Travel Modes Recognition Method Based on Mobile Phone Signaling Data

Ying Xia[1], Jie Tang[1(✉)], Xu Zhang[1], and Hae-young Bae[2]

[1] School of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing, China
{xiaying, zhangx}@cqupt.edu.cn,
S160201056@stu.cqupt.edu.cn
[2] Department of Computer Engineering, Inha University, Incheon, South Korea
hybae@inha.ac.kr

Abstract. With the acceleration of urbanization and motorization, the characteristics and rules of residents' travel are constantly changing. Analysis of this information provides reference and guidance for transportation planning, urban management and residents' travel. With the development of mobile positioning and wireless communications, GPS signals, mobile phone signaling data and other data have established the foundation for obtaining wide-area travel information. This paper proposes a travel mode recognition method based on mobile phone signaling data. In the data preprocessing stage, the method effectively identifies and processes exceptions such as "ping-pong switching" effect and "data drift" effect through time-space threshold filtering, and accurately recognizes key points in the trajectory segmentation stage through feature analyses. In the recognition stage, this method utilizes the road network constraints to improve the calculation of features. The experimental results show that the method can effectively recognize the mode of residents' travel according to the mobile phone signaling data.

Keywords: Travel mode recognition · Mobile phone signaling · Clustering analysis · Data preprocessing · Road network constraints

1 Introduction

Travel mode analysis is one of the important categories in traffic analysis. It provides reference and guidance for traffic planning, urban management and residents' travel. Travel mode refers to a group of vehicles or means used by residents to complete a trip, such as walking, bicycles, motorcycles, cars, taxis, buses and rail transit. GPS signals and mobile phone signaling are the main data sources for travel trajectory analysis. Some researches extract position information from mobile phone GPS and other modules to calculate characteristic parameters, and use machine learning and other methods to recognize modes [1–3]. However, not all mobile devices have enabled GPS module in real time, and such methods are less robust to the analysis of residents with larger time spans and spatial extents.

J. Li et al. (Eds.): SPNCE 2019, LNICST 284, pp. 677–688, 2019.
https://doi.org/10.1007/978-3-030-21373-2_57

The signaling data reflects the user's communication information such as mode, time and location. The location information is obtained by collecting a cell identification number or by a cellular positioning technology. Since the signaling data can reflect the location information in a wide range of time and space, it is widely used in urban computing, especially in the field of travel analysis.

On the other hand, the current methods for analyzing the travel modes use the research framework made up of trajectory data preprocessing, trajectory segmentation and recognition models [2]. But there are still several problems. Firstly, the data anomalies are not fully considered in the preprocessing. The second is that it's not comprehensive enough in the trajectory segmentation, so that only one travel mode is adopted for a single travel. This is often not practical enough. The third is that the feature calculation method needs to be improved to enhance the accuracy of recognition. Therefore, based on the existing methods, this paper improves these three steps of data preprocessing, trajectory segmentation and mode recognition respectively to improve the accuracy of travel mode recognition.

This paper consists of four sections. Section 1 is an introduction. Section 2 states the problem, expounding the necessity of steps such as data preprocessing and trajectory segmentation, and roughly introduces the solutions. Section 3 proposes the recognition framework and proposes the solutions for each stage. Section 4 demonstrates the effectiveness of the recognition method through experimental analysis. Section 5 makes a summary.

2 Statement of Problem

Mobile phone signaling data cannot be directly used for analysis, since it is affected by factors such as terrain fluctuations, building distribution and multipath effects during propagation. And the uneven distribution of base stations makes the signaling data low in positioning accuracy and poor in quality. Therefore, in this paper, data preprocessing, trajectory segmentation and recognition models need to pay special attention to the problems caused by these effects.

