2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Bus Trajectory Identification by Map-Matching

Rudy Raymond
IBM Research – Tokyo
19-21 Nihonbashi Hakozaki-cho,
Chuo-ku, Tokyo, 103-8510, Japan.
Email: rudyhar@jp.ibm.com

Takashi Imamichi
IBM Research – Brazil
Av. Pasteur 138/146, Botafogo,
Rio de Janeiro, 22290-240, Brazil.
Email: tima@br.ibm.com

***Abstract*—We study the problem of identifying vehicle trajectories from the sequences of noisy geospatial-temporal datasets. Nowadays we witness the accumulation of vehicle trajectory datasets in the form of the sequences of GPS points. However, in many cases the sequences of GPS points are sparse and noisy so that identifying the actual trajectories of vehicles is hard. Although there are many advanced map-matching techniques claiming to achieve high accuracy to deal with the problem, only few public datasets that come with ground truth trajectories for supporting the claims. On the other hand, some cities are releasing their bus datasets for real-time monitoring and analytics. Since buses are expected to run on predefined routes, such datasets are highly valuable for map-matching and other pattern recognition applications. Nevertheless, some buses in reality appear not following their predefined routes and behave anomalously. We propose a simple and robust technique based on the combination of map-matching, bag-of-roads, and dimensionality reduction for their route identification. Experiments on datasets of buses in the city of Rio de Janeiro confirm the high accuracy of our method.**

I. Introduction

We have been witnessing the surging amount of geospatial-temporal datasets gathered from various devices, such as, cell phones and vehicles. Although the usage of such datasets with personal information is often limited due to privacy concerns, some cities such as Dublin[1] and Rio de Janeiro have recently released the Global Positioning Systems (GPS) trajectories of their bus fleets for optimizing their daily operations and other analytics applications. Because such trajectories are often labeled (buses usually follow predefined routes and stop at fixed bus stops that are also released for public), they are useful for Artificial Intelligent (AI) algorithms that require a large amount of training datasets, such as, neural networks.

However, the usage of GPS trajectories of bus fleets is limited due to several practical reasons. First, the GPS datasets are noisy and sparse (collected in every several seconds up to several minutes) so that sequences of roads traversed by the vehicles cannot be obtained by simply connecting consecutive GPS points. Fig. 1 depicts an example of a GPS sequence on a road network with its corresponding route. We can see many alternative routes that can explain the GPS sequence under some assumption of noise distribution. Secondly, in reality buses in cities like Rio de Janeiro sometimes behave anomalously from their predefined routes, for example, they skip or visit specific regions, as shown in Fig. 2. This can be caused by special events like festivals or unexpected traffic situations like congestion or road closures. Thirdly, information of official bus routes and bus stops is not always up-to-date and therefore, does not reflect the latest routes, as we also confirmed in Rio de Janeiro. On the other hand, there is a significant percentage of buses that do not report their official bus routes probably due to device failures.

[1]https://data.dublinked.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

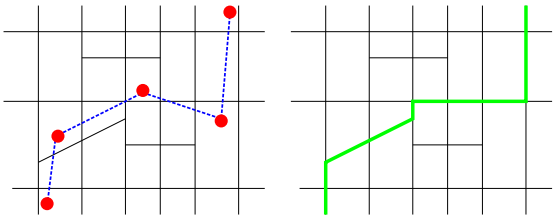


Fig. 1. Left: A road network (black) and a GPS sequence (red). Right: The route corresponding to the GPS sequence in the left (green).

To tackle anomalies of bus trajectories and information of bus routes, we propose a simple but robust and accurate machine learning method incorporating map-matching algorithms. We transform input GPS sequences into sequences of traversed road IDs by map-matching and then treat the traversed road IDs as *bag of features*. Notice that the bag-of-words model is a popular technique in image classification, e.g., [1], [2]. Map-matchings transform a sequence of high-dimensional GPS points into a sequence of roads with lower dimensionality because there are less number of roads traversed by vehicles than their GPS points. We also observe that the number of predefined bus routes is significantly smaller than the total number of roads by several order of magnitude, and thus provides opportunities to exploit dimensionality reduction techniques. We show that applying dimensionality reduction on outputs of map-matchings prior to classification can help achieve both high accuracy and robustness in identifying routes as confirmed with experiments over bus datasets of Rio de Janeiro.

