A FCD–based Method for Traffic Congestion Detection in low-proximity data environment

Ilya S. Ardakani*, Thanaruk Theeramunkong*, and Waree Kongprawechnon*

Sirindhorn International Institute of Technology, Thammasat University, Thailand

Abstract— **Recently GPS data gathered from vehicles on roads, so-called floating car data (FCD), have become one of the popular sources for traffic management. As this source of data becomes more accessible, there are increasing research works conducted in order to extract reliable information from available raw sensory data. Such data from public transportation equipped with GPS sensors provides relatively consistent source of FCD. However, in some cases, due to low-frequency travel routines and the small number of active vehicles, extracted FCD are not sufficient for traffic prediction. This study proposes a method to efficiently identify road congestion using FCD collected from in-campus shuttle buses in low speed routes. In this task, the challenging issue is how to distinguish real congestion from intentional delay, such as passenger loading and temporary parking, since both are similar in terms of low average speed on the road link. The proposed method employs artificial neural network to predict traffic on the links with sparse FCD available. The experimental result demonstrates that our method is capable of producing comparable results with existing methods.**

Index Terms— **Intelligent Transportation Systems; Artificial Neural Networks; Pattern Recognition Traffic Congestion Detection; Expectation Maximization; Gaussian Mixture Models;**

I. Introduction

Being aware of current traffic flow condition benefits every driver on path planning before starting a journey. Overflowing one route often induces congestion which may propagate further back upstream, triggering wide-area traffic congestion. With update information on congestion situation on a certain path, it is possible to make more efficient path planning to avoid trapping into snarled-up traffic, and then reduce local-area and wide-area congestion. While reliability and cost of collecting information become more focused, vehicle tracking technology, particularly Global Positioning System (GPS) for collecting current traffic situation and cloud computing for storing historical traffic flow, has been recently developed to the level of practical use. Predicting traffic flow can be achieved by using such current and historical traffic information. At present, so as an available source of information on traffic situation, mining vehicle's GPS data efficiently and reliably considered to be a well-recognized field of research. In the past, existing works were mainly categorized as follows.

- Traffic and GPS data collection methods
- Public transport arrival time prediction methods
- Traffic congestion detection and prediction

* All authors are with Sirindhorn International Institute of Technology, Thammasat University, 131 Moo 5, Tiwanont Road, Bangkadi, Muang Pathum Thani 12000, Thailand. Email: thanaruk@siit.tu.ac.th, ilya_ardakani@outlook.com, waree@siit.tu.ac.th.

As GPS data become more and more abundant and available, analyzing and mining such large-scale data for extracting useful information became more focus of numerous researchers. At the same time, the computational power grows and fast probabilistic techniques are being introduced in several literatures. Daniel Ashbrook [1] introduced a system to learn significant locations and predict the future movements based on Hidden Markov Models. A similar approach was also applied to transportation. As day by day more vehicles are being equipped with GPS modules, a treasure of data, which could be exploited for myriad of applications, become available. For the case of Intelligent Transportation Systems, GPS data is utilized as Floating Car Data (FCD) which in volume is preferable to conventional Traffic Flow Sensors due to less cost and effort for installation and maintenance as pointed by Leduc in [2].

FCD come with sometimes scarcity or sheer size of data which are challenges for studying, analyzing, and generating of reliable conclusion from collected data. Typically the size of sample data plays an important role on viability of the inferred information from FCD as discussed in [3, 4]. To improve accuracy and efficiency of computation, Li introduced Map-Reduce technique in [5] applied on public buses GPS data. For traffic pattern analysis and prediction K-mean clustering used in [6, 7]. Wu offered improved K-mean clustering in [8] which helped with assigning the correct site or location label to scattered collected GPS position data. Modsching in [9] used map based routes to correct discrepancy of the reported GPS locations. There is also lack of link coverage problem, as buses do not travel through all routes. Abadi in [10] came up with a solution for estimation of traffic flow with limited link data.

FCD collected from public transportation were the spot light of Lipan in [11] which makes it similar case with this study due to likes of shuttle buses and public transport. To use public transport GPS data as FCD, the properties of behavior vehicles' trajectories and responses must be analyzed closely. Since shuttle buses like public buses need to stop and dwell at bus stops the trajectory time depends on both rate of passenger arrival and road congestion. As one application, some researchers conducted research on bus arrival prediction and estimation. Zhu's work clearly demonstrated the dependence on traffic information in [12]. Cheng proposed a model based on passenger count data in [13]. As machine learning approach, Jeong introduced the employment of Artificial Neural Networks with accuracy of around a minute in [14] and an early work in 2006 by Bin [15] used Support Vector Machines.