2.1 Mobile Phone Signaling Data

The signaling data includes fields such as *MSID*, *Data_time*, *CELL_ID*, etc. The *MSID* is the unique identification number of the mobile phone user; *Data_time* indicates the generated time of current signaling; The *CELL_ID* is the connected base station number. The location of the base station is determined when the base station is planned and constructed. Each base station number uniquely determines its coordinates, so the user's location information is hidden in the *CELL_ID*. Therefore, a single piece of signaling data can be defined as Eq. 1.

$$signaling = \langle MSID, Data\_time, CELL\_ID, \ldots \rangle \tag{1}$$

A set of signaling data for a user can be represented as Eq. 2, where the signaling sequence is arranged in chronological order of *Data_time*.

$$userDATA \ = \{signaling_1, signaling_2, signaling_3, \ldots\} \tag{2}$$

2.2 Track Definition

The mobile phone user number (*id*) can be determined by the *MSID* and stored in the *users*, which contains the *MSID*s of all mobile phone users, so the user number set definition is as shown in Eq. 3.

$$users \ = \{id|id \ \in \ MSID\} \tag{3}$$

The location *point* of a single user corresponding to a single piece of signaling data. The location *point* definition is as shown in Eq. 4, where *lng* and *lat* respectively represent the longitude and latitude of the location point corresponding to the piece of signaling data, and t represents the signaling generated time, that is, $t \in Data\_time$. The s is the velocity of each *point*. Use P to collect all *point*s.

$$point \ = \ < id, lng, lat, t\ , s > \tag{4}$$

A trajectory is a sequence of a series of position *point*s of a single user stringed up by time. The trajectory is defined as *track*, shown in Eq. 5. Where $point_i \in P$, $point_i.t > point_{i+1}.t$, and $point_i.id = point_{i+1}.id$.

$$track = \{ < point_1, point_2, \ldots, point_i, \ldots, point_n > \} \tag{5}$$

The sub-track is defined as *travel* in Eq. 6, which contains the sequence of the sub-track *point*s of the corresponding user.

$$travel \subseteq track \tag{6}$$

2.3 Data Preprocessing

Due to factors such as uneven distribution of base stations and obvious terrain differences, the original signaling data has exceptions such as “data drift”, “ping-pong switching”, and backtracking during recording [4]. It is necessary to accurately identify these exceptions in the data preprocessing stage and properly handle them to guarantee the data quality for subsequent operations.

2.4 Track Segmentation

“A trip” of a resident usually corresponds to multiple modes of travel [5], such as the order of “walking” – “transit” – “railway” – “walking”. The different modes are connected by “staying” or “transferring”. Therefore, it is necessary to identify these "key points" (stay points and transfer points) in advance, and then divide the travel trajectory into several sub-tracks. Because different travel modes have large differences in speed change and time consumption, it can be identified by means of travel distance and travel time.

2.5 Travel Mode Recognition

In the study of travel mode recognition, the choice of travel feature variables has a great influence on the accuracy. Literature [6] uses the sensor data collected by the spiral instrument module in the mobile phone to analyze whether the user is in the bus or in the car. Literature [7] selects two travel feature variables, travel distance and maximum acceleration, to identify the four modes of travel including walking, bicycle, bus and car. Through analyzing the historical trajectories and similar research results, it is found that there are obvious differences on 5 feature variables such as travel distance, average speed, median speed, 95% quantile speed and low speed rate in each travel mode [8]. Therefore, the above five eigenvalues of each sub-track are calculated separately, and the SVM, C4.5 decision tree, BP neural network, convolutional neural network and other methods can be used to identify better results [8–10]. In addition, some researches have used the swarm intelligence algorithm such as particle swarm optimization to optimize the recognition mode to solve the problems of local optimum and precocity [8, 11]. On the other hand, most of the above identification methods use the labelled data to train the recognition model and then identify the unknown data. This supervised machine learning approach requires a lot of up-front work to label the mode of travel and lacks adaptability to signaling data in other regions.