The results presented in this paper provide two interesting implications. First, a recent work [3], that describes similar challenges in working with noisy and anomalous GPS trajectories of buses in Rio de Janeiro, proposes the use of Convolutional Neural Network (CNN) to automatically detect outliers and identify bus routes with high accuracy. They also

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

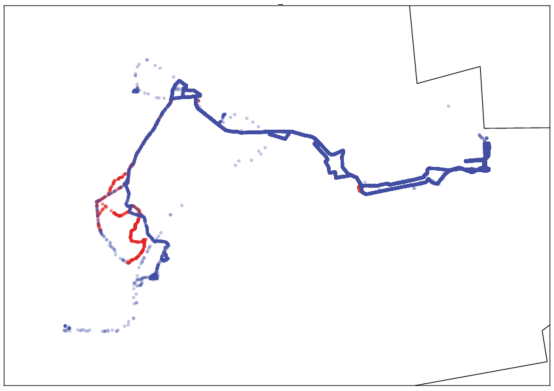


Fig. 2. An example of trajectories of GPS sequences (blue dots) of buses serving route 353 in Rio de Janeiro and their supposed route (red dots). The total length of the trajectories is about 40 kilometers. Notice that most of the time the buses follow the predefined route (as red dots are overlapped by blue dots), but on the lower left we can see parts of the route not traversed (dense red dots), parts of possible anomalous GPS sequences (blurred blue dots). The black line segments inside the rectangle frame represent the coast line.

argue that the usage of simple techniques such as nearest-neighbor classification cannot deal with raw GPS trajectories. On the contrary, we show that map-matching combined with a simple bag-of-words model can achieve up to the same level of high accuracy. Moreover, this is realized with significantly less computational efforts and the results are easier to interpret.

The second implication is in showing the potential of using public buses GPS trajectories for evaluating geospatial-temporal algorithms in large-scale and real-time. Despite the ease of gathering data from various sensors, it seems that there are not many vehicular GPS datasets that come with ground-truth labels available for public. Many proposed geospatial-temporal algorithms like map-matchings were often based on small labeled datasets that are produced with expensive manual labor of labeling. We show that it is possible to process and filter GPS trajectories and evaluate their similarities to official routes for comparison of accuracy levels. We also plan to release tools and datasets based on publicly available bus datasets of Rio de Janeiro in the near future.

A. Related Work

Massive amount of geospatial-temporal sequences including GPS sequences have become popular datasets for various machine learning and pattern recognition algorithms. They can be utilized for ecological research, e.g., tracking movement patterns of wild animals [4], for tracking people behavior [5], and vehicles that are becoming much more important in the context of Simultaneuos Localizations and Mapping (SLAM) [6]. Some of the obstacles in using those datasets include the high noise and irregularity of such GPS sequences. Various techniques have been developed to preprocess them, see e.g. [7] for a survey. Sophisticated techniques to map noisy and sparse GPS sequences of vehicles into road sequences are map-matchings. See e.g. [8] for a survey, where Hidden Markov Model (HMM) is one of the most active fields [9]–[13].

Approaches based on Deep Learning (DL) have also been used for trajectory identification, such as, an end-to-end approach using raw sequences of GPS [3] for anomaly detection and route identification of buses in Rio de Janeiro, which is very similar to our work. However, [3] only utilizes a small fraction of GPS sequences (20 for each bus) and landmarks (bus stops) to identify routes and detect anomaly. The main purpose is to help bus operators by visualizing anomalous buses. We believe map-matching can improve the visualization by providing the comparison of their trajectories instead of sequences of GPS points. Additionally, map-matching can be used to uncover attributes of trajectories that are important for daily traffic management and analytics, e.g., to predict travel time [14] and simulate traffic policies [15]. We also believe that map-matching can be used for encoding geospatial-temporal patterns, similar to [16] that proposes Markov Transition Fields, to enable the use of feature learning and classification in DL.

There have been many representative datasets used to evaluate geospatial-temporal algorithms, such as, *T-drive* [17], [18], the *Seattle datasets* [9], and undisclosed datasets from service providers [19]. The Seattle datasets were designed for accuracy testing in mind with clear ground truth but only contain sequences from one car traversing ad-hoc trajectories. On the other hand, the T-drive datasets are popular for testing scalability due to their large scale but no ground truth labels are not provided. [20] used GPS sequences of several taxis in Shanghai but as the T-drive, there are no ground truth labels. [20] recovered the labels of 70 taxis by manually checking GPS sequences and matching them to roads all performed by two experienced volunteers. [21] tested its algorithms by monitoring 85 volunteers. The closest to ours is [22] that used 4 bus routes in Singapore as ground truth for testing an online map-matching algorithm.