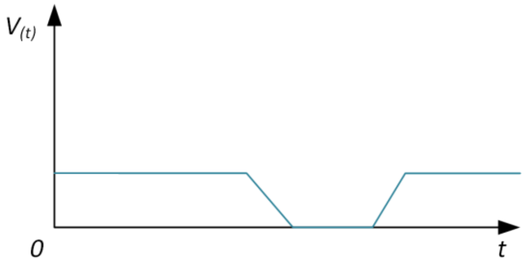

Fig. 1. Plot of velocity function for congested link

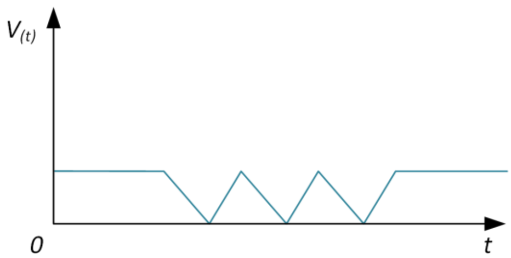

Fig. 2. Plot of velocity function for non-congested link

Finally, on congestion estimation, De Fabritiis and Kerner used FCD to perform online estimation of traffic flow in [16, 17]. They utilized spatial temporal patterns retrieved from FCD to produce speed prediction. Ho [18] used speed pattern recognition to provide a more detailed model for estimation in which average speed calculated between two time stamps were used. As a different approach, Poomrittigul worked on mean speed estimation based on sample speed mean acquired from vehicles on the road segment in [19]. Speed samples reported by GPS device are useful in speed estimation; however, they are infected with more noise at lower speeds depending on sensor's quality.

All methods mentioned above require significant amount of FCD to be collected. Besides, the average speed ranges for different classes of congestion are distant. Therefore, it is suitable to distinguish the congestion patterns based on average velocities. However, in this study we focus on FCD, collected from a special situation, i.e., shuttle buses traveling along campus roads. In this situation, the data are sparse and may not always be available. In addition, commuting velocities are relatively low. As described in the third section, since low speeds and long stops for picking up passengers produce similar average velocity on the links, it is hard to distinguish between congestion and intentional delay. Hence, the goal of this work is to introduce a technique to extract reliable traffic label information from collected FCD for training classifiers which provide short term and long term forecast on traffic congestion.

Organization of this paper is as follows. In Section 2, the problem statement is described in details. Data collection and conditioning methods are discussed in Section 3. Section 4 explains traffic congestion detection algorithms and techniques for analyzing data. Section 5 presents the experimental results and discussions. Finally, conclusion and future works are mentions in Section 6.

II. Problem Statement

This study copes with two main issues. One is how to identify traffic congestion on road segments while GPS data from shuttle buses exists for those segments. The other is how to predict traffic congestion for road segments when GPS information is not available for those segments.

This subsection describes a method to identify congestion on road using the recently gathered GPS data from shuttle buses. The challenge is that nature of vehicle speeds on campus roads are inherently low and dependent on the route and location. Hence simply assigning speed ranges for congestion detection is inaccurate. Average velocity $\boldsymbol{v_{avg}}$ is computed using Eq. (1) where $\boldsymbol{d}$ is total distance travelled and $\boldsymbol{t}$ is travelling time.

$$v_{avg} = \frac{d}{t} \tag{1}$$

Based on velocity function $v_{(t)}$, d can be calculated by Eq. (2), i.e. the area under the $v_{(t)}$ curve.

$$d = \int v_{(t)}\, \delta t \tag{2}$$

In the situation of low velocities along segments of road, different velocity functions of shuttle vehicle expedite the same distances in similar time periods. This results in the same or close average velocities on the road links for different motion patterns belonging to congested and non-congested states as shown in Fig. 1 and Fig. 2, respectively.

Correctly identifying congested links is vital in providing labels for training neural networks for predicting traffic congestions. This study offers a new approach to identify traffic congestions based on spatial-temporal densities of micro speed FCD samples (samples with velocities nearly equal zero).

a. Traffic Congestion Forecast

Using FCD collected from public transport raises another issue in traffic congestion prediction. Shuttle buses carrying FCD collection sensors commute the routes routinely. However, there is not consistently a sensor vehicle available on the concerning road segment. To fill in for mentioned missing information, this work exploits the long term and short term pattern of traffic on road segments to provide the missing congestion states. These patterns are produced based on input features such as day of the week, time of the day, weather index, road segment, and activity of the road. As most of aforementioned features are readily self-explanatory, activity of the road requires for elaboration. Generally, a calendar or casual event on campus also affects the activity on the roads. For instance, in this experiment campus calendar which determines semester breaks, and other irregular individual events should be included in the input features for traffic congestion prediction. But such information is not always readily available. This work proposes a method to use historical average and standard deviation of sample velocities to identify which activity occurs on the roads. Fig. 3 shows travel times in average during a normal active campus period which are higher in contrast to an inactive campus period where travel time averages are much lower as illustrated in Fig. 4.

