IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 2, 2019, accepted February 22, 2019, date of publication March 5, 2019, date of current version May 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903195

When Road Information Meets Data Mining: Precision Detection for Heading and Width of Roads

XUN ZHOU[1], XUELIAN CAI[1], YUEHANG BU[1], XI ZHENG[2], JIONG JIN[3], (Member, IEEE), TOM H. LUAN[4], AND CHANGLE LI[1], (Senior Member, IEEE)

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[3]School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC 3122, Australia
[4]School of Cyber Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Xuelian Cai (xlcai@mail.xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1801266 and Grant 61571350, and in part by the Key Research and Development Program of Shaanxi under Contract 2017KW-004, Contract 2017ZDXM-GY-022, Contract 2018ZDXM-GY-038, and Contract 2018ZDCXL-GY-04-02.

ABSTRACT Real-time road information plays a crucial role in enabling intelligent transportation systems (ITS) applications. With sufficient road information, the map of road topography can be built and updated more easily. Furthermore, many appealing ITS applications can be enabled accordingly. Aiming at improving the quality and update rate of road information, a hot topic today is how to mine information from global positioning systems (GPS) trajectories by the clustering-based methods. Such schemes, however, encounter two challenges: 1) GPS noise and 2) low sampling rate of GPS traces data. As a result, it is difficult to infer road information from these irregular clusters. To tackle the above issues, we directly mine useful road information, heading, and width of roads, for ITS applications from GPS point cloud, i.e., a set of GPS points. First, the distribution of GPS points is discussed and the least squares method (LSM) is demonstrated to be outstanding for mining the heading of the road under a huge number of GPS points. Second, the weighted approximation least squares method is proposed to improve the accuracy of the LSM. Furthermore, combining with relevant distribution features in GPS points, the data distribution variance-road width discrete model is proposed to mine road width from GPS point cloud. Finally, using real-world datasets, we demonstrate that these proposed methods can achieve satisfactory performance in practice.

INDEX TERMS Intelligent transportation systems, road information, GPS data, data mining.

I. INTRODUCTION

Road information is one of the most important resources for intelligent transportation systems (ITS), such as autonomous driving [1], navigation [2], [3] and traffic management [4]–[7]. For instance, Gelso and Sjoberg *et al.* [8] combine road information with traffic data to generate a consistent threat assessment in rear-end near-crashes, which is useful for promoting Collision Avoidance Systems in autonomous driving. Moreover, Liu and Zhang *et al.* [9] propose an instantaneous optimization method based on the recognition of traffic conditions from road information. This method can improve navigation systems of hybrid electric vehicles. Chen *et al.* [10] apply a clustering-based method to mine road information from GPS data, aiming to reconstruct accurate digital maps for autonomous driving. Therefore, aiming to popularize intelligent transportation systems, it is necessary to improve data accuracy and timeliness of road information.

Traditionally, this crucial road information is obtained through ground measurement [11]. It means moving devices equipped with sensors and visual systems are used to collect original information of road segments. This ground measurement is regarded as a time-consuming and expensive method. Recent years, with the research on ad hoc network, various individual mobile devices, such as unmanned aerial vehicles and robots [12], can collect road information simultaneously and share it with each other in this network. This way is effective to improve the collection rate for road information.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhong-Ke Gao.

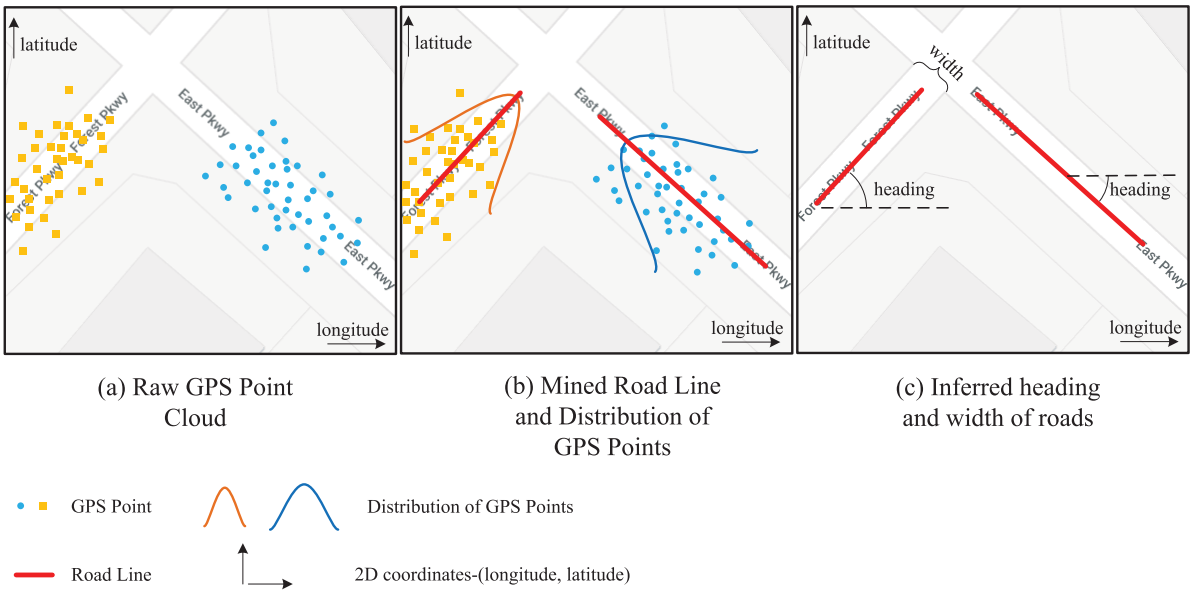


FIGURE 1. Mining road information from GPS point cloud. (a) Raw GPS point cloud. (b) Mined road line and distribution of GPS points. (c) Inferred heading and width of roads.

However, the downside of it is that devices with high-quality sensors, reliable communication systems and a large number of infrastructures are demanded, which greatly increases cost for road information collection.

With wide use of Global Positioning Systems (GPS), a large amount of data about mobile terminal locations and states can be collected and uploaded timely [13], [14]. Compared with information obtained by special devices, such as vehicles equipped with visual systems and unmanned aerial vehicles, acquisition of GPS data is low-cost and even some open databases of GPS data are free [15], [16]. In addition, GPS data are updated frequently with a refresh cycle every several minutes or even every few dozens of seconds, which is beneficial for updating road information timely.

Considering these advantages of GPS data mentioned above, the paper [17] proposes a clustering-based framework that mines road information from GPS traces. By clustering similar vehicular trajectories, this work successfully obtains road structure. The paper [18] carefully divides this framework in [17] into three steps, i.e., clustering, linking and smoothing. To be specific, in the clustering procedure, GPS traces which are similar in heading and position are combined together, trying to generate cluster centres. In the second step, these cluster centres are linked in order to reconstruct a geometric construction of the road. Finally, in the smoothing step, remaining refined work is completed, such as generating arcs and intersections between adjacent road segments. These clustering-based methods reduce the cost of collecting and updating road information in real-time.

However, for these clustering-based methods, there are two challenges: 1) GPS noise; and 2) the low sampling rate of GPS traces data, that impact the reliability of these methods. For example, GSP noise results in the collected GPS data being far from their actual locations and even being not on roads. Also, the low sampling rate leads to temporal uncertainties, because the line between two logged locations is too long to be matched with the real route. These two factors seriously limit use of the clustering-based methods.