Therefore, this paper considers the five travel features as the input variables of the recognition model, which are travel distance, average speed, median speed, 95% quantile speed and low speed rate. And recognize four travels modes of "walking", "rail transit", "buses" and "cars". At the same time, it is considered to use the less-cost unsupervised machine learning method to recognize the travel mode, conducting a cluster analysis to divide the sub-tracks of different travel modes by the internal differences of the travel data, to achieve the purpose of travel mode recognition.

3 Overall Process Design

Based on the above analysis, the overall process of recognize the travel modes of residents using signaling data is designed as shown in Fig. 1.

It mainly includes two stages: data preparation and travel mode recognition. The data preparation stage includes two steps of data preprocessing and track segmentation, and the travel mode recognition stage recognize the travel mode by cluster analysis of the travel feature data.

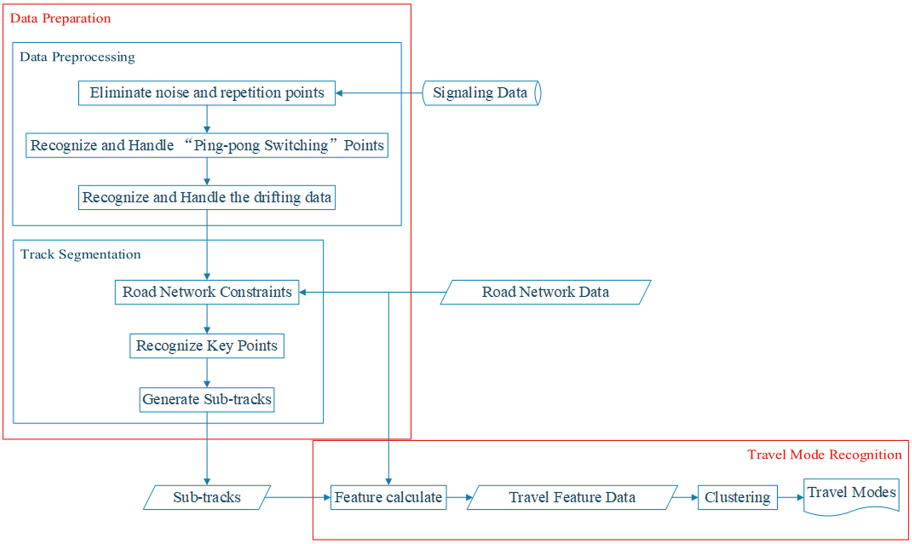


Fig. 1. The overall process of travel mode recognition

3.1 Data Preprocessing

The data preprocessing firstly uses the user number to divide all the signaling data into users and integrates them into the user track data set. Then the time and space thresholds are used to sequentially remove noise points, repetition points, "ping-pong switching" points, and "data drift" points for each user.

User Division. The user division step generates the signaling trajectory data set *userDATA* for each user, using all the given signaling sets, according to the user number *id*. In this process, the signaling data of all users is firstly sorted in ascending order according to the *id*, and it's arranged in the order of time stamp *Data_time* in the same *id* group. After that, the first signaling is read and a trajectory data set *userDATA* belonged to the user is created, then each subsequent signaling will be traversed in turn through the loop: if the user trajectory set of the next signaling does not exist yet, the user's *userDATA* will be created and the current signaling will be added; if it already exists, adds this signaling to the user's *userDATA*. When the next signaling is empty, the loop ends, and finally the *userDATA* collection of different users is generated. The single user's trajectory dataset *userDATA* is defined as the travel trajectory *track* for this user, and each signaling in is defined as a *point*.

Removing Duplicate Points. For a given user trajectory sequence *track*, the repeated points *point** in the set are deleted, as defined by Eq. 7.

$$point\ * \ = \{point|point_i.lng = point_{i-1}.lng,\ \text{AND}\, point_i.lat = point_{i-1}.lat\} \qquad (7)$$

"Ping-Pong Switching" Processing. For a given *track*, if there is a "ping-pong switch", only the first and last *point*s are retained. That is, any adjacent k time-adjacent track *point*s in the *track*, $point_{i+1}$, $point_{i+2}$, …, $point_{i+k}$, set the threshold thr_1. If the travel speed of the k track *points* $s = (d_{i+1,i+2} + d_{i+2,i+3} + \ldots + d_{i+k-1,i+k}) / (t_{i+1,i+2} + t_{i+2,i+3} + \ldots + t_{i+k-1,i+k}) > thr_1$, then $point_{i+1}$, $point_{i+2}$, …, $point_{i+k}$ is the sequence of track *point*s for "ping-pong switching", excluding the i+2th to i+k−1th track *point*s, leaving only $point_i$ and $point_{i+k}$.