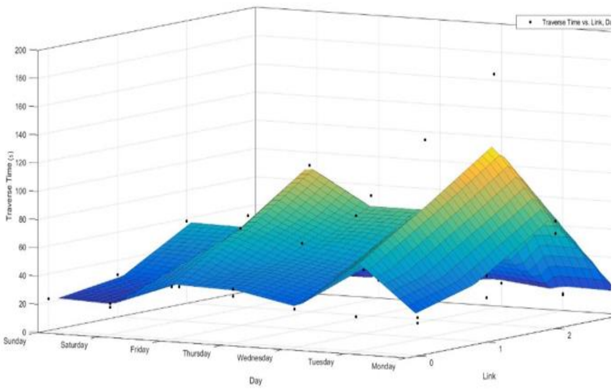

Fig. 3. Average travel time during active campus period

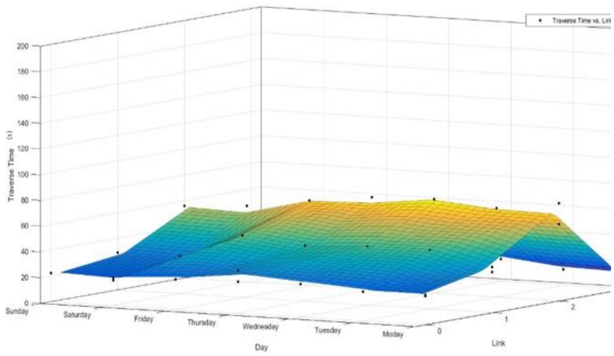

Fig. 4. Average travel time during inactive period (e.g. semester break)

III. Data Preparation

FCD for this experiment come from 30 GPS/GPRS combo modules installed on the campus's shuttle buses. They report their positions every second to a remote server via the GPRS module. Each record represents vehicle ID, latitude, longitude, time, speed and signal strength. The data collected are bus locations in two semesters during November 2014 and February 2015. In total, 160,000 data points are for each semester. As the campus's shuttle routes are reasonably far from any tall building or obstacles, the signal strength remains mostly acceptable and reliable. We also use historical weather data acquired from a meteorological database [20]. Data with rainy weather were assigned with 1 and dry with 0. Feature from time stamp includes day-in-week and time-slot features. For day-in-week feature, a number between 0 and 6 is assigned to Monday to Sunday, in order. For time-slot, we use 20-min slots from 8:00 till 20:00, a usual operation period, with numerical assignments of 0 to 35. The data with time stamp out of this period were ignored for this experiment. Direction of each sample point is of two types: 1 for downstream and -1 for upstream. With this information, a trajectory can be traversed as downstream and upstream.

To set up the path checkpoint matrix, coordinates are chosen either manually or automatically. As a result, path matrix is $N\times 2$ in size where N is the number of check points included in the path. Every consecutive checkpoint pair belongs to a path link with a total of $N-1$ links. In Fig. 5, sample path checkpoints are presented along with GPS samples.