In this paper, the problem about collecting road information can be defined as how to transform a set of raw GPS data to useful road parameters (in Fig. 1). Considering the aforementioned low sampling rate of GPS traces, we mine road information, heading and width of roads, from GPS point cloud, i.e., a set of GPS points including instantaneous locations of devices (in Fig. 1(a)). To be specific, by analysing the Gaussion distribution of GPS points, the least squares method (LSM) is demonstrated to be effective to mine the heading of the road line under a huge number of GPS points (in Fig. 1(b)). Furthermore, the weighted approximation least squares method (WALSM) is applied to improve the accuracy of the LSM under realistic datasets, and the Data Distribution Variance - Road Width Discrete Model (DV-RWDM) is proposed to discover correlation between the road width and the variance of GPS points to mine the road width (in Fig. 1(c)). The real-world GPS data collected by Taiwan, China government from local buses are applied to verify our models. In order to ensure road information can be obtained and updated in real time, execution time of these mining models is 15s (more details of GPS data is in Section IV, Part A). In summary, our contributions in this paper are as follows:

1) The performance of LSM on mining the heading of the road is analysed, then a WALSM is proposed to improve its accuracy.
2) A DV-RWDM is proposed to mine the width of the corresponding road by correlating the variance of GPS points and the width of the road.
3) The performance of our proposed solutions is verified by using the real-world GPS data.

The rest of this paper is organized as follows. In Section II, we introduce background and relevant work. Then we analyse our data mining models in details in Section III. In Section IV, we conduct simulation experiments to evaluate our data mining methods. The conclusion is in Section V.

The notations used are summarized in Table 1.

TABLE 1. Main notations.

Notation	Explanation
x	latitude
y	longitude
x_i	latitude of the point i
y_i	longitude of the point i
a_0	heading of road centerline
a_0'	estimation for a_0
b_0	position of the road centerline
b_0'	estimation for b_0
l_i	perpendicular distance
η	variance of Gaussian distribution
σ	learning rate
$DIS_{threshold}$	threshold value
set	GPS points set

II. BACKGROUND AND RELATED WORK

A. BACKGROUND

1) ROAD INFORMATION

Information about curvature, elevation, width, position, and heading of roads has received much attention in the context of civil engineering as well as for appealing ITS applications [10], [13], [19]. To illustrate, for autonomous driving and navigation, travel direction of moving devices should be consistent with the heading of the road to ensure them safely driving and reaching destinations. Furthermore, the accurate width of the road can support prediction of traffic flow and congestion in short-time.

2) 2D COORDINATES

In order to apply appropriate methodologies to research, a 2D coordinates-(*longitude*, *latitude*) is established to describe road information. For example, in this 2D coordinates, the heading of the road is defined as the angle between this road line and *longitude*-axis (in Fig. 1(c)). For other example, GPS points are the results of GPS data projected on this 2D coordinates and each GPS point includes the instantaneous information of the mobile unit, such as latitude, longitude and speed. These dense GPS points construct unstructured GPS point cloud.

3) ROAD CENTERLINE

In much literature, the road centerline is discussed frequently. The reason is that lanes on the road are often seen as parallel, so from the road centerline, information, such as curvature, elevation, position, and heading of the road can be extracted. Furthermore, combining with the road width, the structure of this road can be recognized completely.

B. RELATED WORK

Mining road information from raw GPS data aims to replace traditional labour-intensive ways for collecting and updating this information. However, two challenges mentioned above restrain applications of the clustering-based methods seriously.

To overcome the challenge of the low sampling rate of GPS traces, existing work focuses on directly mining road information from GPS point cloud. The paper [19] proposes the Spatial-Linear Cluster (SLC) algorithm to obtain a long and thin cluster from the entire set of GPS points and applies a geometric approach to estimate road segments from each cluster. As for the paper, Chen *et al.* [10] attempt to cluster the unstructured GPS point cloud in the area around road centerline and link these cluster centres to recognize the structure of the road segment. Then with prior knowledge on road design, the up-to-date map can be automatically obtained. Corresponding experimental results show these two methods can reconstruct more accurate road information than those methods based on GPS traces.

In contrast to clustering-based methods, aiming to resist GPS noise, another kind of data-driven models is based on the distribution features of GPS data to mine road information. GPS noise is often seen as Gaussian noise in much literature, thus the distribution of GPS points is considered as Gaussian distribution [20]. For instance, the paper [11] proposes to apply the LSM to fit the centerline of road segments, because this method is inspired to fit traces from scattered points under Gaussian noise. In the 2D coordinates-(*longitude*, *latitude*), this road centerline can show the detail information, position and heading, of the road. For other example, the paper [21] models the spread of these GPS traces as a Gaussian distribution. It is beneficial for researchers to explore the width of the road from the distribution of these GPS traces. In summary, by analysing the random distribution of GPS data, these methods mentioned above can establish the relationship between road information and fixed parameters of the random distribution, such as variance.

Inspired by these seminal works mentioned above, in the paper, we devote to mine the road information, heading and width of roads, from raw GPS point data. To be special, the LSM is demonstrated to be effective to mine the heading of the road at first. However, due to the system error in GPS data, the LSM is hard to estimate the accurate position of the road centerline. Then, a WALSM is applied to improve the accuracy of the LSM. Finally, DV-RWDM is proposed to estimate the road width and effectively enrich the information mined from raw GPS data.

III. OUR DATA MINING MODELS

In this part, the 2D coordinates-(*longitude*, *latitude*) is applied to analyse GPS data and mine extra information. Moreover, according to related work [20], [21],

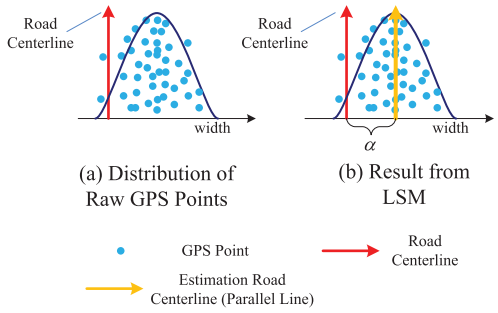


FIGURE 2. Mining road information by LSM. (a) Distribution of raw GPS points. (b) Result from LSM.

the distribution of GPS points is seen as Gaussian distribution. The details about GPS data are in the Section IV part A and relevant statistical method is also applied to verify their Gaussian distribution.

A. THE LEAST SQUARES METHOD AND GAUSSIAN DISTRIBUTION

In the 2D coordinates, let x-axis denotes longitude and y-axis denotes latitude, the centerline of the road segment can be defined as the linear function (1),

$$y = a_0 * x + b_0, \tag{1}$$

where slope a_0 and intercept b_0 define the heading and the position of the road centerline respectively (though the curve road segment may be a complex function, it can be shown as a combination of several linear segments by linking neighbour intersections with linear edges [17]). In much literature these GPS points present Gaussian distribution around the centerline of the road (in Fig. 2(a)), which means as (2) and (3),

$$l_i = \frac{a_0 * x_i + b_0 - y_i}{\sqrt{{a_0}^2 + 1}}, \tag{2}$$

$$p(l_i) = \frac{1}{\eta\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{l_i - \alpha}{\eta})^2), \tag{3}$$

where x_i and y_i denote the longitude and latitude of the ith GPS point, (x_i, y_i), in the data set. In addition, l_i denotes the perpendicular distance between point (x_i, y_i) and the road centerline. (3) means that these GPS points around the road centerline following Gaussian distribution $l \sim N(\alpha, \eta^2)$. Gaussian distribution is leaded by random error in GPS data and bias α is contributed by system error in GPS.

