Identifying Transportation Modes from Raw GPS Data:

Conference Paper · August 2016

DOI: 10.1007/978-981-10-2053-7_35

CITATIONS
29

READS
2,331

7 authors, including:

Mingzhao Li

24 PUBLICATIONS **204** CITATIONS

SEE PROFILE

Zhibiao Huang

JD.com

7 PUBLICATIONS **60** CITATIONS

SEE PROFILE

Identifying Transportation Modes from Raw GPS Data

Qiuhui Zhu[1], Min Zhu[1(✉)], Mingzhao Li[2], Min Fu[1], Zhibiao Huang[3], Qihong Gan[4], and Zhenghao Zhou[5]

[1] College of Computer Science, Sichuan University, Chengdu, China
zhumin@scu.edu.cn
[2] RMIT University, Melbourne, Australia
[3] Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, China
[4] Modern Education Technology Center, Sichuan University, Chengdu, China
[5] High School No. 7, Chengdu, China

Abstract. Raw Global Positioning System (GPS) data can provide rich context information for behaviour understanding and transport planning. However, they are not yet fully understood, and fine-grained identification of transportation mode is required. In this paper, we present a robust framework without geographic information, which can effectively and automatically identify transportation modes including car, bus, bike and walk. Firstly, a trajectory segmentation algorithm is designed to divide raw GPS trajectory into single mode segments. Secondly, several modern features are proposed which are more discriminating than traditional features. At last, an additional postprocessing procedure is adopted with considering the wholeness of trajectory. Based on Random Forest classifier, our framework can achieve a promising accuracy by distance of 82.85 % for identifying transportation modes and especially 91.44 % for car mode.

Keywords: GPS · Transportation mode · Random forest classifier

1 Introduction

Due to ever-growing traffic congestion, human activities have become more complex and associated life trajectories more intensive. User behaviour extraction, trajectory analysis and traffic pattern recognition are particularly significant for service provider and decision maker [3]. Normally, urban transportation modes are classified as road ones (car, bus, bike and walk) and rail ones (subway and train). It's obviously easy for researchers to distinguish rail ones from road ones by using such relatively simple methods as velocity modelling [2]. But our work focuses on identifying different means of road transportation modes which are more complicated depending on the raw GPS data.

In the past several years, researchers collected the data information of transportation modes through questionnaires and telephone interviews recorded by

W. Che et al. (Eds.): ICYCSEE 2016, Part I, CCIS 623, pp. 395–409, 2016.
DOI: 10.1007/978-981-10-2053-7_35

participants, which often resulted in inaccurate and incomplete data under easy overlooked or short trips [10]. Nowadays, urban sensing technology enables us to collect scientific data in a new and innovative way. Being two of the lowest-power sensors available on the phone, accelerometer and GPS are the predominantly used sensors in transportation mode identification. Accelerometer can detect acceleration in the phone's 3 axial directions. However, its readings are phone orientation and position-dependent, as well as vehicle-dependent. To identify transportation modes accurately, sampling rate is typically 10HZ and above. The high sampling rate, 3 axial directions, and position dependence make the classification complicated and increases power consumption [12]. Compared with accelerometer, GPS devices are becoming more popular for urban transportation mode identification in that its advantage of mobility and low sample rate.

In this paper, we present a robust framework to identify urban road transport including car, bus, bike and walk from raw GPS data. The contributions of the paper lies in three aspects: (1) A trajectory segmentation algorithm is designed based on logical assumptions, which can find almost 90 % single mode segments. (2) Several modern features are defined such as acceleration change rate, timeslice type, 85 % percentile velocity and acceleration, which were more discriminating than traditional features. (3) Relying on the wholeness of trajectory, a postprocessing procedure is developed to further improve the precision of mode identification without geographic information.

The remainder of the article is organized as follows. Related work is discussed in Sect. 2. Section 3 describes the dataset for our study. In Sect. 4 the classification model is introduced, followed by a presentation of postprocessing procedure in Sect. 5. In Sect. 6, the result of experiment is reported. Finally, we draw a conclusion, closed with corresponding discussion in Sect. 7.

2 Related Work

Identifying hybrid transportation modes from context information is still a relatively popular study. Biljecki et al. [2], Lin et al. [9], Shin et al. [14], Witayangkurn et al. [17] and Zheng et al. [19], present different approaches for identifying transportation modes. Table 1 shows a summary of the reviewed methods for transportation mode identification using GPS data. As shown in Table 1, a common processing step is applied to divide GPS logs into single mode segments based on criteria, such as transition points which denote a transition of transportation modes from one segment to another.