"Data Drift" Processing. For a given user *track*, if there is drift data, such *point*s are eliminated. That is, for a given *track*, if $point_i$ exists, making the distance between $point_{i+1}$ and $point_i$ and the distance between $point_{i+1}$ and $point_{i+2}$ are both bigger than a given distance threshold thr_2, and the distance between $point_{i+2}$ and $point_i$ is less than 2 * thr_2, then the $point_{\mathrm{i+1}}$ in the middle is a "drift" point, and such *point*s are eliminated.

3.2 Track Segmentation

Since the user may adopt multiple travel modes in his trip, in the travel mode recognition, the user's *travel*, after preprocessing, is not directly used, and the *track* needs to be converted into segments (sub-tracks). This process is to find the key points in the travel trajectory first, then use them to segment a series of continuous data track points and divide the user track into several travel segments indicating different travel modes.

Road Network Constraints. In the communication system, the precise location of the user needs to be estimated using multiple pieces of signaling data and various positioning algorithms. In general, users travel on the road, so the user's travel distance is measured by calculating the road network distance of the travel.

First, use the coordinate generation function to generate the latitude and longitude coordinates of the specified two points as A (*lat*, *lng*), B (*lat*, *lng*). Then use the road network data to solve routes between the two points $path_1$, $path_2$, $path_3$ … $path_{\mathrm{n}}$, and the distances corresponding to each route. Since the adjacent two track points A and B in the travel trajectory are not far apart from each other, the solved distances of the navigation paths of the n paths between A and B are not significantly different, so the shortest navigation route among the n paths is extracted as the road network distance of the two track points. This function's input is the latitudes and longitudes of the two points, and the output is the road network distance of the two points, defined as function d ().

Key Points Recognition. The definition s represents the instantaneous velocity of each point, and d () represents the distance function between adjacent points. The speed calculation of the i+1th point is defined as shown in Eq. 8.

$$s_{i+1} = d(point_i, point_{i+1}) / (point_{i+1}.t - point_i.t) \quad (8)$$

Key points are not exactly one point, but a set of points in a continuous time range. The speed of the stay points is closed to 0, and the speed before and after the stay range is not closed to 0. So the stay points are defined as any adjacent k time-adjacent points $point_{i+1}, point_{i+2}, \ldots, point_{i+k}$ in the *track*, if $point_{i+1}.s = point_{i+2}.s = \ldots = point_{i+k}.s = 0$, and the time spans between $i + 1_{th}$ and $i + k_{th}$ exceeds th_3, all points from $i + 1$ to $i + k$ are stay points.

Another type of key points is the transfer points. Transfer is a form of transition from one mode to another. There are two situations when transferring. One is to transfer to another mode without waiting, such as taking a taxi immediately after walking or getting off the bus. The another needs to wait in its place, such as waiting for a bus after walking or recruiting taxis. The first transferring may not have points with a speed of 0, but there are adjacent points i and i+1, and the difference between the average velocity before i and the average velocity after i+1 is bigger than thr_4. The second transferring has points where the speed is 0, and the continuous time does not exceed the time of thr_5, and the difference of average speed before and after the transfer range exceeds thr_4.