II. Notation and Problem Setting

We describe notations and problem settings. For ease of explanation, we directly model the problem and notation to work with bus trajectories, but it is easy to see that they can be used in the broader context of vehicle trajectory identification.

We consider a set B of buses such that for each bus $b \in B$ its sequence $\mathbf{g}_b = (g_{b,1}, \ldots, g_{b,p_b})$ of GPS points is given, where $g_{b,i}$ consists of the latitude, longitude, and time stamp of i-th GPS point of bus b and the sequence is ordered in the increasing order of the time stamps. A directed road network $G = (V, A)$ is also given where $V = \{v_i \mid i = 1, \ldots, n\}$ and $A = \{a_i \mid i = 1, \ldots, m\}$ denote cross points and road segments, respectively.

Unique to the bus trajectory identification is that each bus is assigned to one or more predefined routes. Let R be a set of the predefined routes such that $\mathbf{r}_i = (r_{i,1}, \ldots, r_{i,q_i})$ represents the sequence of coordinates for each route $i \in R$. Note that there are ℓ predefined routes, i.e., $\ell = |R|$.

We consider two types of classification problems for trajectory identification. The first is to identify the predefined route $i \in R$ of each bus $b \in B$ given its sequence $\mathbf{g}_b$ of GPS points, the directed road network $G(V, A)$, and the set R of predefined routes. We assume that all the bus do not have any labels of the predefined routes. In this problem setting, we assume that all buses strictly follow their assigned routes, and therefore their trajectories can be recovered from the sequence of roads listed in the definition of the route. We call this type of unsupervised classification problem `P1`. It is similar to clustering problems where predefined routes denote the true centers of clusters to evaluate the accuracy of clustering algorithms.

The second is to identify the predefined route of each bus where predefined routes are not reliable; we found the situation in Rio de Janeiro. In this case, we rely on the observation that there should be a group of buses serving the same route, and therefore, we predict the trajectory of an unknown bus by comparing its trajectory to those of buses with known routes. Formally, together with the road network and the GPS sequences of all buses, we are also given the label $l_{b'}$ for $b' \in B'$, where $B' \subset B$ is a subset of the buses and $l_{b'} \in R$ denotes the ID of the predefined route served by bus b'. The task is to identify the label l_b for all $b \in B \setminus B'$. We call this type of supervised classification problem `P2`.

III. The Proposed Trajectory Identification

There are several difficult obstacles to determine the route of bus b directly from its GPS sequence $\mathbf{g}_b$. First, the lengths of GPS sequences vary greatly among buses, and so do the intervals between two consecutive GPS points in the sequence. Moreover, some fraction of GPS points might be corrupted by large amount of noises that make it even more difficult for end-to-end predictors. [3] overcome these obstacles by considering only 20 GPS points of each bus, and therefore might fail to capture anomalies in trajectories if the sampled points are not representative.

Our method transforms the input GPS sequences of both the buses and the predefined routes in the same way. For simplicity, we explain the case for a GPS sequence $\mathbf{g}_b$ of bus b. We first apply map-matching to $\mathbf{g}_b$ to obtain sequences $\mathbf{g}_b^{\text{mm}}$ of road segments. See the next section for more details of map-matching. We then treat the the resulting sequences as *bag of roads*, similarly as in the bag-of-words model for document classification, to classify trajectories with various distances and similarity measures. The bag-of-roads vector $\mathbf{g}_b^{\text{br}}$ is a sparse vector of frequency counts of road segments traversed by bus b, where i-th element $g_{b,i}^{\text{br}}$ of $\mathbf{g}_b^{\text{br}}$ denotes the frequency of bus b traversing road segment $a_i \in A$. Similarly we transform the sequence $\mathbf{r}_j$ of coordinates of predefined route $j \in R$ into a sequence $\mathbf{r}_j^{\text{mm}}$ of roads by map-matching and generate a bag-of-road vector $\mathbf{r}_j^{\text{br}}$. We can easily compare the bag-of-roads vectors of the buses B and the predefined routes R because the vectors have the same dimension $m = |A|$. We further employ a simple dimensionality reduction technique to the bag-of-roads vectors for even faster and accurate trajectory identification. See Table I for the overview of the transformation of data.