In (2) the $\sqrt{{a_0}^2 + 1}$ is constant for each point (x_i, y_i), and we define variable $l_i' = l_i * \sqrt{{a_0}^2 + 1}$. Therefore, $l' \sim N(f(\alpha), \eta'^2)$ where $f(\alpha) = \alpha * \sqrt{{a_0}^2 + 1}$. From the point view of probability theory, the maximum likelihood estimation for parameters (a_0, b_0) can be inferred as follows:

$$\begin{aligned}(a_0', b_0') &= \arg\max_{(a,b)}(\ln \prod_{i=1}^{n} \frac{1}{\eta'\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{l_i' - f(\alpha)}{\eta'})^2)) \\ &= \arg\max_{(a,b)}(-\frac{1}{2\eta'^2} \sum_{i=1}^{n} (l_i' - f(\alpha))^2 - n \ln \eta'\sqrt{2\pi}),\end{aligned} \tag{4}$$

where n is number of GPS point samples and (a_0', b_0') is the maximum likelihood estimation result for (a_0, b_0) from the n samples. Considering $l_i' = a * x_i + b - y_i$, the (a_0', b_0') can be estimated from (5),

$$\begin{aligned}(a_0', b_0') &= \arg\min_{(a,b)}(\sum_{i=1}^{n} (l_i' - f(\alpha))^2) \\ &= \arg\min_{(a,b)}(\sum_{i=1}^{n} (a * x_i + (b - f(\alpha)) - y_i)^2).\end{aligned} \tag{5}$$

However, corresponding estimation for (a_0, b_0) under LSM is shown as (6)

$$(A_0', B_0') = \arg\min_{(A,B)}(\sum_{i=1}^{n} (A * x_i + B - y_i)^2), \tag{6}$$

where $A_0' = a_0'$ and $B_0' = b_0' - f(\alpha)$. It means LSM can be used to fit the heading of the road centerline a_0, and however, the position of the estimation road centerline under LSM has bias $f(\alpha)$ with maximum likelihood estimation position b_0'. In other word, LSM only can be used to fit the parallel line $y = a_0 * x + b_0 - f(\alpha)$, because GPS points are really distributed with this line for $N(0, \eta'^2)$ (in Fig. 2(b)). This bias α is from system error in GPS.

According to the law of large numbers, only when the number of samples tends to be a very large number or to be infinite, the result of maximum likelihood estimation will coincide with the real result. Therefore, the estimation heading of the road centerline A_0' from the LSM only gradually approximates to the real heading of the road centerline a_0 with the number of GPS point samples increasing.

B. WEIGHTED APPROXIMATION LEAST SQUARES METHOD

When the number of GPS points tends to be infinite, under the law of larger numbers, the maximum likelihood estimation distribution of these GPS points coincides with their real Gaussian distribution. In other words, the estimation A_0' from the LSM coincides with the heading a_0 of the real road centerline under this situation. However, the amount of data is insufficient to meet this condition in the real world. Therefore, improvements should be required for the LSM.

First of all, the influence from these GPS points for the estimated result in (6) is analysed. Under the gradient descent method, the calculating process of (6) can be shown as its partial derivative in (7) and (8),

$$\frac{f(A, B)}{\partial A} = \sum_{i=1}^{n} x_i(A * x_i + B - y_i) = 0, \tag{7}$$

$$\frac{f(A, B)}{\partial B} = \sum_{i=1}^{n} (A * x_i + B - y_i) = 0, \tag{8}$$

where the (A_0', B_0') are unique determined by the (7) and (8) meeting the requirement in (6). It means that starting from any initial value (A, B), the unique parameters (A_0', B_0') which corresponds to the minimum value in (6) can be obtained in

finite time by the gradient descent. In addition, if the $(a_0, b_0 - f(\alpha))$ is assumed as the initial value of (A, B), it is unable to meet the (7) and (8), since there are differences between the maximum likelihood estimation and real results. Its updated parameters are shown as,

$$A_0' = a_0 - \sum_{i=1}^{n} x_i(a_0 * x_i + (b_0 - f(\alpha)) - y_i), \quad (9)$$

$$B_0' = b_0 - \sum_{i=1}^{n} (a_0 * x_i + (b_0 - f(\alpha)) - y_i), \quad (10)$$

which also denotes differences between real results and estimation results. In (9), x_i denotes the longitude and the content in parentheses denotes the distance L_i' between GPS point (x_i, y_i) and the parallel line $(a_0, b_0 - f(\alpha))$. When this GPS point (x_i, y_i) with a distance L_i', which can find a symmetric point (x_i, y_i') with distance $-L_i'$, it cannot contribute update for parameter (A_0', B_0'). However, in the realistic finite datasets when a point with large distance appears, it is hard to find its symmetric point and easy for this point to contribute a large error for the estimation result.

Then, in order to reduce the differences between real results and maximum likelihood estimation, a weighted factor $\omega_i = 2/(1 + \exp(L_i^2/2\sigma))$ is used to update locations of GPS point i by the (11) and (12),

$$x_i := x_i + I(a_0)(1 - \omega_i) * L_i * sin(arc(a_0)), \quad (11)$$

$$y_i := y_i - I(a_0)(1 - \omega_i) * L_i * cos(arc(a_0)), \quad (12)$$

where L_i is the perpendicular distance between the point (x_i, y_i) and the initial parallel line $(a_0, b_0 - f(\alpha))$, and $I(a_0)$ denotes that if a_0 is greater than 0, the $I(a_0)$ is 1, otherwise, $I(a_0)$ is -1. The σ is an hyper-parameter which is related to the learning rate. Actually, the (11) and (12) reduce the perpendicular distance between the point (x_i, y_i) and the parallel line $(a_0, b_0 - f(\alpha))$ to $\omega_i * L_i$. The weighted factor ω_i is chosen from the view that its value is approximately equal to 1 when L_i is relatively small, and its value is approximately equal to zero when L_i is relatively large. Due to the factor ω_i is only related with attribute L_i, if the corresponding symmetric points (x_i, y_i') and (x_i, y_i) can be matched, they cannot update the parameters (A, B). As for the remaining points, with the distance L_i increasing, the factor ω_i is increasing to reduce its influence.

Finally, this improved LSM is named Weighted Approximation Least Squares Method (WALSM) whose process is shown in Fig. 3, and its pseudocode is shown in Algorithm 1.

The input dataset D includes m GPS points collected from the corresponding road segment. The related parameters, learning rate σ and threshold value $DIS_{threshold}$, are input at first. Parameter n denotes the iteration times. The function $L(\bullet)$ denotes using the LSM to calculate parameters (a_n, b_n) from m samples. The function $distance_n(\bullet)$ denotes the method in (2) which calculates the distance for GPS point samples with line (a_n, b_n). Firstly, the LSM is used to estimate the initial road line (a_1, b_1) as benchmark and calculate the initial distance L_i for each point (x_i, y_i) (in Fig. 3(a)). Then, based on the weighted factor ω_i of each GPS point sample, locations of these points are adjusted and used as new input for updating parameters (a_n, b_n) and L_i (in Fig. 3(b)-3(c)). When all the GPS points cluster together, i.e., distance L_i of each point is no more than the $DIS_{threshold}$, the iteration stops. Its output (a^*, b^*) of the Algorithm 1 corresponds to the (A_0', B_0') in (6) and A_0' is the estimation heading of the road centerline.

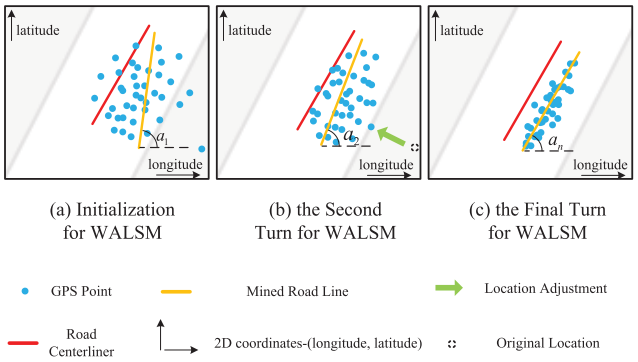


FIGURE 3. Mining the road heading by WALSM. (a) Initialization for WALSM. (b) The Second turn for WALSM. (c) The final turn for WALSM.

