Traffic Data Processing in Vehicular Sensor Networks

Xu Li†, Wei Shu♦, Minglu Li†, Pei-En Luo†, Hongyu Huang† and Min-You Wu†

†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

♦Department of Electrical and Computer Engineering, The University of New Mexico, USA

Abstract-**The existing vehicular sensors of taxi companies in most of cities can be used for traffic monitoring, however sensors are always set with a long sampling interval because of communication cost saving and network congestion avoidance. In this paper, we focus on the traffic data processing in vehicular sensor networks providing sparse and incomplete information. A performance evaluation study has been carried out in Shanghai by utilizing the sensors installed on 4000 taxis. Two types of traffic status estimation algorithms, the link-based and the vehicle-based, are introduced based on such data basis. The results from large-scale testing cases show that the traffic status can be fairly well estimated based on these imperfect data and we demonstrate the feasibility of such application in most of cities.**

Keywords-vehicular sensor networks; GPS; Data processing; Intelligent Transportation System(ITS);

I. Introduction

Currently, taxi companies often equip GPS-based sensors on their taxis for effective vehicle dispatching in many cities. These GPS-based mobile sensors can constitute a vehicle-based mobile sensor network and the sensing data can be collected through a vehicular ad hoc network or through a GSM communication network. The taxi companies, however, often set sensor with long sampling interval, such as 1-2 minutes, because they want to reduce communication cost and are only interested in the general locations of their taxis for vehicle dispatching.

In this paper, we focus on the data processing issue in vehicular sensor networks for traffic monitoring [1]. This is a fundamental problem need to be solved. The vehicular sensor networks have an advantage of cost saving than the traditional stationary sensors, such as loop detectors, and/or video cameras, which lead to high cost of infrastructure and maintenance [2]. Overall, available sensor data are only a byproduct of taxi companies, not designed specifically for traffic monitoring. We carried out a performance evaluation study in urban area of Shanghai by utilizing the sensors installed in about 4000 taxis. Sensors can provide longitude and latitude coordinates, timestamp, etc. The average sampling interval is 61 to 129 seconds. Two traffic status estimation algorithms, the link-based (***LBA***) and the vehicle-based (***VBA***), are introduced to compute the real-time mean speed for every segment of roads. Totally 56 testing cases from Aug, 2006 to May, 2007 have been analyzed. The testing results show that the traffic status can be fairly estimated based on these imperfect data provided by these vehicular sensor networks.

II. Related Work

Several works on mobile sensor for traffic monitoring have been carried out in recent years [3][4][5][6][7][8]][11]. Most of them focused on highways or freeways, where traffic light delay is not an issue because there is no intersections on highways and vehicles have certain routes to follow on highways [2]. This is different from the counterpart in urban area [3]. Meanwhile, they mostly assumed that the sensor is set with high sampling rate, such as 1Hz, inevitably implying a considerable communication cost that might cancel the benefits of infrastructure cost saving. A comparison of traffic measurement system with stationary and mobile sensors can be found in [4]. In [5], an algorithm for the arterial road speed estimation was proposed by using taxis equipped GPS sensors in Guangzhou, China. This work is based on a fine-grain data sampling model and only proposes the methodology of how to use sensor data to estimate the traffic status. To our best knowledge, performance and verification of the algorithm has not been reported. Work in [6] [7] uses buses to monitor the arterial traffic status. Similarly, the work is also based on short sampling interval, ranging from every one second to at most 30 seconds. Both works are mainly restricted to arterial routes, and usually not applicable to the fine-grain streets and roads in a metropolitan area, such as Shanghai city. In [8], authors drove a single vehicle for collecting GPS data along a pre-specified loop route repeatedly with the sampling interval of 4 to 10 seconds. Compared to our work, the realistic vehicular sensor networks can cover the entire road network of the city, including arterial and inferior roads. Overall, the existing work is experimental study and only proposed the methodology of such idea; the feasibility and the real testing of accuracy are seldom to be found.

Procedure $\boldsymbol{Estimate}_{LBA}(L_i,t_k)$ //Link-based traffic status estimation

Input: L_i, t_k; and all sensor data pairs collected around t_k

Output: $V_i(t_k)$ // the mean traffic speed of L_i at time t_k

1) Construct a set of data pairs to be used

$\wp(L_i, t_k) = \{ p(s, t_1, t_2) \mid (t_1,t_2) \subseteq (t_k-\tau,t_k+\tau)$

and $Include(L_i, \psi_1,\psi_2) \equiv true \}$

where, τ is a predefined constant used to handle asynchronous data sample timing.