A. Map-Matching of GPS Sequences

The GPS data may contain the measurement error and the interval of measurement may be long and irregular. We adopt the state-of-the-art map-matching algorithm with the HMM in [12] because it is known to be robust to such cases. The HMM-based map-matching generates hidden states (road segments) around GPS points and looks for the sequence of hidden states with the highest likelihood, which consists of the initial, the emission, and the transition probabilities.

We basically follow the implementation in [12], [13], but simplify the transition probability as follows

$$\frac{1}{\beta} \exp\left(-\frac{d^*(s, s')}{\beta}\right),$$

where β is a parameter and $d^*(s, s')$ denotes the shortest path distance from a midpoint s of a road segment to another s' on road network $G(V, A)$. Intuitively, two consecutive GPS points will likely be matched to road segments having shortest driving distance. We verified the simplification does not make the precision worse through preliminary experiments. Throughout the experiments in this paper, we configure the map-matching parameters by adjusting parameters in [12] appropriately with $w_{\text{turn}} = 0$ (the turn cost in [12]) of the transition probability.

The output of map-matching of $\mathbf{g}_b$ is a sequence $\mathbf{g}_b^{\text{mm}}$ of connected roads that best explained the GPS sequence. Since the number of roads is at most the number of road segments s in the road network $G(V, A)$, we can consider the map-matching as a method to reduce the dimension of the vector of GPS sequence.

B. Route Similarity and Comparison

We measure how close the bag-of-roads vectors of the buses to those of the predefined routes for the problem setting `P1`. For example, we can compute the L_p distance, for $p = 1, 2, \infty$, between $\mathbf{g}_b^{\text{br}}$ and $\mathbf{r}_j^{\text{br}}$, and output the prediction of the route of bus b by

$$\arg\min_{j \in R} \|\mathbf{g}_b^{\text{br}} - \mathbf{r}_j^{\text{br}}\|_p. \tag{1}$$

Another way is to compute the cosine similarity between bus b to each $j \in R$. Namely, The prediction for the route of bus b is

$$\arg\max_{j \in R} \frac{\mathbf{g}_b^{\text{br}} \cdot \mathbf{r}_j^{\text{br}}}{\|\mathbf{g}_b^{\text{br}}\| \|\mathbf{r}_j^{\text{br}}\|}. \tag{2}$$

The similarity and distance computation in Eqs. (1) and (2) has a problem because the dimension $m = |A|$ of the bag-of-roads vectors $\mathbf{g}_b^{\text{br}}$ and $\mathbf{r}_j^{\text{br}}$ is still high. However, we observe that $|R| = \ell \ll |A| = m$ holds, e.g., m is more than 174K in Rio de Janeiro, while ℓ is at most 500. Thus, we first project the bag-of-roads vectors into the subspace spanned by the bag-of-roads vectors of the predefined routes and then compute the similarity and distance.

Let $\mathbf{R} = (\mathbf{r}_1^{\text{br}}, \ldots, \mathbf{r}_k^{\text{br}})$ be a matrix of dimension $m \times \ell$ whose i-th column is the bag-of-road vector $\mathbf{r}_i^{\text{br}}$ of $i \in R$.

TABLE I
OVERVIEW OF THE TRANSFORMATION OF DATA.

Input	GPS sequence (seq. of coordinates)		Map-matching (seq. of roads)		Bag-of-roads (vec. of freq., dim = m)		Dimensionality reduction (vec. of freq., dim = $d \ll m$)
Bus GPS data	$\mathbf{g}_b$	$\Rightarrow$	$\mathbf{g}_b^{\text{mm}}$	$\Rightarrow$	$\mathbf{g}_b^{\text{br}}$	$\Rightarrow$	$\tilde{\mathbf{g}}_b^{\text{br}}$
Predefined routes	$\mathbf{r}_j$	$\Rightarrow$	$\mathbf{r}_j^{\text{mm}}$	$\Rightarrow$	$\mathbf{r}_j^{\text{br}}$	$\Rightarrow$	$\tilde{\mathbf{r}}_j^{\text{br}}$

In this case, since the rank of $\mathbf{R}$ is at most ℓ, we can apply dimensionality reduction as follows. By Singular Value Decomposition (SVD), we can find a decomposition of $\mathbf{R}$ such that

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U}$ is a $m \times m$ matrix whose columns are left-singular vectors of $\mathbf{R}$, $\mathbf{V}$ is a $\ell \times \ell$ matrix whose columns are right-singular vectors of $\mathbf{R}$, and $\mathbf{\Sigma}$ is a $m \times \ell$ rectangular diagonal matrix that has at most k non-zero elements sorted in the non-increasing order.