Algorithm 1 Weighted Approximation Least Squares Method

Input: (1) GPS points dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, (2) σ, (3) $DIS_{threshold}$
output: (a^*, b^*)
/*Initialization*/
$n = 1$
$D_n = D$, $(a_n, b_n) = L(D_n)$
for inputCoutput pair (x_i, y_i) in D_n **do**
  $L_i = distance_n(x_i, y_i)$
end for
  /*Weighting and approximating*/
  while $max(abs(L)) >= DIS_{threshold}$ **do**
    for inputCoutput pair (x_i, y_i) in D_n **do**
      if $a_n >= 0$ **then**
        $x_i := x_i + (1 - \omega_i) * L_i * sin(arc(a_n))$
        $y_i := y_i - (1 - \omega_i) * L_i * cos(arc(a_n))$
      else
        $x_i := x_i - (1 - \omega_i) * L_i * sin(arc(a_n))$
        $y_i := y_i + (1 - \omega_i) * L_i * cos(arc(a_n))$
      end if
    end for
    $n \leftarrow n + 1$
    $D_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
    $(a_n, b_n) = L(D_n)$
    for inputCoutput pair (x_i, y_i) in D_n **do**
      $L_i := distance_n(x_i, y_i)$
    end for
  end while
  $(a^*, b^*) := \frac{1}{n} \sum_{i=1}^{n} (a_i, b_i)$

C. DISTRIBUTION VARIANCE - ROAD WIDTH DISCRETE MODEL

Considering GPS points around the road centerline with Gaussian distribution, it is inspired to correlate distribution variance of GPS points with the road width. Furthermore, because the width of the lane is fixed in majority countries, number of lanes of these road segments can also be estimated. In this part, a Distribution Variance - Road Width Discrete Model (DV-RWDM) is proposed to estimated the road width.

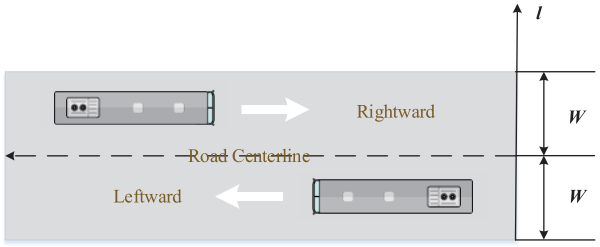


FIGURE 4. Bi-directional road.

As shown in Fig. 4, it supposes that the road is bi-directional and the width of each unidirectional (leftward and rightward) road is W. In much literature, distribution of GPS points on the unidirectional road can be regarded as Gaussian distribution around the road centerline, shown as follows,

$$l_{left} \sim N(\frac{W}{2} + \alpha, \sigma^2_{left}), \tag{13}$$

$$l_{right} \sim N(-\frac{W}{2} + \alpha, \sigma^2_{right}), \tag{14}$$

where l denotes the perpendicular distance between GPS point and corresponding bi-directional road centerline. Furthermore, (13) and (14) denote GPS points on the leftward road, i.e., set_{left} and the GPS points on the rightward road, i.e., set_{right} are around the road centerline with Gaussian distribution ($(\frac{W}{2})$ and $(-\frac{W}{2})$ are locations of leftward road centerline and right road centerline respectively, and bias α is contributed by system error in GPS noise). In addition, the GPS points set_{road} on the bi-directional road is the combination of set_{left} and set_{right}, i.e., $set_{road} = (set_{left} \cup set_{right})$.

The distribution probability density function of GPS points on the bi-directional road can be show as follows (15):

$$f_{road}(l) = a_{left} f_{left}(l) + a_{right} f_{right}(l), \tag{15}$$

where a_{left} and a_{right} relate with number of GPS points on each direction and $f(l)$ is the probability density function. Actually, $a_{left} = \frac{set_{left}}{set_{left}+set_{right}}$ and $a_{right} = \frac{set_{right}}{set_{left}+set_{right}}$. From the definition of variance, the variance $D(l_{road})$ can be shown as (16):

$$D(l_{road}) = \int_{-\infty}^{+\infty} [l - \mathrm{E}(l_{road})]^2 f_{road}(l) dl, \tag{16}$$

where $E(l)$ denotes the expectation of distance l. By combining with (15), (16) can be extended as (17):

$$D(l_{road}) = \int_{-\infty}^{+\infty} [l - \mathrm{E}(l_{road})]^2 (a_{left} f_{left}(l) + a_{right} f_{right}(l)) dl, \tag{17}$$

where a_{left} and a_{right} is unfixed, because the total number of data set_{left} and data set_{right} may be unbalanced. Considering to simplify (17), data preprocessing may be applied to make the a_{left} equal to a_{right}, and both are set as 0.5. For example, GPS data is often matched with the identifiable attributes, such as the ID of mobile devices, so these attributes can be useful for removing the redundant data, and making the scale of these two datasets set_{left} and set_{right} equal. As a result, (17) can be converted to

$$\begin{aligned} D(l_{road}) &= \frac{1}{2} \int_{-\infty}^{+\infty} l^2 (f_{left}(l) + f_{right}(l)) dl - E(l_{road})^2 \\ &= \frac{1}{2} [E(l^2_{left}) + E(l^2_{right})] - E(l_{road})^2. \end{aligned} \tag{18}$$

and because the equation $E(l^2) = D(l) + E(l)^2$, finally, we get (19):

$$D(l_{road}) = \frac{1}{2}(\sigma^2_{left} + \sigma^2_{right}) + \frac{W^2}{4}. \tag{19}$$

It is effective to combine the road with W and variance of GPS points on the road. The W can be estimated by the following formula:

$$W = 2\sqrt{D(l_{road}) - 1/2[D(l_{left}) + D(l_{right})]}, \tag{20}$$

which is the DV-RWDM model proposed in this paper. It is well known, the width of the lane is set to 3.5 meters in most countries, so it is easy to get the road lanes numbern from the road width W. In this paper, $D(l_{road})$, $D(l_{left})$ and $D(l_{right})$ can be calculated based on the road line estimated by the WALSM.

IV. EXPERIMENTS

In this section, realistic GPS point data are used to verify our methods. Firstly, the real-world GPS data are shown, and distribution of these data has been demonstrated to be Gaussian distribution by the statistical approach Kolmogorov-Smirnove (K-S) test. Then, real-world GPS point data are put into the WALSM and the DV-RWDM to mine the heading and the width of the corresponding road.

A. EXPERIMENTAL SETUP AND INITIAL FINDINGS

1) RAW DATA AND DATA PREPROCESSING

Real-world GPS data gathered from buses in Taiwan, China are used for our study. These GPS data start on April 1, 2013 and end on April 30, 2013, collected every 20 seconds from buses which spread across the Taiwan, China (these real GPS datasets provided by our partners to support this project. Now, other relevant work is still progressing based on

TABLE 2. Example of the raw data.

Longitude	Latitude	Speed	Orientation	GPS time
120.3096967	22.63766833	39	272	064951
120.3069467	22.63763500	43	272	225421
120.3103700	22.63762333	37	271	004551
120.3084467	22.63762667	35	269	004611
120.3084467	22.63762667	35	269	004611
120.3105833	22.63768000	38	272	023808
...	...	...	...	...

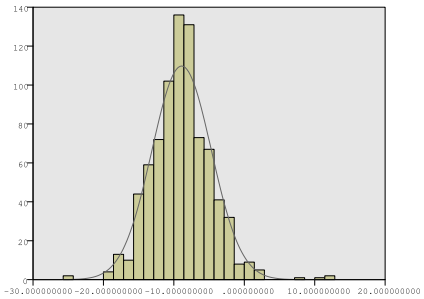
these datasets. When the relevant work is completed, we will try our best to open these datasets). Table 2 shows the example of GPS data, which mainly includes these contents as follows:

1) Latitude and Longitude: indicating bus's location uploaded based on the scheme of WGS 84;
2) Speed: indicating the speed (km/h) of the bus when the information is uploaded;
3) Orientation: indicating the moving direction of bus (from 0-360 degrees);
4) Timestamp: indicating the time when the information is uploaded. (Its format is hhmmss, for instance 105020 indicates the time of 50 minutes and 20 seconds past 10 am).

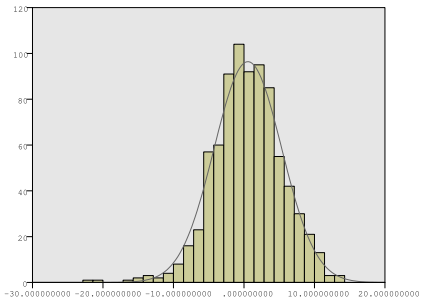
FIGURE 5. Projection of GPS points.