2) Find the length of L_i: $e = \boldsymbol{r}(k_B(L_i), k_E(L_i))$

3) For each sensor s with $p(s, t_1, t_2) \in \wp(L_i, t_k)$

- compute its road distance over L_i as $l(L_i,s)$.
  - if $(\psi_1 \in L_i \wedge \psi_2 \in L_i)$ $l(L_i,s) = \boldsymbol{r}(\psi_1,\psi_2)$
  - if $(\psi_1 \notin L_i \wedge \psi_2 \notin L_i)$ $l(L_i,s) = \boldsymbol{r}(k_B(L_i), k_E(L_i))$
  - if $(\psi_1 \in L_i \wedge \psi_2 \in A(k_E(L_i)))$ $l(L_i,s) = \boldsymbol{r}(\psi_1, k_E(L_i))$
  - if $(\psi_1 \in L_i \wedge \psi_2 \in A(k_B(L_i)))$ $l(L_i,s) = \boldsymbol{r}(\psi_1, k_B(L_i))$
  - if $(\psi_1 \in A(k_E(L_i)) \wedge \psi_2 \in L_i)$ $l(L_i,s) = \boldsymbol{r}(\psi_2, k_E(L_i))$
  - if $(\psi_1 \in A(k_B(L_i)) \wedge \psi_2 \in L_i)$ $l(L_i,s) = \boldsymbol{r}(\psi_2, k_B(L_i))$
- compute weight of *average mobile speed* contributed
  - $\beta_s = l(L_i,s) / e$; // where $\beta_s \leq 1$
- accumulate weighted mean speed:
  - $R = R + \beta_s * r(\psi_1, \psi_2) / (t_2 - t_1)$
  - $U = U + \beta_s$

4) Compute the mean traffic speed of L_i around t_k:

$V_i(t_k) = R / U$;

Boolean Function $\boldsymbol{Include}(L_i, \psi_1,\psi_2)$

Input: L_i, ψ_1,ψ_2; Output: *true* or *false*

$= (\psi_1 \in A(k_B(L_i)) \wedge \psi_2 \in A(k_E(L_i)) \vee (\psi_1 \in A(k_E(L_i)) \wedge \psi_2 \in A(k_B(L_i))$

Figure 1. Link-based algorithm for traffic status estimation

III. ALGORITHMS FOR TRAFFIC ESTIMATION

A. Macroscopic traffic-flow theory

The macroscopic traffic-flow model includes three key characteristics, that is, flow rate, mean speed and density [1]. The public always tends to consider more in terms of mean speed rather than flow rate or density in evaluating the quality of their trips. In this paper, mean speed is also used as a performance measure.

A road network consists of a set of roads embedded in a predefined geographical region, such as metropolitan of Shanghai. A ***link*** is a road segment between two intersections (called ***node***), and a ***road*** consists of several ordered links, but all of them share the same road name [9].

In general, a pair of data sampled consecutively by a same sensor is defined as :

$$p(s, t_1, t_2) = \{\mathbf{s}, \mathbf{t_1}, \boldsymbol{\psi_1}, \mathbf{t_2}, \boldsymbol{\psi_2}\} \qquad \textbf{(1)}$$

where ψ_1 and ψ_2 are obtained by map-matching from the consecutive data samples at t_1 and t_2, respectively. The process of locating sensing data onto a road network map due to the well-known error of GPS device is called *map-matching* [10]. A sensor *s*, with an vehicle, will have its ***Average Mobile Speed*** (**AMS**) during interval (t_1,t_2), denoted as:

$$v(s, t_1, t_2) = \boldsymbol{r}(\boldsymbol{\psi_1}, \boldsymbol{\psi_2}) / (\mathbf{t_2} - \mathbf{t_1}) \qquad \textbf{(2)}$$

where $\boldsymbol{r}(\psi_1,\psi_2)$ is the length of road being traveled between ψ_1 and ψ_2.