Letting $\mathbf{U}_d$ be the top-d (where $1 \leq d \leq \ell$) left singular vectors in the above SVD of $\mathbf{R}$, we compute the projection of $\mathbf{g}_b^{\text{br}}$ into the subspace spanned by $\mathbf{U}_d$ to obtain $\tilde{\mathbf{g}}_b^{\text{br}} = \mathbf{U}_d^T \mathbf{g}_b^{\text{br}}$ for each $b \in B$. Similarly, we also compute the projection of $\mathbf{r}_j^{\text{br}}$ to obtain $\tilde{\mathbf{r}}_j^{\text{br}} = \mathbf{U}_d^T \mathbf{r}_j^{\text{br}}$ for $j \in R$. Notice that both $\tilde{\mathbf{g}}_b^{\text{br}} \in \mathbb{R}^d$ and $\tilde{\mathbf{r}}_j^{\text{br}} \in \mathbb{R}^d$ hold, and therefore, we can compute the prediction of the route in Eqs. (1) and (2) faster by substituting $\mathbf{g}_b^{\text{br}}$ and $\mathbf{r}_j^{\text{br}}$ with $\tilde{\mathbf{g}}_b^{\text{br}}$ and $\tilde{\mathbf{r}}_j^{\text{br}}$, respectively, because $d \leq \ell \ll m$.

To deal with the problem setting `P2`, we compute nearest neighbors of an unlabeled bus among the labeled buses by using the bag-of-roads vectors. Namely, for a set B' of buses whose predefined routes are known, we predict the label of $b \in B \setminus B'$ by L_p distance

$$\arg\min_{b' \in B'} \|\tilde{\mathbf{g}}_b^{\text{br}} - \tilde{\mathbf{g}}_{b'}^{\text{br}}\|_p. \tag{3}$$

We call it *supervised nearest neighbor*. Notice that it can be performed efficiently with data structures for nearest-neighbor computations, such as, the Ball Tree or the KD Tree.

IV. EXPERIMENTS

A. Datasets

We explain the details of the datasets we use for the experiments. For the road network, we use the *OpenStreetMap (OSM)* [23] that is also available for public. The numbers of cross points and road segments are 85872 and 174323, respectively, in the area we extracted. The GPS sequences and predefined routes of buses in Rio de Janeiro are available for public too. There are several types of data as follows.

Bus GPS data[2]: Real-time GPS data of buses. The data consists of records of time, vehicle number, bus line number, coordinates, velocity and direction. We use the time and coordinates for map-matching, and the bus line number (when available) for testing accuracy of trajectory identification.

We collected the bus GPS data every minute from 00:00 on Feb 15, 2016 to 12:14 on Feb 17, 2016. The data contains 6963 buses, 450 bus lines and 12.8 million records in total. Around 20% of records are lack of the bus line number (the entries are left empty). Note that we are not sure whether the records without bus line number were out of service or just missing data, but we suspect this is a typical operational problem. See Table II for the top 5 bus lines including the records without bus line numbers.

TABLE II
TOP 5 BUS LINES IN THE BUS GPS DATA. RECORDS WITHOUT BUS LINE NUMBER ARE COUNTED AS 'EMPTY.'

bus line	count	proportion
empty	2574509	20.11 %
864	129332	1.01 %
371	96931	0.76 %
803	92741	0.72 %
908	89089	0.70 %
others	9821611	76.70 %
total	12804213	100.00 %

Old bus route data[3]: The trajectories of the predefined bus line routes last updated in April 4, 2014. It is represented by sequence of coordinates. There are 489 bus lines in the data, but because the city has modified the bus lines and the old bus route data is not necessarily up-to-date.

New bus route data[4]: Various data about the predefined bus lines in the same sequence-of-coordinates format to update the previous old bus route data on January 29, 2016. There are 375 bus lines included. The datasets also contain bus-stop data[5] that we do not use.

We apply map-matching to the bus GPS data, old and new bus route data to transform them into sequences of roads.

B. Results

We performed experiments on the problem `P1` and `P2` of identifying trajectory of buses to measure accuracy.