Sub-track Generation. After recognizing a certain set of key points, the *track* can be divided into two different *travel*s by the first key point and the last key point. Therefore, the user's travel *track* is divided into several segments of *travel* using key points. Then users' travel segment set T is generated, which contains the all chronological sorted users' sub-tracks. The definition is as shown in Eq. 9. And each sub-track *travel* is unique numbered by *tID = id_i*, that is, a combination of user number and sequence number. The set T contains all the travel sub-tracks of all users and is distinguished by *tID*. The five features of each *travel* are calculated and included in the set.

$$T = < travel_1, travel_2, travel_3, \ldots, travel_n > \tag{9}$$

3.3 Travel Mode Recognition

Feature Calculation. The travel distance is calculated by using the cumulative method, which is the sum of the road network distances between the two points. The travel distance D is calculated as shown in Eq. 10.

$$D = \sum\nolimits_{i=1}^{n} d(i+1, i) \tag{10}$$

Since the speed of travel is not evenly distributed, the average speed cannot be calculated by dividing the total travel distance by the total travel time, but the average of the speeds between adjacent sets of track points. The calculation method of the average travel speed $\bar{S}$ is as shown in Eq. 11.

$$\bar{S} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{d(i+1, i)}{t_{i+1} - t_i} \tag{11}$$

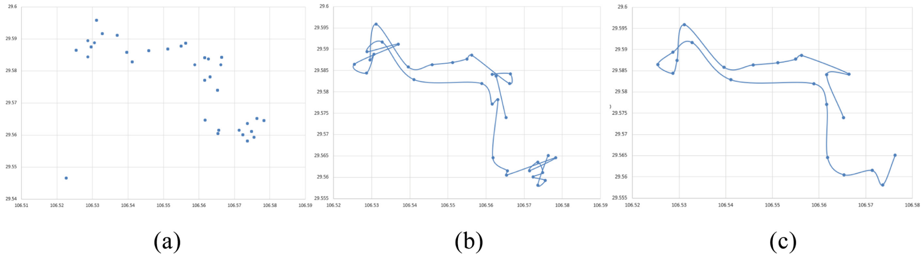
The speed set S of travel is arranged as S_N from small to large, and the median speed takes the middle number of the speed set S_N. When the number is even, the median speed takes the average of the middle two speeds. The 95% quantile speed takes the speed stands at the 95% position of the speed set S_N. The low speed rate takes the ratio of the speed less than 30 km/h in the set S. Therefore, the data set T is a table of n rows and 6 columns, each row represents a *travel*, and each column represents the feature value of the corresponding *travel* as the *travel* number *tID*, the travel distance *tTLen*, the average speed *tAvS*, the median speed *tMdS*, and the 95% quantile speed *tNFS* and low speed rate *tLSR*.

Recognition Method. The core of the recognition model is to use the inherent differences of the data to divide the data T into four different categories, each representing different modes of travel, namely walking, cars, buses and rail transit. Due to the improvement of feature extraction and eigenvalue calculation, T has significant differences in the five feature values. Therefore, the data set T can be effectively divided. In this paper, K-means clustering algorithm is used to divide the data T to reflect the rationality of the data preprocessing and data preparation process. Then, based on the analysis of the characteristics of the travel segments in the clusters, the travel mode can be inferred.

4 Experiments

The signaling data used in the experiment was provided by Chongqing Transport Planning and Research Institute. It contains 11587 records as signaling information of 14 users during the week from May 11 to May 17, 2015. The experiment was performed on the MATLAB platform. This paper uses three fields: *MSID*, *Data_time* and *CELL_ID*.