Fig. 5 is the projection of GPS points of two buses on electronic map (blue and yellow points respectively represent instantaneous locations of the two bueses). In order to improve the quality of data, the raw GPS data needs to be cleaned. The specific steps are as follows:

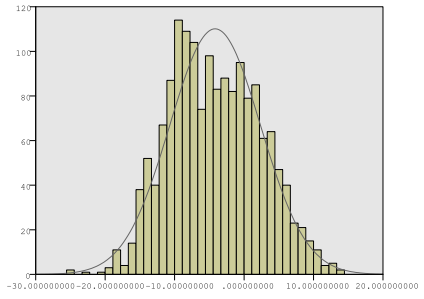
1) Remove the repeating data and replace the missing/wrong data with the mean of their closest surrounding values. The LSM is also used to estimate initial distance L_i of each point. If the initial distance is more than 20 meters, the corresponding point should be removed (in normal condition, GPS errors are within 20 meters);
2) Extract GPS points on single direction road. Because traffic flow on the each direction of the bi-direction roads is uneven and independent. Therefore, we utilize the attribute orientation in GPS data to extract GPS points on each direction, which is necessary to accurately estimate the distribution of GPS points on each direction.



(a)



(b)



(c)

FIGURE 6. GPS data error characteristics. (a) Statistical histogram of unidirectional (leftward) GPS points. (b) Statistical histogram of unidirectional (rightward) GPS points. (c) Statistical histogram of bi-directional GPS points.

2) GPS DATA CHARACTERISTICS

In this part, characteristics of GPS errors and its distribution are discussed. Just as the results mentioned in much literature, the GPS noise from lots of random factors lead to about 20 meters error for GPS data. Combining with central limit theorem, under the noise, GPS points tend to follow the Gaussian distribution around the road centerline. The details are explained below:

1) GPS errors characteristics: In Fig. 2 and Fig. 3, the random distribution of GPS points and bias between road centerline and centre of this distribution have been shown. In fact, there are two kind of GPS errors: system

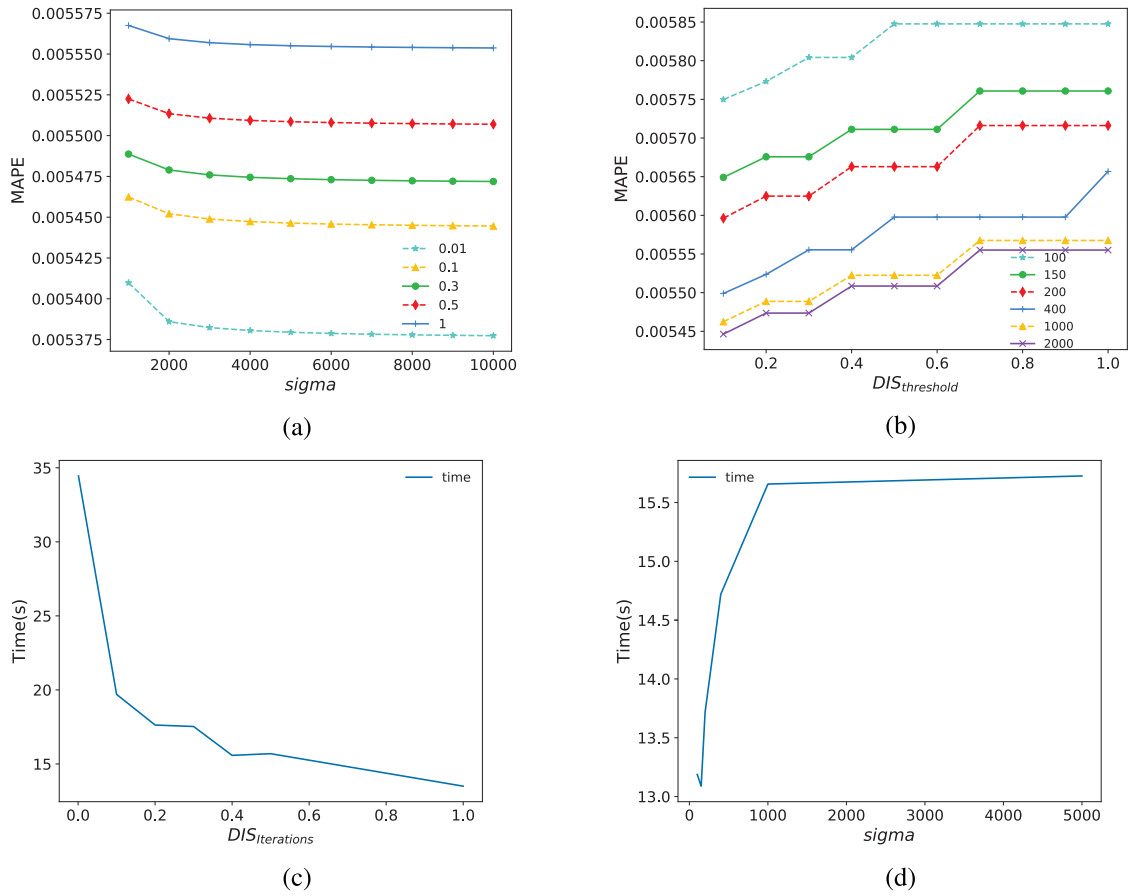


FIGURE 7. Relationship between the value of parameters and MAPE/TIME.

error and random error. System error leads to the same deviation on locations for all GPS points. Random error makes these points randomly distribute. The system error is caused by satellite ephemeris error, ionospheric refraction and so on [22]. The random error is caused by multipath effect, receiver's noise and so on.

2) GPS data distribution characteristics: there are lots of random factors, such as weather and multipath effects, which effect the distribution of GPS points. Furthermore, these random factors tend to force these GPS points following Gaussian distribution with road centerline according to the central infinite theorem [11], [20], [23]. In addition, system error forces this Gaussian distribution with expectation $f(\alpha)$ which is related system error α.

In this paper, the distribution of GPS points is analysed by the statistical approach K-S test. Fig. 6 indicates the comparison of the Gaussian distribution curve and the statistical histograms of GPS points. Furthermore, Fig. 6(a), 6(b) and 6(c) corresponds to the scenario in Fig. 4 respectively. The K-S test is performed on GPS points set_{left}, set_{right} and set_{road} to test whether the distribution of measured points is a known theoretical distribution. Table 3 shows the results of the K-S test for the three data sets (if the value, Asymp. Sig., is greater than 0.1, samples in corresponding data set are following the Gaussian distribution). From the test results in Table 3, the distribution of GPS points on the unidirectional roads coincides with Gaussian distribution. However, GPS points on the bi-directional road are unable to meet this assumption. In the paper, [20] a mixture Gaussian distribution is proposed to account for GPS points distribution on the bi-directional road.

TABLE 3. The result of K-S test.

	Number of samples	Asymp. Sig.
bi-directional	1624	0.00
unidirectional(left)	812	0.265
unidirectional(right)	812	0.546

B. EXPERIMENTAL RESULTS

1) PARAMETERS INITIALIZATION FOR WALSM

It is important to select proper parameters in the WALSM for achieving desired results. In this paper, there are two parameters in the WALSM: σ and $DIS_{threshold}$. We use real data as a training set to analyse different results from different parameters selection. Mean Absolute Percentage

Error (MAPE) is used to measure the error. Fig. 7 indicates MAPE and program execution time with different parameters.

Fig. 7(a) shows the relationship between σ and MAPE, each curve with a constant value of the $DIS_{threshold}$. From this figure, it is clear that with the σ increasing, MAPE decreases gradually. Note that these values of MAPE decrease rapidly with σ from 1000 to 2000 and when σ is more than 2000, the variation of MAPE is not obvious.