In the unsupervised problem `P1`, for each $b \in B$ we are given its sequence of GPS points $\mathbf{g}_b$ (from the bus GPS data), the road network $G(V, A)$ (from the OSM), and the set of predefined routes R, and its corresponding bag-of-roads vectors $\mathbf{R}$ (from the new bus route data). We identify the trajectory of b by computing the *Euclid* (L_2) distance and the *Cosine* similarity between the bag-of-roads vector $\mathbf{g}_b^{\text{br}}$ with each bag-of-roads vector $\mathbf{r}_j^{\text{br}}$ for $j \in R$, as in Eqs. (1) and (2). The result is presented in Table III where the accuracy by taking the top-k nearest predefined routes according to the Euclid distance and Cosine similarity are shown. We can observe that Cosine achieve higher accuracy than Euclid,

[2]http://data.rio/dataset/gps-de-onibus

[3]http://data.rio/dataset/pontos-dos-percursos-de-onibus

[4]http://data.rio/dataset/onibus-gtfs

[5]http://data.rio/dataset/pontos-de-parada-de-onibus

TABLE III
TOP-k UNSUPERVISED NEAREST NEIGHBOR PREDICTION ACCURACY OF PROBLEM P1

Distance Measure	k = 1	2	4	8	16	32
Euclid	0.75	0.79	0.81	0.82	0.82	0.82
Cosine	0.82	0.88	0.91	0.94	0.96	0.96

TABLE IV
TOP-1 NEAREST NEIGHBOR ACCURACY AND DIMENSION d OF BAG-OF-ROADS VECTORS

Method	Dimension d = 4	8	16	32	64	128	256	375
Unsup	0.03	0.05	0.14	0.28	0.57	0.76	0.82	0.82
Sup	0.72	0.86	0.90	0.92	0.93	0.95	0.95	0.95

which reaffirms the fitness of cosine similarities for bag-of-words models [24]. When only the top-1 nearest neighbor route is used, the accuracy of Cosine is 0.82, while that of Euclid is 0.75. We vary $k = 2, \ldots, 32$, and as the result the accuracy of Cosine and Euclid saturate at 0.96 and 0.82, respectively.

Table IV shows how the dimension d of bag-of-roads vectors relates to the accuracy of trajectory identification for both unsupervised and supervised prediction. At the row of "Unsup" (Unsupervised), which corresponds to the Cosine row in Table III, when we vary $d = 4, \ldots, 256$, the accuracy becomes 0.82 when $d = 256$, where the optimal accuracy is 0.82 at $d = 375$ as in Table III (the size of R is 375). The row of "Sup" (Supervised) in Table IV shows the accuracy of route identification from other buses with known routes, i.e., for Problem P2, with the supervised nearest neighbor as in Eq. (3). We evaluate the accuracy by 10-fold cross validation: we randomly partition the data into 10 parts, use 9 of them as known routes and the rest for testing. We find that the average accuracy of "Sup" is 0.95 when dimensionality reduction is not applied, which is much higher than that obtained in Problem P1. Similar to the row of "Unsup" in the table, we test how the accuracy changes as the dimension $d = 4, \ldots, 256$. We observe that the accuracy is ≥ 0.9 when $d \geq 16$, which is much larger than "Unsup".

In the supervised problem P2, R is not reliable so that Eqs. (1) and (2) are not appropriate for trajectory identification. However, we can utilize labels from other known buses because buses serving a particular route will usually follow similar trajectories. In this case, we can rely Euclidean distance measures of supervised nearest neighbor to identify their trajectories.

We also evaluate the supervised nearest neighbor based on the old bus route data, where let $\mathbf{R}_{\text{old}}$ be a matrix of bag-of-roads vectors of the predefined routes of the old route data. From the SVD of $\mathbf{R}_{\text{old}}$, we compute the projection $\tilde{\mathbf{g}}_b^{\text{br}}$ and apply Eq. (3) to identify b's route. As we did in Table IV based on the new bus route data, we evaluate the effect of the dimensionality reduction of $\mathbf{R}_{\text{old}}$ and observe that it does not affect accuracy by much: The accuracy of "Sup" when no dimensionality reduction is performed is still 0.95, and the accuracies of "Sup" when $d = 4, 8$ are $0.79, 0.87$, respectively, and when $d \geq 16$ the accuracy is ≥ 0.91, which are not that different from those in the row "Sup" of Table IV.

