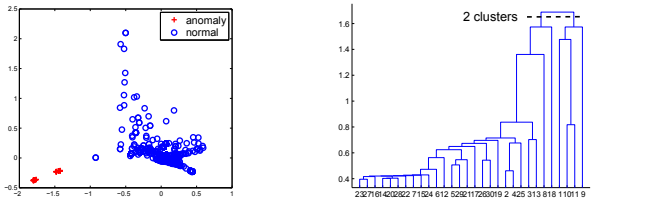
400488 (24.007)	401561(24.477)	400611(24.917)	400284(25.767)	400041(26.027)	400165(26.641)
2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:40	2/8/2008 10:45	2/8/2008 10:45	2/8/2008 10:50
	2/8/2008 10:50	2/8/2008 10:45	2/8/2008 10:50		
	2/8/2008 10:55	2/8/2008 10:50	2/8/2008 10:55		
		2/8/2008 10:55			

(d) Final report produced by associating all views: *flow*, *speed* and *duration* view



(a) Station 400488 (b) Station 401561 (c) Station 400611 (d) Station 400284 (e) Station 400041

Figure 7: The Isomap data plot for each of interested PeMS stations. The blue circle indicates normal pattern and the red cross indicates the anomalous pattern. For each station, we can observe that the anomalous points are well separated from the normal points.



(a) *Duration* view (GPS) (b) Dendrogram of *duration* view

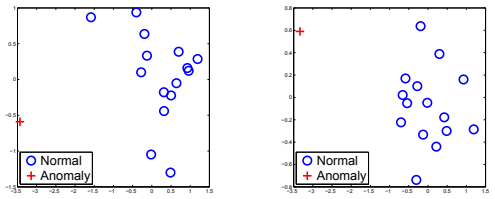
Figure 6: *Duration* view results from GPS phones (a) the Isomap data plots of *duration* view (b) dendrogram of hierarchical clustering of *duration* view.

subsection 3.2.3 to confirm the anomalies in all our detection procedures.

4.2.2 Evaluation via the Caltrans PeMS Data

In the second series of experiments, we evaluate the proposed anomaly detection scheme using the *Caltrans PeMS* data. As a preliminary study, we only focus on the data that are collected near the postmile 26.641. The proposed method detects an anomalous event at 1:00 PM, which coincides with the ground truth. Figure 8 shows the Isomap data plots for *flow* and *speed* views. Both views show that there is an unusual pattern on the left which is the accident happened at 1:00 PM of December 14, 2007. The view association confirms the finding from these two views.

Given the *Caltrans PeMS* data, a multiple-day dataset, we can test how the proposed method is used for online anomaly detection: detecting anomalous events on a later day given the training of previous days. We would like to answer the following questions: i) Given a location and a specific moment, can we detect anomalies in that moment using previous days' data for training? ii) Following the previous question, will the result be influenced if some weekend days are added in the training?



(a) *Speed* view (b) *Flow* view

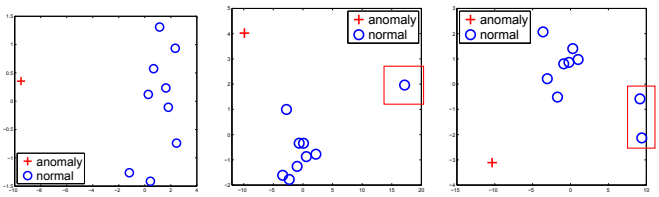
Figure 8: The anomaly detection results for Dec. 14, 2007. Both views can detect the anomaly correctly which is the accident happened at 1:00 PM. The anomalous pattern is clearly separated from other normal patterns.

Table 5: Anomaly detection results using the *speed* data on previous days.

Weekdays	One-day weekend	Two-day weekend
12/14/2007 13:00	12/9/2007 13:00	12/8/2007 13:00 12/9/2007 13:00

In order to answer the above questions, we design the experiments as follows: choosing the previous days' data for training, but (i) including only the data at the same time with the moment that we want to detect; ii) including only the data contains no accident; iii) using the previous 10 days' data; iv) detecting by hourly basis instead of daily basis. These previous days' data is used to judge if the data from this moment is considered as an anomaly or not. We use only the *speed* information in this part of study.

Table 5 shows anomaly detection using previous days' data. If we include only the weekdays for training, then we can correctly capture the anomalous event; on the other hand, if we include the weekdays' as well as the weekend's data for training, the detection result is no longer correct. The weekend's data are detected as anomalies in this case. We have more than one kind of "anomalies" and the anomaly detection procedure may detect either one of the anomalies. The data plot results are in Figure 9. Figure 9 (a) shows the detection result given only weekdays for training. Apparently, the anomaly is correctly detected in this case. Figure 9(b) and (c) show how the result is influenced by the weekend's data. The weekend's data (red square) are detected as anomalies in this case, while the real anomaly (red cross) is the one on the left.



(a) weekdays (b) one-day weekend (c) two-day weekend

Figure 9: The Isomap plot results: (a) using only weekdays' data for training, (b) using weekdays' and Sunday's data for training, and (c) using weekdays' and the whole weekend's data for training.

4.2.3 Computation Time

All experiments were performed on regular Pentium-5 machine. All computation regarding to one view was done less than five seconds on average.

4.2.4 Discussion

We would like to discuss several issues related to the proposed method and the focused problem. Some possible future extensions to this work will also be mentioned.