Precise identification of transportation mode is attributed to the high quality recognition of transition points. Many existing approaches of finding transition points require fine-grained acceleration data or geographic data. Usually, fine-grained acceleration data is generated by accelerometer embedded in mobile phone. Shin et al. [14] detected walking activity through acceleration data as a separator to partition the data stream into other activity segments. With the increase in sampling rate and time complexity, the accuracy of transportation mode identification can not be significantly improved. In addition, many

researchers explored transition points relying on geographic data instead of acceleration data. For instance, Liao et al. [8] segmented multi-modal trajectories by analysing the proximity to potential transition locations such as bus stops. Biljecki et al. [2] used OpenStreetMap data to help the segmentation process in a two-step process, partition of trajectories to single-journey segments based on two meaningful locations, and segmentation of journeys into single-mode segments. Geographic data such as road networks, bus stops and parking lots are not widely used by current approaches, because it can add to the cost and complexity of the system and increase calculation consumption. It is beneficial to develop approaches that do not rely on such data. Mountain and Raper [11] indicated that transition points mainly appeared in a rapid and sustained change in direction or speed when one user ceased one activity and began another. Zheng et al. [19] found transition points by a logical assumption that the start point and end point of walk segment can be a transition point in very high probability. Compared with their researches, we design a novel processing method which is robust for noise and perform better in finding transition points.

For each segment generated by transition points, most of the previous work was accomplished by building classification models with extracting significant features. Many of these models regard velocity as the significant feature for mode identification. Bolbol et al. [4] concluded the velocity variable could contribute positively to the classification. Due to the measurements of noise, researchers noted that approximately maximum values should be used [13,16]. Zheng et al. [18] proposed the method which is still robust for noise using top two maximum values of velocity. Besides, Stenneth et al. [15] derived features related to transportation network to improve classification effectiveness. In spite of high accuracy, their work needs great calculation consumptions. Zheng et al. [18] considered features that characterize changes in movement direction, velocity and acceleration. However, modern cities show more characteristics along with the development of times, such as traffic congestion, changes of people's behavior. Our work extract more powerful features to achieve high quality transportation mode identification, which can also fix and enhance traditional methods.

Table 1. A summary of the reviewed methods for transportation mode identification using GPS data

Study	Sensor	Geographic information	Modes	Accuracy
Zheng et al. [18]	GPS	No	4	76.2 %
Witayangkurn et al. [17]	GPS	No	5	77.4 %
Lin et al. [9]	GPS	Yes	4	76.3 %

3 Data Preprocessing

In this section, we first introduce the GPS trajectory dataset in Geolife project from Microsoft Research Asia and define several terms used in this paper. Then we describe the procedure of trajectory segmentation in detail.

3.1 Data Survey

The GPS trajectory dataset used in this paper was collected in Geolife project [18,20,21] from Microsoft Research Asia by 182 users in a period of over five years (from April 2007 to August 2012). The majority of the data was created in Beijing, the capital city of China, which has an integrated urban land use and a composite transportation network including the complex road network. This dataset recorded a broad range of users' outdoor movements, including not only life routines like go home and go to work but also some entertainments and sports activities. In each day of data collection, every user can label their data in the following way, 2011/03/19, 06:43:50-06:52:14, bus. Each trajectory in this dataset is represented by a sequence of time-stamped point. Every point contains the specific information of latitude, longitude and time. The uneven time sampling rate is set to be 2 s or 5 s. 73 users of total 182 users have labelled their trajectories with transportation mode. In this experiment, we use the dataset of version 1.3 which contains abundant information about transportation modes.

3.2 Travel Survey Definitions

In order to better understand GPS trajectories, some terms have been defined to describe different fragments of the trajectory. The total trajectory about a specific user in one day is called a trip. A trip is consist of a number of segments (such as car segment, walk segment, etc.). A new segment is generated when a user changes transportation mode or the time between two consecutive points exceed the specified threshold. A transition point is the point whose previous and posterior points belong to different segment. For instance, in Fig. 1, a transition point is generated when a user changes transportation mode from bus to walk. The time interval of two consecutive point is 2 s or 5 s, in Fig. 1 Δt is 2 s.

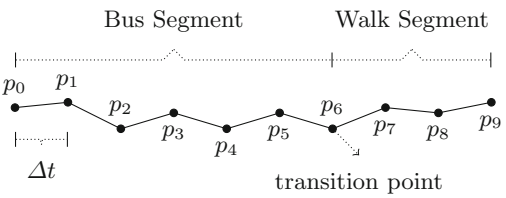


Fig. 1. Trip, segment and transition point

3.3 Data Preprocessing Procedure

Just as Mountain and Raper [11] stated that, the situation of one user may have ceased one activity and begun another mainly appears in a rapid and sustained change in direction or speed. We adopt the concept and presented a new trajectory segmentation method. Our approach is comprised of two portions, label specification and segmentation procedure. Algorithm 1 gives a detailed description of our trajectory segmentation procedure.