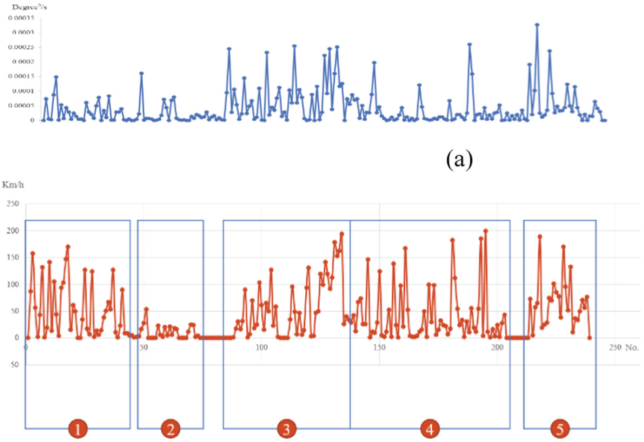
Data Preprocessing. After data preprocessing, exceptions are eliminated. The results are shown in Fig. 2. The Fig. 2(a) is the trajectory point of a user on the day of May 12, 2015. Figure 2(b) is the user's trajectory connected with the points that removes the repeated positioning and drift points, and Fig. 2(c) is the user's trajectory after removing the "ping-pong switching". Through data preprocessing, the user trajectory is smoother and clearer without losing the spatial distribution information of the location, and the real travel route is more effectively expressed.



(a) (b) (c)

Fig. 2. The preprocessing results

Road Network Constraints. The distance between two location points is calculated by calling the Baidu Map API's calculation method (*RouteMatrix* API). Firstly, the interface address of the Baidu Map API is accessed. The latitudes and longitudes of the two points, the return text type, the user's AK are transmitted and request a response. Then, the required distance data is extract according to the returned text result. The system automatically calculates the nearest route between the two points. The returning distance is the road network distance between the two points, and the time consuming is the time required for a corresponding travel mode.



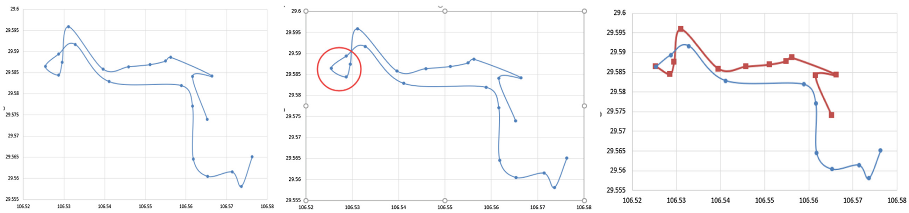
(a)

(b)

Fig. 3. The speed change after road network constraints

Figure 3 shows the change of the track speed of the user 4 before and after the road network constraint. Figure 3(a) shows the velocity distribution generated by directly calculating the Euclidean distance using the coordinates of the user, and Fig. 3(b) shows the speed distribution of the user after utilizing the road network constraint. The two distributions consistently have a similar trend in the speed change, but the distance calculation method is improved by the road network constraint in the Fig. 3(b), getting a better reflection of the difference of different travel modes in the speed distribution, and the key points can be more intuitively reflected in Fig. 3(b).

Track Segmentation. Figure 3(b) above shows the user's travel trajectory divided into five sub-tracks using indicators such as speed. Figure 4 below shows the trajectory segmentation process and results for User 1. The velocity turning occurs in the area in the red circle in Fig. 4(b) and the leftmost point's velocity approaches zero, so this point is recognized as a key point. As shown in Fig. 4(c), the key points split the track into two sub-tracks. The original data contains 14 users' signaling data, segmented resulting in 33 sub-tracks.



(a) (b) (c)

Fig. 4. Key point recognition result

Travel Mode Recognition. Through the road network constraint and eigenvalue calculation, the 33 sub-tracks include five travel characteristics. Because the difference of features among similar sub-tracks is obvious, the K-means clustering algorithm is implemented to divide the 33 *travel*, which target number of the cluster is set as 4. The result is shown in Table 1. The number of travel segments in each cluster is shown in the second column. The analysis of the travel segments in each cluster shows that cluster 1 has a long travel distance, a low speed rate, and an unstable speed. The travel mode can be estimated as buses. The travel segments in cluster 2 have a uniform velocity and a low rate of low velocity, which can be inferred to be rail transit. Cluster 3 has the characteristics of long travel distance, high speed and stability, and can be inferred to the travel mode of cars. The remaining cluster 4 has a short travel distance and a low speed rate, corresponding to the walking mode. The clustering results show that the characteristics of each category are obviously different, and the data characteristics of the same category are similar, which embodies the rationality of data preprocessing, track segmentation, feature selection and eigenvalue calculation.