In order to estimate the traffic status around time t_k, we need to utilize data collected from a group of associated sensors. More precisely, we use data pair $p(s, t_1, t_2)$ as input of the traffic estimation algorithm. For link L_i with length of l_i, let the ***Mean Traffic Speed (MTS)*** of link L_i at time t_k be denoted as $\boldsymbol{V_i}(t_k)$, which can be obtained by following algorithms with sensor data. If $p(s, t_1, t_2)$ is used for computing MTS of L_i around t_k, we say $v(s, t_1, t_2)$ is a ***Speed Element*** (***SE***) for L_i. The definition of MTS is given as follows:

$$V_i\ (t_k) = \sum_{v \in O_i(t_k)} \left(\frac{l_i^v}{\sum_{v \in O_i(t_k)} l_i^v} \times v\right) \qquad \textbf{(3)}$$

where $V_i(t_k)$ denotes MTS of L_i around time t_k obtained by traffic status estimation algorithms with sensor data, v represents a ***SE*** and $O_i(k)$ is the set of SEs, and l_i^v denotes the length of the segment of L_i which v covers. In addition, we aggregate sensor data from $(t_k-\tau,t_k+\tau)$ for calculation of $\boldsymbol{V_i(t_k)}$ to handle asynchronous data sample timing. More precisely, $p(s, t_1, t_2)$ can be used to calculate $\boldsymbol{V_i(t_k)}$ when $(t_1, t_2) \subseteq (t_k-\tau,t_k+\tau)$, where τ is a predefined constant.

In addition, we analyze the real traffic flow by videotaping to get measurement of MTS, which is regarded as the ***real value*** of MTS (***RMTS***). The formula used is as follows:

$$V_i^R(t_k) = \frac{l_i}{\frac{1}{|C_i(t_k)|} \sum_{c \in C_i(t_k)} \Delta t_c} \qquad \textbf{(4)}$$

where c denotes a vehicle which travels L_i with the time cost Δt_c around time t_k. A vehicle that enters into link L_i between $(t_k-\tau,t_k+\tau)$ is included in a set of vehicles, $C_i(t_k)$, where $|C_i(t_k)|$ is the size of $C_i(t_k)$.

B. Link-based traffic status estimation

First we describe the basic idea of ***Link-based Algorithm*** (***LBA***): LBA is proposed with an assumption that given a link, pairs of sensor data either starting or ending around this link can best reflect traffic status of this particular link. Based on such assumption, given a particular link L_i, LBA

Procedure $\boldsymbol{Estimate}_{VBA}(\mathrm{t_k})$ //Vehicle-based traffic estimation

Input: t_k; and all sensor data pairs collected around t_k

Output: $\boldsymbol{V_i}(t_k)$ for every links // mean traffic speeds at t_k

1) For each link L_i, find the length of L_i: $e_i = \boldsymbol{r}(k_B(L_i), k_E(L_i))$
2) Construct a set of data pairs to be used
   $\wp(t_k) = \{ p(s, t_1, t_2) \mid (t_1,t_2) \subseteq (t_k-\tau, t_k+\tau) \}$
3) For each sensor s with $p(s, t_1, t_2) \in \wp(t_k)$,
   - compute weight of *average mobile speed*
     $\boldsymbol{v}(s, t_1, t_2) = r(\psi_1, \psi_2) / (t_2 - t_1)$
   - construct an ordered list of links
     $Q(s, t_k) = [L_1, L_2, \ldots, L_q]$, *where* $q \geq 1$

   $\boldsymbol{Q}(s, t_k)$ is ***an ordered list of links*** over which sensor s moved through during time (t_k, t_{k+1}).

   - for every $L_i \in Q(s, t_k)$
     i) compute road distance traveled over L_i as $l(L_i,s)$,
        - $l(L_i,s) = \boldsymbol{r}(\psi_a, \psi_b)$, where
          - $\psi_a = (L_i = L_1)\ ?\ \psi_1 : k_B(L_i)$
          - $\psi_b = (L_i = L_q)\ ?\ \psi_2 : k_E(L_i)$
     ii) compute weight of *average mobile speed* contributed
        - $\beta_s(L_i) = l(L_i,s) / e_i$; // where $\beta_s(L_i) \leq 1$
     iii) accumulate weighted mean speed:
        - $R(L_i) = R(L_i) + \beta_s(L_i) * \boldsymbol{v}(s, t_1, t_2)$
        - $U(L_i) = U(L_i) + \beta_s(L_i)$
4) For each link L_i, compute the mean traffic around t_k:
   - $\boldsymbol{V_i}(t_k) = R(L_i) / U(L_i)$;

Figure 2. Vehicle-based algorithm for traffic estimation

only aggregates pairs of sensing data from link L_i as well as links adjacent to either of intersection nodes of L_i for every time duration $(t_k-\tau, t_k+\tau)$. Note that, every link L_i has two intersection node $k_B(L_i) = k_1$ and $k_E(L_i) = k_2$. Every intersection node w defines an ***adjacent set of links*** as:

$$\boldsymbol{A(w)} = \{ L_i \mid k_B(L_i) = w \vee k_E(L_i) = w \} \quad \textbf{(5)}$$

The Link-based algorithm is presented as a procedure $\boldsymbol{Estimate}_{\mathbf{LBA}}(L_i, t_k)$ in Figure 1.