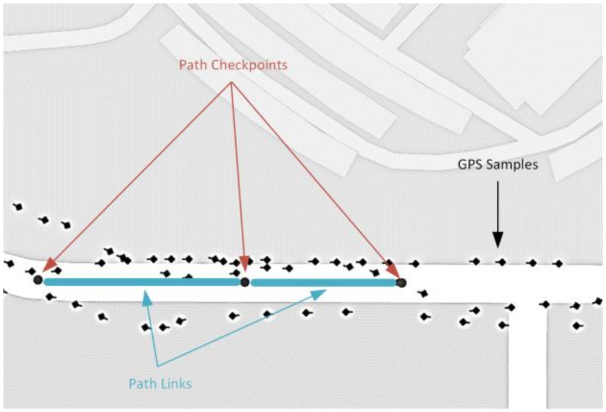

Fig. 5. Path checkpoints and links presented along with GPS samples

For congestion identification experiment each of collected micro-speed GPS samples is classified based on the nearest path's check point. It is to be noted that for finding the nearest point of each path checkpoint to the trajectory coordinates KD-Tree algorithm was used [21]. There are N location based bins. Data rows are of <*PathCheckpointIndex*, *TimeSlotIndex*, *WeekDayIndex*, *DistancetoCheckpoint*, *SampledSpeed*, *Direction*, *BusId*> where '*PathCheckpointIndex*' is the index of the nearest path checkpoint to the sampled GPS point. '*DistancetoCheckpoint'* is the distance between the checkpoint and the reported GPS location. This is calculated using Haversine function, $H(p_1, p_2)$, given in Eq. (3) and (4), where r_e is average earth's radius, x and y are longitude and latitude of the location respectively in radians. The unit will be based on unit of earth's radius measurement. '*SampledSpeed*' is NMEA reported near zero speed which is originally in knots and converted to SI unit of Km/h.

$$H(p_2, p_1) = 2r_e \sqrt{\sin^2\left(\frac{\Delta y}{2}\right) + \cos(y_{p_2})\cos(y_{p_1})\sin^{-1}\left(\frac{\Delta x}{2}\right)} \quad (3)$$

$$\Delta y = y_{p_2} - y_{p_1}, \qquad \Delta x = x_{p_2} - x_{p_1} \tag{4}$$

For computing historical average and standard deviation of the road segments velocities, average speed of sensor carrying vehicle on road links obtained using Eq. (5), where t_{tail} and t_{head} are time stamps acquired upon sensor vehicle passing respecting checkpoints.

$$v_{avg} = \frac{H(P_{head}, P_{tail})}{t_{tail} - t_{head}} \tag{5}$$

It is noted that for finding the nearest point of each path checkpoint to the trajectory coordinates KD-Tree algorithm was used. Data vectors consist of <*PathLinkIndex*, *TimeSlotIndex*, *WeekDayIndex*, *WeatherIndex*, *AverageSpeed*, *Direction*, *BusId*> features where '*PathLinkIndex*' is the index of the traversed path link.

IV. Methodology

Consisting of two parts, the first part describes our method to produce congestion labels while the second part presents our algorithm that performs congestion forecast. These two parts are combined and tested in our experiments shown in Section 5.

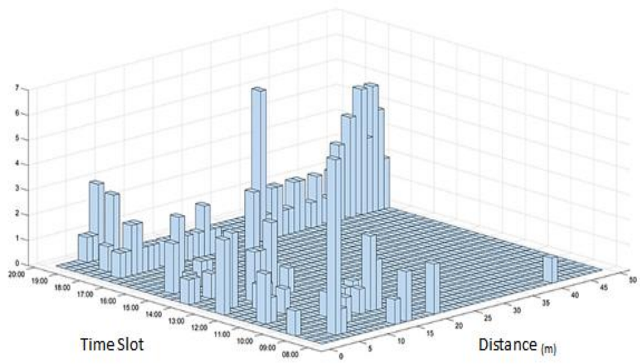

Fig. 6. Histogram of micro-speed incidents of a sensor vehicle for a day

a. *Traffic Congestion Identification*

This study estimates occurrences of micro-speed (near zero speed) incidents belonging to traffic congestion rather than natural incidents (bus-stop, speed bump, and so on) using log likelihood results. To compute the density parameters of both incidents, Expectation-Maximization (EM) algorithm on multivariate Gaussian mixture models (GMM) [22] is used. The multivariate Gaussian mixture basically models three-dimensional spatial-temporal micro-speed incidents. To reduce the complexity even further, spatial dimensions are reduced to scalar distance for nearest checkpoint. For time dimension, we utilize 20-min slots. The histogram of micro-speed incidents for a day is shown in Fig. 6.

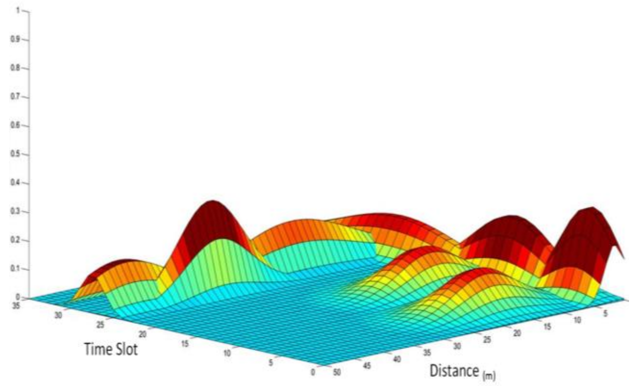

Fig. 7. Gaussian Mixture Model of micro-speed incidents at checkpoint 1 downstream direction

Each region with high micro-speed density can be assigned to a multivariate Gaussian with $\mu_{t,d}$ and $\Sigma_{t,d}$. As stretch of density over time is more than distance for bus-stops or natural stop occurring more often at around the similar locations, multiple densities can be combined forming a single bigger density with a similar distance mean since the distribution on time axis is not a concern. The covariance of these distributions would be a diagonal matrix with higher greater than one ratio of time Eigen value over relatively very small distance Eigen value. The approximate Gaussian mixture model is shown in Fig. 7.