Fig. 7(b) shows the relationship between $DIS_{threshold}$ and MAPE, and each curve with a constant value of σ. From the picture, it is clear to see with $DIS_{threshold}$ increasing, MAPE increases gradually. Note that these curves look like steps and in three variation ranges of $DIS_{threshold}$, i.e., 0.2-0.3, 0.4-0.5 and 0.7-1, MAPE is invariant.

Fig. 7(c) shows the relationship between program execution time and $DIS_{threshold}$. It can be observed there is a negative correlation between program execution time and $DIS_{threshold}$, especially when $DIS_{threshold}$ is from 0 to 0.1.

Fig. 7(d) shows the relationship between program execution time and σ. It can be observed there is a positive correlation between program execution time and σ. However, when σ is more than 1000, increase of program execution time is not obvious.

From Fig. 7(c) and 7(d), compared with the parameter $DIS_{threshold}$, the variance of σ slightly influences program execution time. To be specific, in this paper, the σ is set as 2000, because starting from $\sigma = 2000$, MAPE changes hardly from Fig. 7(a). In addition, program execution time even hardly changes from $\sigma = 2000$ in Fig. 7(d). As for the parameter $DIS_{threshold}$, it is adjusted to trade off the relationship between estimation accuracy and program execution time. In order to meet the real-time applications, program execution time should be less than data update cycle. In this paper, the GPS data of buses updates each 20 seconds, so the program execution time is set under 15 seconds to provide enough time to update related road information. Furthermore, considering the curves shape in Fig. 7(b), where MAPE hardly changes from $DIS_{threshold} = 0.7$ to $DIS_{threshold} = 1$, $DIS_{threshold}$ is set as 1 to obtain the shorter program execution time. When the traffic flow is stable, $DIS_{threshold}$ can be less than 1 to obtain more accurate road information and reduce the update frequency for road information.

2) MINING THE HEADING OF THE ROAD BY WALSM

These parameters used in the WALSM is ($\sigma = 2000$, $DIS_{threshold} = 1$). In Fig. 8, the projection of the original GPS points and the projections of these updated GPS points generated by each iteration are shown in Fig. 8(a)-8(e), respectively. From it, we can see that the GPS points on the road are gradually concentrated to the road line with iterations in Algorithm 1.

Then in Fig. 9 the proposed WALSM is compared with the LSM under three real datasets from different road segments. Fig. 9(1a)-9(3b) show these results in three different road

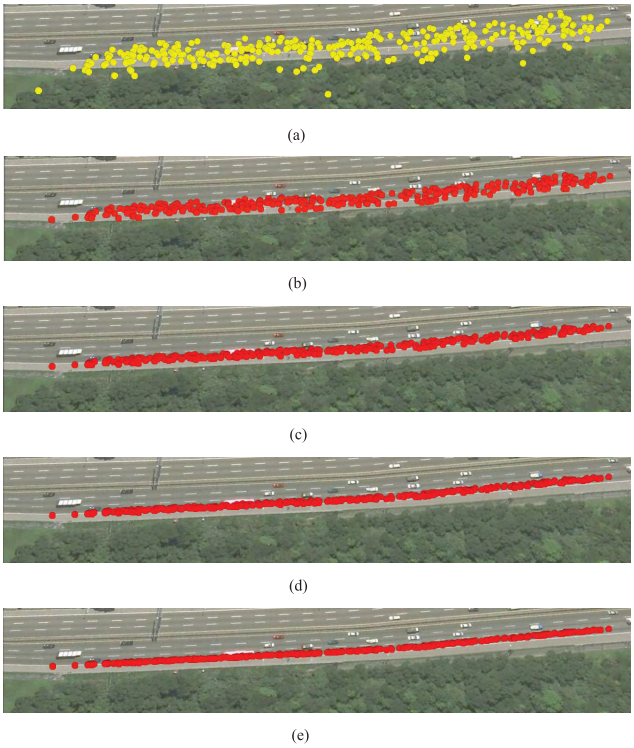


FIGURE 8. Projections of the original GPS points and the updated GPS points.

segments, and the part (a) and the part (b) in each subfigure represent the different unidirectional situations in a same bi-directional road. In each figure, the upper limit value on x-axis denotes the iteration times for the WALSM. The y-axis a_0 is the heading of the road. Note that the curve named pureLSM presents the result of the LSM and the curve named Real_Road presents the heading of the road centerline. We can see that results of the WALSM can achieve convergence, and it outperforms the LSM.

In Fig. 10 and 11, we show the impact of various values of σ and $DIS_{threshold}$ on the WALSM. GPS points collected from two road segments are chosen to discuss the problem, corresponding to Fig. 10 and 11 respectively. Besides the combination of parameters ($\sigma = 2000$, $DIS_{threshold} = 1$) (green), we add three curves ($\sigma = 2000$, $DIS_{threshold} = 0.01$) (cyan), ($\sigma = 100$, $DIS_{threshold} = 1$) (red) and ($\sigma = 100$, $DIS_{threshold} = 0.01$) (yellow). The upper limit value on x-axis of each curve denotes its iteration times, where all the GPS points meet the stop condition in the WALSM, i.e., distance L_i of each point is no more than $DIS_{threshold}$. For example, in Fig. 10, iterations of the WALSM ($\sigma = 2000$, $DIS_{threshold} = 1$, green curve) is 6 and it of the WALSM ($\sigma = 100$, $DIS_{threshold} = 0.01$, yellow curve) is 11.

By expressing pairwise comparisons, we discuss the impact as follows:

1) By comparing with these curves on the same value of σ, i.e., green and cyan, red and yellow, the variation of $DIS_{threshold}$ hardly changes the distance between these