C. Vehicle-based traffic status estimation

The basic idea of ***Vehicle-based Algorithm*** (***VBA***) is described as follows: Compared to LBA, VBA utilizes every available data pairs and disseminates them back to all links traveled to estimate the MTS. Thus, a sensor moving with a vehicle may travel over one or more links, which again can be associated with one or more roads.

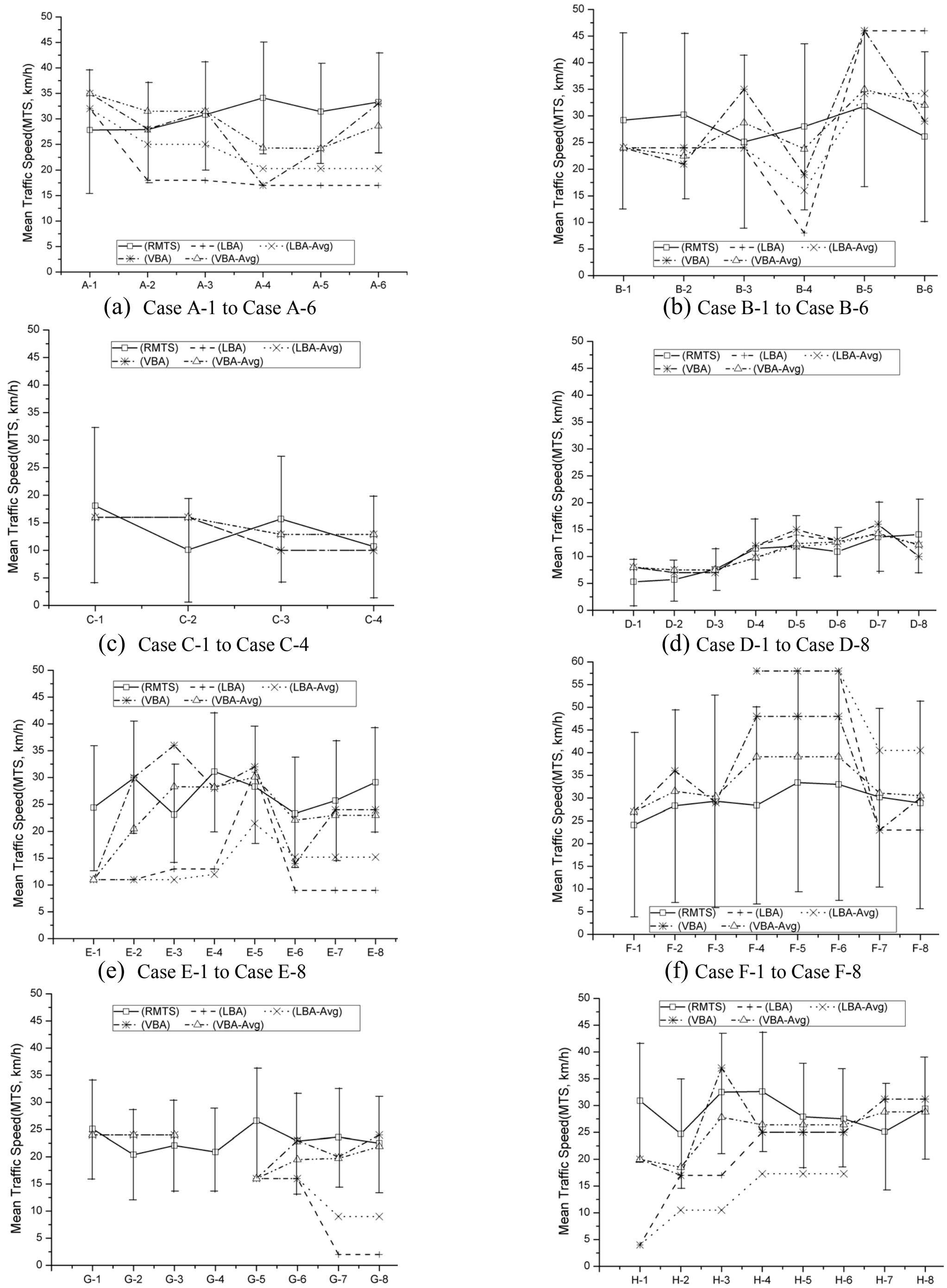
The realistic data background can explain the methodology of VBA that long sampling interval makes two data of a data pair always far from each other. Thus, VBA can make the most use of sensor data to calculate traffic status whereas LBA only uses portion of sensor data. As same as LBA, VBA also calculates $\boldsymbol{V_i}(t_k)$ by using Equ (3). The Vehicle-based algorithm for traffic status estimation is presented as a procedure $Estimate_{VBA}(t_k)$ in Figure 2.