Traffic congestion densities are identified by significantly larger distance Eigen value in their covariance matrices Σ_c. With distance mean μ_d and distance covariance σ_d^2 computed for traffic congestion C and natural stops S, negative log likelihood values are obtained by Eq. (6) and (7) for each micro speed incident z_i in the corresponding time slot. The minimum value determines the nature of samples in each component. Negative log likelihood function is shown in (6) and (7).

$$\arg\min_{S,C} -\log\, p(z_i | \mu_d\,, \sigma_d^2) \tag{6}$$

$$p(z_i | \mu_c, \sigma_d^2) = \mathcal{N}(\mu_d, \sigma_d^2) \tag{7}$$

b. *Traffic Congestion Prediction*

Given the congestion densities identified by the method in the previous session, the severity of the congestion or its size can be identified using a classification method described in this section. In our method, a three-class grading <*Heavy*: 1, *Medium*: 2, *Free*: 3> is applied to express traffic congestion. This grading is done based on variance and frequency of micro speed samples. This approach greatly depends on the geometry of the road links and dynamics of sensor carrying vehicles. For instance, in this work, densities graded in accord with criterion in (8) where L is road link's length in meters, N_z is number micro speed incidents and γ is free parameter.

$$N_z \times \frac{L}{\gamma \sigma_d^2}, \;\; \gamma > 0 \tag{8}$$

The input features are <*PathLinkIndex*, *TimeSlotIndex*, *WeekDayIndex*, *WeatherIndex*, *HistoricalLinkSpeedMean*, *HistoricalLinkSpeedStd*>. Historical means and standard deviations could be weighted towards most recent samples. In this experiment, a week window of previous samples is used for averaging. Historical input features are intended to provide information about campus activity state to the predictor. These inputs could be configured to address either short-term or long-term dependencies based on the application.

For prediction, an artificial neural network with hidden layers and non-linear activation functions is trained using stochastic gradient descent. The network is trained and tested on data collected for a single semester. Size of held-out test set is 15% of whole dataset. Moreover, an extended held-out test is performed using input data from different semester. This test was done to observe effects of campus calendar changes on prediction results.

V. Experiment Results and Discussion

In order to train the classifier, the training target labels are identified in advance. Then neural network is trained to achieve the best results on validation set. Finally, held-out test and extended tests are performed to evaluate the performance of the classifier.

a. Congestion Components Identification

For verifying the practical performance of the method introduced, one-month micro-speed incidents of around 49000 collected samples of speed data near checkpoints clustered using EM algorithm with range of five components and lowest BIC Score used to verify the best number of components. The result of checkpoints, 1 and 2, is discussed here. The mean and diagonal components of the covariance matrices for checkpoint 2 are shown in Table I.

TABLE I
Mean and covariance of GMM components for checkpoint 2

Component Number	Time Mean	Distance Mean	Time Eigen Value	Distance Eigen Value
1	29.3878	28.8851	10.6799	66.1546
2	29.1126	10.4923	7.8871	31.4381
3	11.4075	29.5156	43.3763	40.5304
4	13.0453	13.7467	40.6414	54.3935
5	19.2030	45.5161	65.6773	5.4409

As mentioned before, components with high ratio of time Eigen value to distance Eigen value, demonstrate the fixed location stops like bus-stops. Component 5 is such location distance at 45 meters from check point. The same criteria apply to peak time traffic congestions. Components 1 and 2 are of that feature. We observe very close time mean and high ratio of distance Eigen value over time Eigen value. It clearly shows the rush hour at time slot 29 which translates to around 18:00. It is noted that components 1 and 2 could be combined to represent a single traffic congestion component.