- In this work, to obtain the final anomaly detection result we simply intersect two or more views for view association. We can however to have a more robust view association procedure. First, we need to have criteria to judge which view is trustworthy, such as in this case, we can assume the *flow* and the *speed* views more trustworthy than the *duration* view because the *flow* and the *speed* views are provided by the sensors directly and the *duration* view is a derived feature based on several sensor readings. Second, we can consider different types of view association. For instance, we can consider adjust the portion of anomalies in each view so that the intersection is maximized. It is one of our future research topics.
- We studied both of the *duration* view and the *speed* view. It seems that they may share some correlation between each other, hence not appropriate to be used simultaneously to detect anomalies. The *duration* feature is defined as dividing the distance of consecutive stations by the *speed*. In this case study, the distance of consecutive stations may not remain constant (ranged from a half to one mile); therefore, the *duration* and *speed* may not refer to the same concept.
- We use Isomap mainly for data representation and visualization. To perform data clustering, we can either use hierarchical clustering method or perform regular data clustering such as K-means in the dimensionality-reduced (intrinsic) space. The difference between them is that the hierarchical clustering gives relatively stable result, compared to, e.g., the one from K-means if applied to the intrinsic space.
- In our study, we apply the 90-10 principle to confirm whether or not a group of data are considered as anomalies. In this work, all the results satisfy this principle. However, we should consider some adjustment once the principle is not fitted to the clustering result. We can try the following options:
  - Still stick with the 90-10 principle, but we move the points from the major group or the minor group or vice versa for the points closest to the group boundaries until 90-10 rule is satisfied.
  - Let the hierarchical clustering result or the clustering result in the intrinsic space (based on Isomap) decides the result itself.
  - Apply another flexible principle such as cutting data into two portions p and $1-p$ where p is between 0 and 1.

5. CONCLUSION

We proposed a method to detect anomalies in ITS data for traffic analysis. Different from previous anomaly detection approaches, we mainly focused on automatically generating an anomaly report, in this case, the incident report that shall

be helpful for drivers and police to act accordingly for the incident. In this report, the incident location, the moment of incident, and more importantly how the incident affects the traffic, for how long are all included based on our detection result. Therefore the police patrol can decide the significance of the incident and make appropriate judgment about whether or not they should go to the incident location for investigation and recovery and which incident they should take care first if more than one incident occurs at similar moments. Regarding to the methodology, we detect anomalies based on a view association approach given the multi-view information where each single view is used to detect anomalies, and the results from different views are combined for the final detection result. The method has many benefits if compared to previous anomaly detectors: 1) it needs little parameter tuning; 2) it needs no clean data training as the initial step; 3) it can work efficiently such as the algorithm can easily implemented in a parallel fashion. We evaluated the proposed method on *Mobile Century* and *PeMS* datasets. The evaluation shows the proposed method is effective at detecting incidents from the data. Even though we focused only on anomaly detection in ITS data in this work, it would not be surprising if the proposed method is easily generalized to other types of applications where anomaly detection is necessary. Investigating the above possibility is one of our future research plans.

6. REFERENCES

[1] A. Agovic, A. Banerjee, A. Ganguly, and V. Protopopescu. Anomaly detection using manifold embedding and its applications in transportation corridors. *Intell. Data Anal.*, 13(3):435–455, Aug. 2009.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 92–100, 1998.

[3] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *Intelligent Transportation Systems, IEEE Transactions on*, 5(1):42–53, 2004.

[4] S. Chawla, Y. Zheng, and J. Hu. Inferring the root cause in road traffic anomalies. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 141–150. IEEE, 2012.

[5] C. Chen, D. Zhang, P. S. Castro, N. Li, L. Sun, and S. Li. Real-time detection of anomalous taxi trajectories from GPS traces. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pages 63–74. Springer, 2012.

[6] T. F. Cox and M. A. A. Cox. *Multidimensional Scalling, Second Edition*. Chapman & Hall/CRC, 2000.

[7] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for unix processes. In *Security and Privacy, 1996. Proceedings., 1996 IEEE Symposium on*, pages 120–128. IEEE, 1996.

[8] Z. Fu, W. Hu, and T. Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–602. IEEE, 2005.

[9] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.

[10] F. Jiang, Y. Wu, and A. K. Katsaggelos. A dynamic hierarchical clustering method for trajectory-based unusual video event detection. *Image Processing, IEEE Transactions on*, 18(4):907–913, 2009.

[11] I. T. Jolliffe. *Principal component analysis*, volume 487. Springer-Verlag New York, 1986.

[12] A. Kind, M. P. Stoecklin, and X. Dimitropoulos. Histogram-based traffic anomaly detection. *Network and Service Management, IEEE Transactions on*, 6(2):110–121, 2009.

[13] P. G. Michalopoulos, R. D. Jacobson, C. A. Anderson, and T. B. DeBruycker. Automatic incident detection through video image processing. *Traffic engineering and control*, 34(2), 1993.

[14] P. Mirchandani and L. Head. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies*, 9(6):415–432, 2001.

[15] C. Piciarelli and G. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15):1835–1842, 2006.

[16] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[17] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of pca for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review*, 35(1):109–120, 2007.

[18] P. H. A. Sneath. *Numerical taxonomy : the principles and practice of numerical classification*. W.H. Freeman, San Francisco, 1973.

[19] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[20] S. Thajchayapong and J. A. Barria. Anomaly detection using microscopic traffic variables on freeway segments. *CD-ROM, Transportation Research Board of the National Academies*, 2010.

[21] D. Valerio, T. Witek, F. Ricciato, R. Pilz, and W. Wiedermann. Road traffic estimation from cellular network monitoring: a hands-on investigation. In *Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pages 3035–3039. IEEE, 2009.

[22] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen. Data-driven intelligent transportation systems: A survey. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4):1624–1639, 2011.

[23] Z. Zhang, K. Huang, and T. Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1135–1138. IEEE, 2006.