IV. Testing and Performance

A. Testing results of traffic status estimation algorithms

Testing was carried out on several links which belong to different types of roads, including arterial and inferior roads: ZhaoJiaBang road (LinkID=20822, Case A-1 to A-6, Date:2006-8-11 10:10-10:40, 117m, arterial short link; LinkID=16935, Case B-1 to B-6, Date:2006-10-23 09:25-10:00, 296m, arterial long link); FengLin road (LinkID=8373, Case C-1 to C-4, Date:2006-10-24 09:25-09:50, 99m, inferior short link); W.TianMu road (LinkID=5322, Case D-1 to D-8, Date:2007-05-20 09:35-10:15, 154m, arterial short link); HengFeng road (LinkID=3942, Case E-1 to E-8, Date:2007-05-20 10:30-11:10, 244m, arterial long link); ChangShou road (LinkID=4517, Case F-1 to F-8, Date:2007-05-20 12:05-12:45, 277m, arterial long link); JiaoZhou road (LinkID=22167, Case G-1 to G-8, Date:2007-05-20 14:20-15:00, 261m, inferior long link); HuaShan road (LinkID=4611, Case H-1 to H-8, Date:2007-05-20 16:20-17:00, 301m, arterial long link), some cases are not consecutive in time series. In addition, Case A, B and C are based on average sampling interval of 129 seconds while the rest of cases are based on average 61 seconds. Map-matching algorithm is adopted from [9].

We do not present the comparison of performances between our algorithms and existing mechanism in previous work because they are based on different assumptions and data basis. We mainly carried out a performance evaluation study with large scale testing cases from real world, which aims to demonstrate the feasibility of such application. We focus on MTS of unidirection for links around t_k, τ=2.5min. Meantime, we calculate average of results at t_k and t_{k-1}. The average of LBA is denoted as LBA-Avg when the average of VBA is denoted as VBA-Avg. It aims to explore how many improvements can be made with historical information. First, we describe how to estimate traffic status. When algorithms begin to run, for every calculating approach, the first time t_k which has data pairs for calculating is regarded as their respective "initial time". The results of LBA-Avg and VBA-Avg are the same with LBA and VBA



(a) Case A-1 to Case A-6

(b) Case B-1 to Case B-6

(c) Case C-1 to Case C-4

(d) Case D-1 to Case D-8

(e) Case E-1 to Case E-8

(f) Case F-1 to Case F-8

(g) Case G-1 to Case G-8

(h) Case H-1 to Case H-8

Figure 3. Testing results of 16 cases on three different links

respectively at "initial time" because of no historical information. If no data pairs to be used for calculating around t_k, we use the latest historical MTS of $t_{k'}$ as result of t_k when $(t_k - t_{k'}) < 15$min.

Next, we explain the methodology of testing. Figure 3(a)-(h) show that the RMTSs of links often have large standard deviations (SD), which are regarded as real MTSs, so we tend to evaluate our results of algorithms by using the following criterion: If difference of calculating result and RMTS is less than SD of RMTS, we regard it as a reasonable result. Thus, the VBA and VBA-Avg have reasonable results in most of 56 testing cases but many results of LBA and LBA-Avg are not satisfactory.

For some cases, LBA can only use historical result because of no SEs for LBA to calculate MTS. The data background explains this situation that most of sensor data pairs have a long time interval that makes them not on the same link or on the links connected directly with each other. So these data pairs cannot be used in LBA while they can contribute to VBA. Moreover, we found it is very effective method to use latest historical results while no SEs for calculating in current case. The result around t_k can keep valid for the following time because traffic flow do not have a considerable change in a short time, e.g., we deem that it keeps valid in 15 minutes in our work. Meanwhile, the average error of VBA-Avg can be within only 17.3%, a fairly accurate estimation of traffic status, as shown in Table I. It demonstrates feasibility of such vehicular sensor networks that is not designed for traffic monitoring.