The results for components 3 and 4 are less definitive, however the component 4 is great representation of right turn cross which also gets lightly congested in morning and noon time. The component 3 also coincides with an unsigned bus stop location at which sometimes bus stops. After all, it is simply a more preferable choice to combine the components with the close means on either of axis for online detection. As traffic congestions could be appearing anytime anywhere, while there has to be less likely chance of labeling long stops at the same location as components with high time Eigen value. There are the same results for checkpoint 1 as shown in Fig. 8 where the component 4 has a relatively small distance Eigen value, i.e. 6, and time Eigen value of 60 identified as bus-stop with average distance of around five (5). Two heavy congestion clusters 1, and 2 are clearly identified at time around 29 or 18:00 which agrees with checkpoint 2 pointed above. Spherical shape of two is caused by overlapping the bus-stop location.

b. Traffic Congestion Prediction

There are 4351 vectors of 6 dimensions available for this experiment. 653 of the samples are held-out for testing. Training is performed on 2828 samples and model trained model is validated at each epoch on 870 data points. Trained network was able to achieve best accuracy of 92.6% and lowest of 87% in held-out test. In comparison, De Fabritiis in [16] was able to obtain results in range of 2-8% for 15-min prediction and 3-16% for 30-min forecast. Projection windows in this work are varying from 20-mins to 40-mins depending on availability of the shuttle bus on corresponding links. Therefore, the obtained results are relatively acceptable. Fig. 9 illustrated the prediction results averaged over days. In comparison with average links speeds shown in Fig. 10, prediction contours show a great similarity with links speeds contours in time axis during rush hours. However, there is not much Class-1 congestion state predicted for link 3 in contrast with its very low speeds. This demonstrated the classifier's success in distinguishing the low speed from congestion.

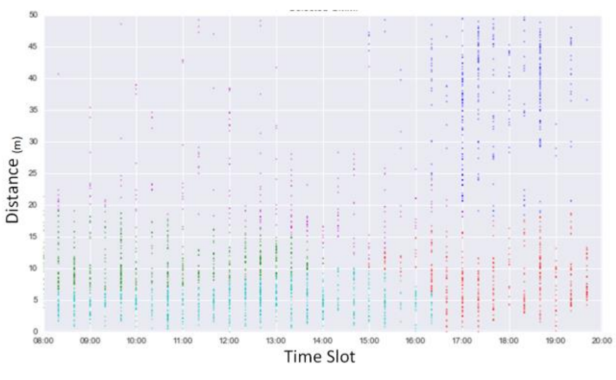

Fig. 8. GMM produced by EM algorithm with 5 components at checkpoint 1

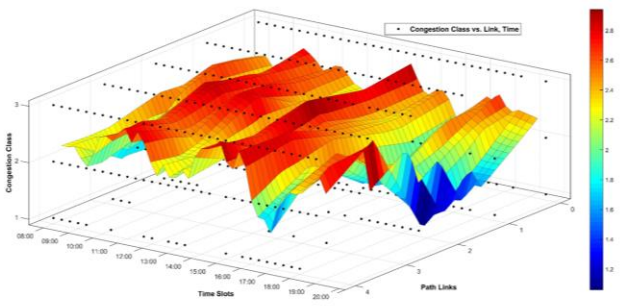

Fig. 9. Fitted surface plot of predicted congestion classes against time and link averaged over days

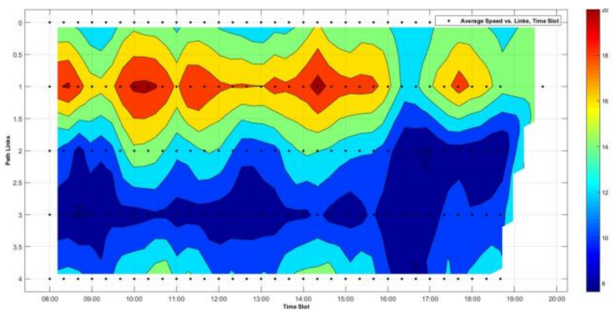

Fig. 10. Contour plot of links' speed against time, averaged over days

In addition, in the extended test, the classifier achieved an accuracy of 77.3% predicting the congestion state of road links in the next semester. For this test dataset, we observed that 3,002 data points and 680 data points were misclassified. The classifier failed to detect any Class-1 congestion along high error in Class-2 predictions. Interestingly, all Class-1 targets predicted to be Class-2 and errors in Class-2 predicted to be Class-3. This outcome may provide the explanation on the source of failure. Average speeds in the training semester where lower than extended test's data from second semester. This verifies the effect of historical average speed input in the classification. This is a drawback for this method as classifier should be retrained in events of pronounced changes in calendar activities. Although, comparing obtained results with case of classifier with no historical input provided, there is about 20-30% advantage. Experiment showed that the accuracy of classifier significantly reduces to approximately 50-60% by removing the historical links speeds features and labeling classes using speed ranges.