Table 1. Numbers in each cluster

Cluster number	Number of travels
1	5
2	1
3	26
4	1
Total	33
Missing	0

Through cluster analysis, the 33 users travel segments include 5 bus travel, 26 travel by car, 1 travel by walking and 1 travel left by rail transit, which embodies the effectiveness of the method of clustering and analyzing travel modes. Table 2 shows the recognition result of 14 users' signaling data by using the travel mode recognition method proposed in this paper.

Table 2. Recognition Result for 14 users

User number	Travel modes	User number	Travel modes
01	Bus-Car-Car-Bus-Car	08	Car
02	Walking	09	Car
03	Car-Car-Bus	10	Bus
04	Car	11	Car
05	Car-Rail Transit-Car	12	Car
06	Car	13	Car
07	Car	14	Bus-Car

5 Conclusion

This paper proposes a travel mode recognition method based on mobile phone signaling data. In order to solve the data quality problems caused by uneven distribution of base stations and terrain differences, this method firstly recognizes and processes the exceptions in the data preprocessing stage through time and space threshold screening. In order to compensate for the calculation error caused by the inaccurate positioning of the base station, it is proposed to improve the feature calculation method by using the road network constraint, to improve the accuracy of the distance and speed. At the same time, the key points are accurately recognized through feature analysis, and the trajectory is segmented by using these key points to solve the “One Travel, One Mode” problem in traditional travel analysis. Finally, this paper uses the unsupervised machine learning method to cluster the sub-tracks after segmentation and combine the indicators of the data samples to recognize the travel modes. The experimental results show that the method can effectively recognize the travel mode according to the mobile phone signaling data. Due to the limited data samples, the modes are mainly recognized through the combination of unsupervised learning and empirical analysis. Subsequent research will conduct collecting volunteer data sets to validate the proposed method.

References

1. van Dijk, J.: Identifying activity-travel points from GPS-data with multiple moving windows. Comput. Environ. Urban Syst. **70**, 84–101 (2018)
2. Fang, Z., Jian-yu, L., Jin-jun, T., et al.: Identifying activities and trips with GPS data. IET Intell. Transp. Syst. **12**(8), 884–890 (2018)
3. Shafique, M.A., Hato, E.: Use of acceleration data for transportation mode prediction. Transportation **42**(1), 163–188 (2015)
4. Liu, Z., et al.: Traffic travel mode recognition method based on mobile phone grid data, CN108171973A (2018)
5. Wang, L., Zuo, Z.Y., Fu, J.H.: Travel mode character analysis and recognition based on SVM. J. Transp. Syst. Eng. Inf. Technol. **14**(3), 70–75 (2014)
6. Heydary, M.H., Pimpale, P., Panangadan, A.: Automatic identification of use of public transportation from mobile sensor data. In: Green Technologies Conference (GreenTech), pp. 189–196. IEEE (2018)

7. Zheng, Y., Liu, L., Wang, L., et al.: Learning transportation mode from raw GPS data for geographic applications on the web. In: Proceedings of the 17th International Conference on World Wide Web, pp. 247–256. ACM (2008)
8. Linlin, W.U., Biao, Y., Peng, J.: Research on travel mode identification of university students based on IPOS-SVM. Comput. Eng. **44**(1), 193–198 (2018)
9. Xiao, G., Juan, Z., Zhang, C.: Travel mode detection based on GPS track data and Bayesian networks. Comput. Environ. Urban Syst. **54**, 14–22 (2015)
10. Dabiri, S., Heaslip, K.: Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C: Emerg. Technol. **86**, 360–371 (2018)
11. Li, Z., Bo, C., Sun, J., et al.: Travel mode recognition based on particle swarm optimization and support vector machine. Appl. Res. Comput. **33**(12), 1–5 (2016)