Table I also shows the average percentage of difference between results of four algorithms and RMTSs. It is shown that the performance of LBA-Avg and VBA-Avg are better than LBA and VBA, respectively. So we conclude that the historical based results are better for traffic status estimation than only using current results, especially in such a situation that there are few SEs used for calculating and these SEs have abnormal value, such as taxis stop for taking passenger for few minutes. Our testing results do not reflect such situation because these links are always with strict traffic surveillance in daytime, vehicles cannot stop free and easy. In case F-4, the result of LBA is about 58km/h, this is because a vehicle separately sent two data in 10s with high speed while in the middle of link 4517. Thus, even 58km/h is reasonable for this vehicle to travel link 4517, it still has a large difference with the RMTS which is used to describe the *whole* traffic flow of this link.

In addition, we pay more attention on accuracy of LBA and VBA between different sampling intervals, as shown in Table II. It can be found that while sampling interval is shorten from 129 seconds to 61 seconds, the accuracy of algorithms did not change much. This means that even if based on incomplete information of data basis due to long sampling interval, we still make a relatively accurate estimation of traffic status. To model the relationship between accuracy of traffic estimation and the sampling interval, more cases need to be tested. We take this issue as part of future work.

TABLE I. Statistic results of 4 algorithms

LinkID	LBA	LBA-Avg	VBA	VBA-Avg
20822	39.5%	26.5%	17.2%	17.8%
16935	39.2%	20.7%	29.3%	17.6%
8373	28.2%	27.1%	28.2%	27.1%
5322	21.2%	16.3%	22.2%	17.2%
3942	54.2%	47.4%	24.7%	20.1%
4517	59.5%	65.5%	28.2%	13.5%
22167	63.1%	47.9%	13.2%	14.9%
4611	34.7%	55.7%	19.1%	15.0%
Total	**41.4%**	**37.4%**	**22.6%**	**17.3%**

TABLE II. Statistic results between different sampling intervals

Average Sampling Interval	LBA	VBA
129 seconds	35.6%	24.9%
61 seconds	46.5%	21.5%

V. CONCLUSION

In this paper, we focus on utilizing the existing vehicular sensors of taxi companies for traffic monitoring. The sensor used is set with long sampling interval because of low communication cost and avoidance of network congestion. We adopted two types of traffic status estimation algorithms, the link-based and the vehicle-based. The testing result shows that the traffic status can be fairly well estimated and demonstrates the feasibility of such application.

ACKNOWLEDGMENTS

This research was partially supported by Natural Science Foundation of China grant No.60573138 and 60773091, the National Grand Fundamental Research 973 Program of China under Grant No.2006CB303000, the National 863 Program grant No.2005aa122320, and National Science Foundation of USA grant CNR-0626380.

REFERENCES

[1] D.Ni. Determining traffic-flow characteristics by definition for application in ITS. *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 181-187, 2007.

[2] Y.Cho. Estimating velocity fields on a freeway from low-resolution videos. *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp.463-469, 2007.

[3] C.Nanthawichit, et al. Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway. *Transportation dat*a research, no 1855, pp.49-59, 2003.

[4] R.Kolbl. Probe vehicle: a comparison of motorway performance measure with other motorway flow detection techniques. *IEE Conference Publication*, no. 486, pp. 182-186, 2002.

[5] L.Zou, et al. Arterial speed studies with taxi equipped with global positioning receivers as probe vehicle. *International Conference on Wireless Communications, Networking, and Mobile Computing (WCNM)*, 2005.

[6] R.L.Bertini, S.Tantiyanugulchai. Transit buses as traffic probes: Use of geolocation data for empirical evaluation. *Transportation Research Record (Journal)*, no. 1870, pp.35-45, 2004.

[7] P.Chakroborty, S.Kikuchi. Using bus travel time data to estimate travel times on urban corridors. *Transportation Research Record (Journal)*, no. 1870, pp.18-25, 2004.

[8] J. Yoon. Surface Street Traffic Estimation. *The 5th Int'l Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2007.

[9] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[10] M.A.Quddus, W.Y.Ochieng. Integrity of map-matching algorithms. *Transportation Research Part C: Emerging Technologies*, vol.14, no. 4, pp.283-302, 2006.

[11] M.Chen. Dynamic Freeway Travel Time Prediction Using Probe Vehicle Data. *Transportation Research Record (Journal)*, 2001.