VI. Conclusion

In summary, this study offered a new approach towards identification of traffic congestion using Expectation Maximization (EM) algorithm and Gaussian Mixture Models (GMM). Given appropriate sampling periods, the proposed method is able to efficiently and reliably identify the congestions from regular delays in motion of special purpose sensor carrying vehicles. Moreover, this method could be extended for identifying bus-stops with more frequent passenger and their peak usage time window. However, there are also downsides to this method such as sampling rate margins and uneven sampling rates. This method needs at least a minimum limit for period of sampling for reliable detection. In addition, road link lengths are also should be considered considerably larger than components geometry in the mixture model. Otherwise, it is difficult to identify the congestion components.

Regarding the congestion forecast, this work exploited the congestion patterns and their dependencies on the recent speed profile. This profile expressed to the classifier in terms of average and standard deviation of short-term or long-term speed data. Focusing on the results, proposed method achieved results comparable to parallel existing methods using a smaller dataset. Yet, there exist few disadvantages to this method. Significant changes in calendar, greatly affects the prediction results. As it observed in extended testing, change in semester resulted in 15% drop in accuracy. So, the classifier needed to be retrained. An alternative method is providing calendar information to the classifier. However, a more complex solution is employing long-short term memory network for prediction.

Acknowledgement

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech Program. It is also partly supported by National Research University Project of Commission of Higher Education (CHE), Thailand as well as Center of Excellence (CoE) Initiatives funded by Thammasat University. Besides, special thanks go to Mr. Kobkrit Viriyayudhakorn, Mr. Phalakron Nilkhet, TAIST Tokyo Tech and National Research University Project for providing the Thammasat University's Shuttle Buses' GPS data.

References

1. Ashbrook, Daniel, and Thad Starner. "Learning significant locations and predicting user movement with GPS." In Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on, pp. 101-108. IEEE, 2002.
2. Leduc, Guillaume. "Road traffic data: Collection methods and applications." Working Papers on Energy, Transport and Climate Change 1 (2008): 55.
3. Long Cheu, Ruey, Chi Xie, and Der-Horng Lee. "Probe vehicle population and sample size for arterial speed estimation." Computer-Aided Civil and Infrastructure Engineering 17, no. 1 (2002): 53-60.
4. Hong, Jun, Xuedan Zhang, Zhongya Wei, Li Li, and Yong Ren. "Spatial and temporal analysis of probe vehicle-based sampling for real-time traffic information system." In Intelligent Vehicles Symposium, 2007 IEEE, pp. 1234-1239. IEEE, 2007.
5. Li, Dapeng, Yu Haitao, Xiaohua Zhou, and Mengdan Gao: Map-reduce for calibrating massive bus trajectory data. In: ITS Telecommunications (ITST), 2013 13th International Conference on, pp. 44-49. IEEE, 2013.
6. Weijermars, Wilhelmina Adriana Maria. Analysis of urban traffic patterns using clustering. University of Twente, 2007.