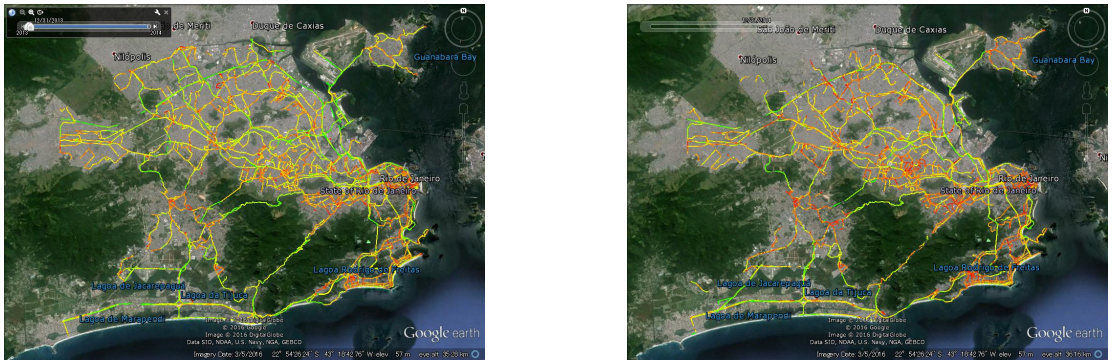
C. Discussion

We compare our approach with [3] that copes with the same problem with ours and use the same source of datasets. [3] reported that the best prediction model based on Convolutional Neural Network (CNN) had accuracy of 94.9%, which is close to the accuracy of our simple and robust Sup-NN. The CNN was trained with more than 20GB of bus data set, while our supervised nearest neighbor is built upon much less data set despite its robustness and high accuracy. Interestingly, [3] has tried to implement K-nearest neighbors and K-means using bus stops as bag-of-features but failed to achieve high accuracy (exact numbers were not reported). Notice also that despite based on the bag-of-word models that ignore the sequential information of roads in the trajectories, our proposed method achieved quite high accuracy. We have not explored the reasons in depth but we believe it is because map-matching algorithms already take into account sequential information in their outputs. We tried to construct feature vectors by bag-of-words of consecutive road IDs (for partially imposing sequential information), but the accuracy did not improve and often decreased due to high dimensionality problems.

We can use matched roads to detect spatial and temporal outliers like [3], but map-matching also has important complementary applications to analyze the urban traffic condition. Once we obtain the correspondence between GPS points and road segments, we can infer features of the road segments, such as, the mean velocities of vehicles traversing the road segments. Figures 3 (a) and (b) (created with Google Earth™) are the heatmaps of mean velocity of road segments in Rio de Janeiro at 10:00 and 18:00 on Feb 16, 2016, from map-matching of the bus GPS data. For each GPS point of a bus and its matched road segment, we assign its velocity to the road segment. If a road segment has at least five velocity values, we calculate their mean value as *mean velocity*. Line segments with red color imply that the mean velocity is around 0, i.e., traffic congestion, those with green color imply those with mean velocity over 50 km/h, i.e., free-flow traffic, and those with yellow color are around 20–30 km/h. We can observe different traffic patterns in the city depending on the time. For example, one can notice that there were significant bus trajectories (and delays) in the morning around Nilopolis and Sao Joao de Meriti (upper middle of the figures), but not in the afternoon.

V. CONCLUDING REMARKS

We showed a robust and accurate bus trajectory identification based on the combination of map-matching and bag-of-words model. Our choice of map-matching is the one based on Hidden Markov Models because they are proven effective for noisy and sparse GPS sequences. However, one should be able to utilize other types of map-matchings and test their accuracy



(a) 10:00–10:30 (b) 18:00–18:30

Fig. 3. Heatmaps of mean velocity of buses on Feb 16, 2016 in two different periods: in the morning (a) and in the afternoon (b).

on the datasets. We showed that even though the route data may not be up-to-date, trajectory identification is still possible and can be performed accurately.

Finally, apart from the purpose of generic testing for geospatial-temporal algorithms, the datasets themselves are important for practical reasons. For example, there have been many studies that focus on the punctuality of bus operations [14], [25]. As cities like Rio de Janeiro have started releasing datasets of their bus fleets, we believe they are ideal for both scalability and accuracy testings: they are abundant, available almost real-time, and come with ground truth labels.