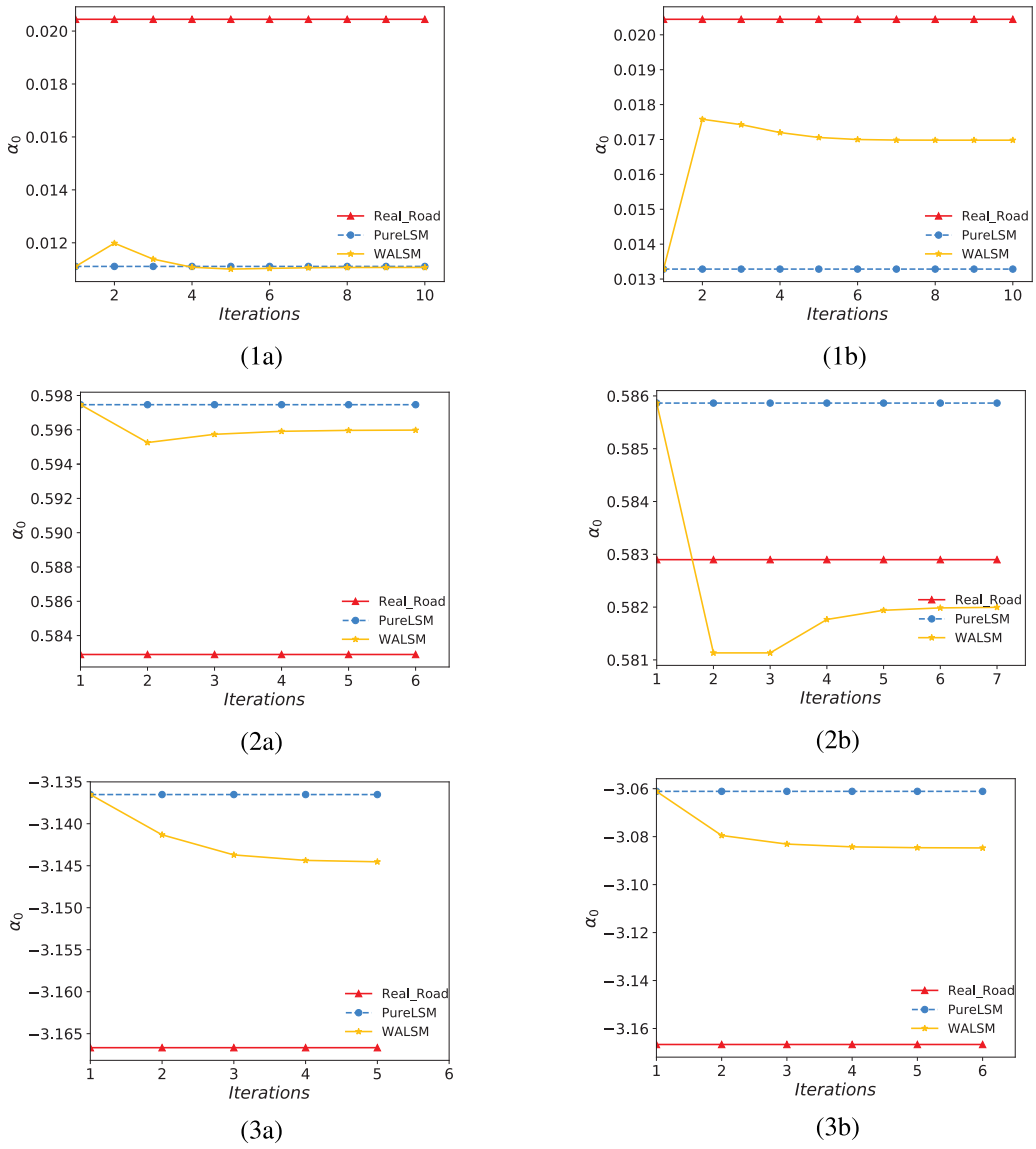


FIGURE 9. Comparison between the WALSM and the LSM (α_0 is the heading of the road).

curves and Real_Road. In fact, $DIS_{threshold}$ is used to control iteration time of the WALSM. When there are enough iterations, the accuracy of the WALSM will not change, even under different values of σ.

2) By comparing with these curves on the same value of $DIS_{threshold}$, i.e., green and red, cyan and yellow, the variation of σ changes the distance between curves and Real_Road obviously. From the two figures, increasing the value of σ can reduce the distance between curves and real road. It means σ easily impacts the accuracy of the WALSM. However, it hardly impacts iteration time of the WALSM.
3) Therefore, σ mainly impacts the accuracy of the WALSM and by combining with Fig. 7(a), when the σ is more than 2000, the impact becomes slight. Other parameter $DIS_{threshold}$ mainly impacts the iteration time which is relevant with program execution time.

In this paper, the proposed WALSM is one of main data-driven methods and can be used to mine the heading of the corresponding road from GPS data directly. It is a low cost and effective way. In addition, from simulations on real GPS data, the proposed WALSM can fit the heading of the corresponding road with accuracy about $10^{-1}\sim10^{-3}$ (or $1\sim10^{-1}$ degree), which meets the accuracy of traditional navigation applications about $10\sim10^{-1}$ degree in the paper [2].

3) ESTIMATING LANES NUMBER BY DV-RWDM

From the Distribution Variance - Road Width Discrete Model (DV-RWDM), road width can be estimated by (20). As we all know, the lane width is generally constant, such as 3.5m, so when the road contains 2 lanes, 3 lanes, 4 lanes or 5 lanes, the $W^2/4$ in (20) is 12.25,27.5625,49,76.5625 respectively. To verify that the DV-RWDM can be used to effectively

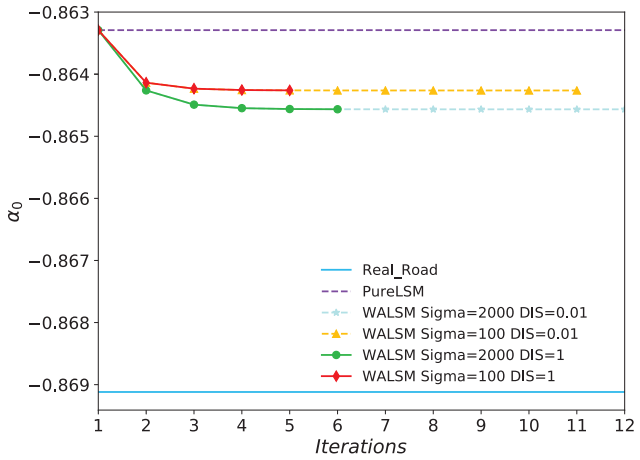


FIGURE 10. Impact of σ and $DIS_{threshold}$ on the WALSM (1).

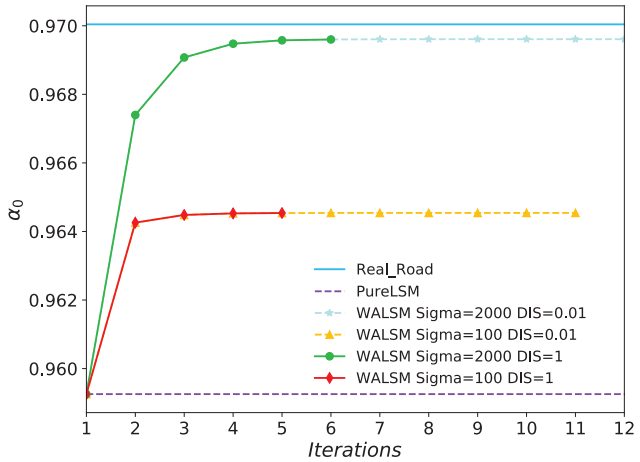


FIGURE 11. Impact of σ and $DIS_{threshold}$ on the WALSM (2).

mine the road width W and number of lanes, the real-world GPS points collected from three different road segments are input into this method. The corresponding results are shown in Table 4.

TABLE 4. The validation of the DV-RWDM.

$D(l_{left})$	$D(l_{right})$	$D(l_{road})$	$Width$	$L_{estimated}$	L_{actual}
17.75	23.01	42.71	9.45m	3	3
36.23	49.32	102.42	15.45m	4	4
57.87	72.65	132.86	16.45m	5	5

The estimated number of lanes $L_{estimated}$ and the actual number of lanes L_{actual} are shown in Table 4, where three different scenarios are discussed corresponding to 3, 4 and 5 lanes respectively. In the simulation, the road width W is estimated at first by (20). Then the width of the lane is set as constant 3.5m, and the estimated number of lanes should be 3, 4 and 5 lanes respectively. The estimated results coincide with the actual number of lanes of each road.

V. CONCLUSION

In this paper, by exploring the potential relationship between GPS points data and real roads structure, we successfully mine road information for raw GPS data. To be specific, we demonstrate the LSM can effectively estimate the real heading of road from raw GPS data. Aiming to improve the accuracy of the LSM, we further propose the WALSM to estimate the more accurate heading of the road under finite GPS point samples. In the WALSM, the weighted factor can be used to resist the negative impact from GPS noise during the estimation process. Simulation results based on real data show the WALSM outperforms the LSM in real world. Particularly, according to the discussion about the newly proposed WALSM, it can be seen that parameters σ and $DIS_{threshold}$ impact the accuracy and program execution time of this method, respectively. It means the performance of WALSM can be adjusted according to requirements on accuracy and timeliness for different applications. Furthermore, analysis of distribution of GPS points helps us infer the DV-RWDN to mine the width of the road from raw GPS data. Moreover, based on the constant width of the lane, we can estimate the number of lanes on the corresponding roads. The potential road information in raw GPS data can be utilized to promote development of ITS, such as updating electronic map, deploying RSUs and recognizing traffic bottleneck.

In the future work, combining with ad hoc network and intelligent sensors on vehicles or infrastructures, real positions of mobile devices can be collected. By calculating differences between their real positions and corresponding measured positions from GPS, system error in this position systems can be estimated. It means more accurate distribution of GPS points can be obtained to help determine positions of road segments from raw GPS data.

REFERENCES

[1] X. Chen, W. Shen, M. Dai, Z. Cao, J. Jin, and A. Kapoor, "Robust adaptive sliding-mode observer using RBF neural network for lithium-ion battery state of charge estimation in electric vehicles," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1936–1947, Apr. 2016.