7. Ashbrook, Daniel, and Thad Starner. "Learning significant locations and predicting user movement with GPS." In Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on, pp. 101-108. IEEE, 2002.
8. Wu, Xueying, and Chunlong Yao. "Application of improved K-means clustering algorithm in transit data collection." In Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on, vol. 7, pp. 3028-3030. IEEE, 2010.
9. Modsching, Marko, Ronny Kramer, and Klaus ten Hagen. "Field trial on GPS Accuracy in a medium size city: The influence of built-up." In 3rd Workshop on Positioning, Navigation and Communication, pp. 209-218. 2006.
10. Abadi, Afshin, Tooraj Rajabioun, and Petros Ioannou. "Traffic flow prediction for road transportation networks with limited traffic data." Intelligent Transportation Systems, IEEE Transactions on 16, no. 2 (2015): 653-662.
11. Lipan, Florin, and Adrian Groza. "Mining traffic patterns from public transportation GPS data." In Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing, pp. 123-126. IEEE, 2010.
12. Zhu, Tongyu, Jian Dong, Jian Huang, Songsong Pang, and Bowen Du. "The bus arrival time service based on dynamic traffic information." In Application of Information and Communication Technologies (AICT), 2012 6th International Conference on, pp. 1-6. IEEE, 2012.
13. Cheng, Shaowu, Baoyi Liu, and Botao Zhai. "Bus arrival time prediction model based on APC data." (2010): 165-169.
14. Jeong, Ranhee, and Laurence R. Rilett. "Bus arrival time prediction using artificial neural network model." In Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on, pp. 988-993. IEEE, 2004.
15. Bin, Yu, Yang Zhongzhen, and Yao Baozhen. "Bus arrival time prediction using support vector machines." Journal of Intelligent Transportation Systems 10, no. 4 (2006): 151-158.
16. De Fabritiis, Corrado, Roberto Ragona, and Gaetano Valenti. "Traffic estimation and prediction based on real time floating car data." In Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on, pp. 197-203. IEEE, 2008.
17. Kerner, B. S., C. Demir, R. G. Herrtwich, S. L. Klenov, H. Rehborn, and A. Haug. "Traffic state detection with floating car data in road networks." In Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE, pp. 44-49. IEEE, 2005.
18. Ho, Yao Hua, Yao Chuan Wu, Meng Chang Chen, Tsun-Jui Wen, and Yeali S. Sun. "GPS Data Based Urban Guidance." In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, pp. 703-708. IEEE, 2011.
19. Poomrittigul, Suvit, Setha Pan-ngum, Kunchit Phiu-Nual, Wasan Pattara-atikom, and Panita Pongpaibool. "Mean travel speed estimation using GPS data without ID number on inner city road." In ITS Telecommunications, 2008. ITST 2008. 8th In-ternational Conference on, pp. 56-61. IEEE, 2008.
20. The Weather Underground, LLC., http://www.wunderground.com/history/
21. Bentley, Jon Louis. "Multidimensional binary search trees used for associative searching." Communications of the ACM 18, no. 9 (1975): 509-517.
22. Moon, Tood K. "The expectation-maximization algorithm." Signal processing magazine, IEEE 13, no. 6 (1996): 47-60.

Ilya Ardakani graduated master of engineering degree in Information Communication and Technology for Embedded Systems at Sirindhorn International Institute of Technology, Thammasat University in July 2016. He has a bachelor degree in Mechatronics Engineering from Assumptions University of Thailand. He is now continuing his study in a doctoral program at Tohoku University. He is interested in machine learning, image processing, and control systems research areas.

Thanaruk Theeramunkong is currently a professor at School of Information, Computer and Communication Technology at Sirindhorn International Institute of Technology (SIIT) at Thammasat University, Bangkok, Thailand. He is also the Program Director of Information and Communication Technology for Embedded Systems (ICTES) at TAIST Tokyo Tech, National Science and Technology Development Agency (NSTDA). He received his bachelor degree in Electric and Electronics Engineering, master and doctoral degrees in Computer Science from Tokyo Institute of Technology. He was a research associate at Japan Advanced Institute of Science and Technology in Japan and a MIS manager at C.P. Seven Eleven Public Co., Ltd. in Thailand. He got several awards, including the Very Good Research Award in engineering field from Thammasat University in **2008**, **2009** and **2010.** Recently, in **2014**, he has received the National Outstanding Researcher Award in the field of Information Technology and Communication Arts. He also got several best paper awards from conferences and societies, including the Japanese Society for Artificial Intelligence, PAKDD workshops, and KICSS. In **2015**, he also got a Gold Medal with the Congratulations of the Jury from the **43**rd International Exhibition of Inventions of Geneva for the inventions of automatic semantic-based multi-document summarization and application to public hearing. His research interests are natural language processing, data mining, text mining, machine learning and applications to service science. He is also a member of the Steering Committee of the Pacific-Asia Conferences on Knowledge Discovery and Data Mining (PAKDD). He is an associate editor of the Institute of Electronics, Information and Communication Engineers (IEICE). He is the author of more than **40** papers in a number of journals with impact factors and more than **100** conference papers.

Waree Kongprawechnon is currently an associate professor at School of Information, Computer and Communication Technology at Sirindhorn International Institute of Technology (SIIT) at Thammasat University, Bangkok, Thailand. She received her bachelor degree in Electrical Engineering with a First Class Honor from Chulalongkorn University, Thailand in 1992, her master degree in Control Engineering from Osaka University in 1995, and her doctoral degree in Mathematical Engineering and Information Physics from the University of Tokyo in 1998. Her research interests are theory in $H\infty$ control, Control theory, Robust control, System identification, Modeling, Adaptive control, Learning control, Neural network and Fuzzy control. She received a number of research grants from many institutes including Thailand Research Fund, Thailand National Electronics and Computer Technology Center (NECTEC), Thammasat University, the Council on Higher Education (CHE) and so forth. She published more than 3**0** papers in a number of journals with impact factors and more than 75 conference papers.