Acknowledgment

A part of this research was supported by CREST, JST.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. of ECCV Int. Workshop on Statistical Learning in Computer Vision*, no. 2004/010, 2004, pp. 1–22.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[3] A. Bessa, F. de Mesentier Silva, R. F. Nogueira, E. Bertini, and J. Freire, "RioBusData: Outlier detection in bus routes of Rio de Janeiro," 2015, arXiv:1601.06128 [cs.HC].

[4] J.-G. Lee, J. Han, X. Li, and H. Gonzalez, "TraClass: Trajectory classification using hierarchical region-based and trajectory-based clustering," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1081–1094, 2008.

[5] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proc. of the 18th Int. Conf. on World Wide Web*, 2009, pp. 791–800.

[6] J. D. Carlson, "Mapping large, urban environments with GPS-aided SLAM," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, 2010.

[7] Y. Zheng and X. Zhou, *Computing with Spatial Trajectories*, 1st ed. Springer-Verlag New York, 2011.

[8] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intelligent Systems and Technology*, vol. 6, no. 3, p. Article No. 29, 2015.

[9] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proc. of 17th ACM SIGSPATIAL Int. Conf. Adv. in Geographic Information Systems*, 2009, pp. 336–343.

[10] G. Wang and R. Zimmermann, "Eddy: An error-bounded delay-bounded real-time map matching algorithm using HMM and online Viterbi decoder," in *Proc. of 22nd ACM SIGSPATIAL Int. Conf. on Adv. in Geographic Information Systems*, 2014, pp. 33–42.

[11] R. Raymond, T. Morimura, T. Osogami, and N. Hirosue, "Map matching with hidden Markov model on sampled road network," in *Proc. of 21st Int. Conf. Pattern Recognition*, 2012, pp. 2242–2245.

[12] T. Osogami and R. Raymond, "Map matching with inverse reinforcement learning," in *Proc. of 23rd Int. Joint Conf. Artificial Intelligence*, 2013, pp. 2547–2553.

[13] T. Imamichi, T. Osogami, and R. Raymond, "Truncating shortest path search for efficient map-matching," in *Proc. of 25th Int. Joint Conf. Artificial Intelligence*, 2016, pp. 589–595.

[14] M. Kormáksson, L. Barbosa, M. R. Vieira, and B. Zadrozny, "Bus travel time predictions using additive models," in *Proc. of 2014 IEEE Int. Conf. on Data Mining*. IEEE, 2014, pp. 875–880.

[15] T. Osogami, T. Imamichi, H. Mizuta, T. Suzumura, and T. Idé, "Toward simulating entire cities with behavioral models of traffic," *IBM Journal of Research and Development*, vol. 57, no. 5, pp. 6:1–6:10, 2013.

[16] Z. Wang and T. Oates, "Spatially encoding temporal correlations to classify temporal data using convolutional neural networks," 2015, arXiv:1509.07481 [cs.LG].

[17] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proc. of 18th SIGSPATIAL Int. Conf. Adv. in Geographic Information Systems*, 2010, pp. 99–108.

[18] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. of 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2011, pp. 316–324.

[19] Y. Li, Q. Huang, M. Kerber, L. Zhang, and L. Guibas, "Large-scale joint map matching of GPS traces," in *Proc. of the 21st ACM SIGSPATIAL Int. Conf. on Adv. in Geographic Information Systems*, 2013, pp. 214–223.

[20] J. Yang and L. Meng, *Progress in Location-Based Services 2014*. Springer International Publishing, 2015, ch. Feature Selection in Conditional Random Fields for Map Matching of GPS Trajectories, pp. 121–135.

[21] D. Delling, A. V. Goldberg, M. Goldszmidt, J. Krumm, K. Talwar, and R. F. Werneck, "Navigation made personal: Inferring driving preferences from GPS traces," in *Proc. of the 23rd SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, no. 31, 2015, pp. 31:1–31:9.

[22] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet, "Online map-matching based on hidden Markov model for real-time traffic sensing applications," in *Proc. of 15th Int. IEEE Conf. Intelligent Transportation Systems*, 2012, pp. 776–781.

[23] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[24] J. Choi, H. Cho, J. Kwac, and L. S. Davis, "Toward sparse coding on cosine distance," in *Proc. 22nd Int. Conf. on Pattern Recognition*, 2014, pp. 4423–4428.

[25] E. I. Diab and A. M. El-Geneidy, "Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability," *Public Transport*, vol. 4, no. 3, pp. 209–231, 2013.