[2] K. Gade, "The seven ways to find heading," *J. Navigat.*, vol. 69, no. 5, pp. 955–970, 2016.

[3] T. S. J. Darwish and K. A. Bakar, "Fog based intelligent transportation big data analytics in the internet of vehicles environment: Motivations, architecture, challenges, and critical issues," *IEEE Access*, vol. 6, pp. 15679–15701, 2018.

[4] H. Xie *et al.*, "A hybrid method combining Markov prediction and fuzzy classification for driving condition recognition," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10411–10424, Nov. 2018.

[5] R. S. Bali, N. Kumar, and J. J. P. C. Rodrigues, "Clustering in vehicular ad hoc networks: Taxonomy, challenges and solutions," *Veh. Commun.*, vol. 1, no. 3, pp. 134–152, Mar. 2014.

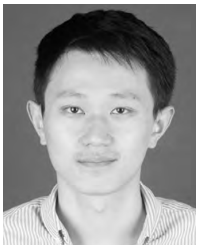
[6] S. Edelkamp and S. Schrodl, "Route planning and map inference with global positioning traces," *Computer Science in Perspective*. Berlin, Germany: Springer, 2003, pp. 128–151.

[7] S. Keykhaie and A. Mahmoudifar, "Study of connectivity in a vehicular ad hoc network with random node speed distribution," in *Proc. IEEE 6th NTMS*, Mar./Apr. 2014, pp. 1–4.

[8] E. R. Gelso and J. Sjoberg, "Consistent threat assessment in rear-end near-crashes using BTN and TTB metrics, road information and naturalistic traffic data," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 1, pp. 74–89, Jan. 2017.

[9] G. Liu and J. Zhang, "An energy management of plug-in hybrid electric vehicles based on driver behavior and road information," *J. Intell. Fuzzy Syst.*, vol. 33, no. 5, pp. 3009–3020, 2017.

[10] C. Chen, C. Lu, and Q. Huang, "City-scale map creation and updating using GPS collections," in *Proc. ACM 22nd SIGKDD*, 2016, pp. 1465–1474.
[11] T. Guo, K. Iwamura, and M. Koga, "Towards high accuracy road maps generation from massive GPS traces data," in *Proc. IEEE IGARSS*, Jul. 2007, pp. 667–670.
[12] J. Jung, S.-M. Lee, and H. Myung, "Indoor mobile robot localization and mapping based on ambient magnetic fields and aiding radio sources," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 7, pp. 1922–1934, Jul. 2015.
[13] P.-J. Tseng, C.-C. Hung, Y.-H. Chuang, K. Kao, W.-H. Chen, and C.-Y. Chiang, "Scaling the real-time traffic sensing with GPS equipped probe vehicles," in *Proc. IEEE 79th VTC Spring*, May 2014, pp. 1–5.
[14] C. Chen, C. Lee, and C. Lo, "Vehicle localization and velocity estimation based on mobile phone sensing," *IEEE Access*, vol. 4, pp. 813–817, 2016.
[15] (2015). *Here 360*. [Online]. Available: http://360.here.com
[16] (2016). *Openstreetmap*. [Online]. Available: http://openstreetmap.org
[17] S. Schroedl and K. Wagstaff, "Mining GPS traces for map refinement," *Data Min. Knowl. Discovery*, vol. 9, no. 1, pp. 59–87, 2004.
[18] W. Stewart and E. Nebot, "Automated process for generating digitised maps through GPS data compression," in *Proc. ACRA*, 2007, pp. 59–87.
[19] H. Li, L. Kulik, and K. Ramamohanarao, "Automatic generation and validation of road maps from GPS trajectory data sets," in *Proc. ACM 25th CIKM*, 2016, pp. 1523–1532.
[20] Y. Chen and J. Krumm, "Probabilistic modeling of traffic lanes from GPS traces," in *Proc. ACM 18th SIGSPATIAL*, 2010, pp. 81–88.
[21] L. Zhang, F. Thiemann, and M. Sester, "Integration of GPS traces with road map," in *Proc. IWCTS*, 2010, pp. 17–22.
[22] S. Miura, L.-T. Hsu, F. Chen, and S. Kamijo, "GPS error correction with pseudorange evaluation using three-dimensional maps," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3104–3115, Dec. 2015.
[23] L. Cao and J. Krumm, "From GPS traces to a routable road map," in *Proc. ACM 17th SIGSPATIAL*, 2009, pp. 3–12.



XUN ZHOU received the B.Eng. degree in communication engineering and the M.E. degree in electronics and communications engineering from Xidian University, Xi'an, China, in 2011 and 2018, respectively. His current research interests include intelligent transportation systems, machine learning, and evolutionary algorithms.

XUELIAN CAI received the Ph.D. degree in communications and information system from Xidian University, China, in 2013, where she is currently an Associate Professor with the School of Telecommunications Engineering. She was a Visiting Scholar with the University of Technology Sydney. Her research interests include vehicular networks, mobile ad hoc networks, and wireless sensor networks.

YUEHANG BU received the B.Eng. degree in communication engineering from the Hefei University of Technology, China, in 2017. He is currently pursuing the M.Sc. degree in communications and information system with Xidian University, Xi'an, China. His current research interests include traffic data mining and deep learning.

XI ZHENG received the bachelor's degree in computer information system from Fudan University, the master's degree in computer and information science from the University of New South Wales, and the Ph.D. degree in software engineering from The University of Texas at Austin. He is the Chief Solution Architect with Menulog, Australia. He is currently an Assistant Professor/Lecturer in software engineering with Macquarie University. He has specialized in service computing, the IoT security, and reliability analysis. He has over 40 high-quality publications in top journals and conferences (PerCOM, ICSE, ICCPS, the IEEE Systems Journal, *ACM Transactions on Embedded Computing Systems*. He has received the Best Paper Award from the Australian Distributed Computing and Doctoral Conference, in 2017. He has also received the Deakin Research outstanding Award, in 2016. He is a reviewer of top journals and conferences (IEEE Systems Journal, *ACM Transactions on Design Automation of Electronic Systems*, *Pervasive and Mobile Computing*, the IEEE Transaction on Cloud Computing, and PerCOM).

JIONG JIN (M'11) received the B.E. degree (Hons.) in computer engineering from Nanyang Technological University, Singapore, in 2006, and the Ph.D. degree in electrical and electronic engineering from the University of Melbourne, Australia, in 2011. He is currently a Senior Lecturer with the Faculty of Science, Engineering and Technology, School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC, Australia. Prior to joining the Swinburne University of Technology, he was a Research Fellow of the Department of Electrical and Electronic Engineering, University of Melbourne, from 2011 to 2013. His research interests include network design and optimization, fog and edge computing, robotics and automation, the Internet of Things, and cyber-physical systems and their applications in smart manufacturing and smart cities.

TOM H. LUAN received the B.Eng. degree from Xi'an Jiao Tong University, China, in 2004, the M.Phil. degree from The Hong Kong University of Science and Technology, in 2007, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2012. He is currently a Professor with the School of Cyber Engineering, Xidian University, Xi'an, China. He has authored/co-authored over 40 journal papers and 30 technical papers in conference proceedings. He holds one U.S. patent. His research interests include content distribution and media streaming in vehicular ad hoc networks, and peer-to-peer networking, and the protocol design and performance evaluation of wireless cloud computing and edge computing. He has served as a TPC member of the IEEE GLOBECOM, ICC, PIMRC. He is a Technical Reviewer of multiple IEEE Transactions including TMC, TPDS, TVT, TWC, and ITS.

CHANGLE LI (M'09–SM'16) received the Ph.D. degree in communications and information system from Xidian University, China, in 2005, where he is currently a Professor with the State Key Laboratory of Integrated Services Networks. He has conducted a Postdoctoral research in Canada and the National Institute of information and Communications Technology, Japan. He was a Visiting Scholar with the University of Technology Sydney. His research interests include intelligent transportation systems, vehicular networks, mobile ad hoc networks, and wireless sensor networks.