Label Specification. In this procedure, we first specify the points as *walk-point* if its velocity and acceleration is under the appointed threshold values, or *non-walk-point*. Owing to the error specification caused by noise points, it would lead to the following observations. First, the abnormal points may appear in the trajectory, which are usually far away from the nearest adjacent points. Second, the sharp-pointed points may occur with sharp turnings which would deviate from the trajectory trend and generate a zigzag segment. The second phenomenon usually happens when GPS points accumulated together, such as walk segments. The aforementioned two phenomenons would result in some points with high speed which belong to walk segment actually, thus may classify the walk segment as the bus segment or other modes mistakenly. So we put forward label revision method to eliminate the abnormal points specification. The loop in line 9 to 19 of Algorithm 1 describes the procedure in detail. The state of point (walk point or non-work point) depends on the states of its previous and posterior points. Our procedure for label specification can not only keep high points specification, but also maintain the trajectories original.

Segmentation. In segmentation procedure, we adopt the concept that the transition points usually appear in the situation that velocity encounter with sudden change. So we first get candidate transition points collection and many segments. Line 20 to 27 in Algorithm 1 describes the processing procedure in detail. Then, the segments with a length under the specified value d_{thd} should be merged with its nearby segments. Firstly, the merged segment usually continues almost 90 s through statistical analysis. This statistic time is consistent with the maximum tolerable time of pedestrians in red light. Most of pedestrians achieve 50 m with normal walk speed in 90 s. An example of this practice is the situation that, a car stops for a few seconds, then goes on the trip with high velocity, and the interval usually does not exceed 90 s. Secondly, the common users would not frequently change their transportation modes within such a short distance. For instance, within a short distance, it is impossible for a person to take the following transition, Bus→Walk→Bus→Walk→Bus. Hence this type of segments generated by interval should be merged with its nearby segments.

Algorithm 1. Trajectory Segmentation algorithm

Input: GPS logs T, velocity threshold v_{thd}, acceleration threshold a_{thd}, distance threshold d_{thd}, positive integer N represent point number, $scale$ represent coefficient of proportionality, $0 < scale \leq 1$, candidate transition points collection CTP, segment collection $CSEG$ divided by CTP

Output: a set of segment

function SEGALG($T, v_{thd}, a_{thd}, d_{thd}, N, scale$)
  for each $p_i \in T$ **do**
    if $p_i.v < v_{thd}$ **and** $p_i.a < a_{thd}$ **then**
      label p_i as walk-point
    else
      label p_i as non-walk-point
    end if
  end for
  Initialize positive integer $M \leftarrow \lceil (N * scale) \rceil$
  repeat
    for each $p_i \in T$ **do**
      if at least M in N of both adjacent previous points and posterior points of p_i is labelled as walk-point **then**
        label p_i as walk-point
      end if
      if at least M in N of both adjacent previous points and posterior points of p_i is labelled as non-walk-point **then**
        label p_i as non-walk-point
      end if
    end for
  until all points' label keep unchanged
  Initialize candidate transition points collection CTP
  $CTP \leftarrow \phi$
  for each $p_i \in T$ **do**
    if at least M in N of adjacent points in front of p_i is labelled as non-walk-point **and** at least M in N of adjacent posterior points of p_i is labelled as walk-point **or** opposite **then**
      $CTP \leftarrow CTP \cup p_i$
    end if
  end for
  generate a segment collection $CSEG$ based on the transition points collection CTP
  for each segment $seg_j \in CSEG$ **do**
    if distance of segment $seg_j < d_{thd}$ **then**
      merge seg_j with its previous and posterior segments into one segment
    end if
  end for
end function

4 Methodologies

This section is organized as follows. Firstly, features used to identify transportation modes are extracted from raw GPS logs. Secondly, the remainder of analysis in this section will focus on the inference model.

4.1 Feature Selection

Timeslice type (TS). The dataset was collected in Beijing which has a complex road network. Individual activity makes their moving trajectories interweave together due to daily routine. According to the time-statistical analysis, the rush hours mainly distributed in the time slot 7:00–10:00 and 16:00–21:00. During these two timeslices, people are more likely to encounter traffic congestions. When the average velocity of car is as slow as bike or in other uncommon situations, transportation modes may be labelled as other improper modes, then this mistaken information will result in inaccurate mode identification. Therefore, we divide the whole daily time into two timeslice types, as T_busy and T_idle. Specifically, we denote the timeslice type value of segment as T_busy if its timeslice falls into the specified time slots described above, otherwise, the type will be set to T_idle.

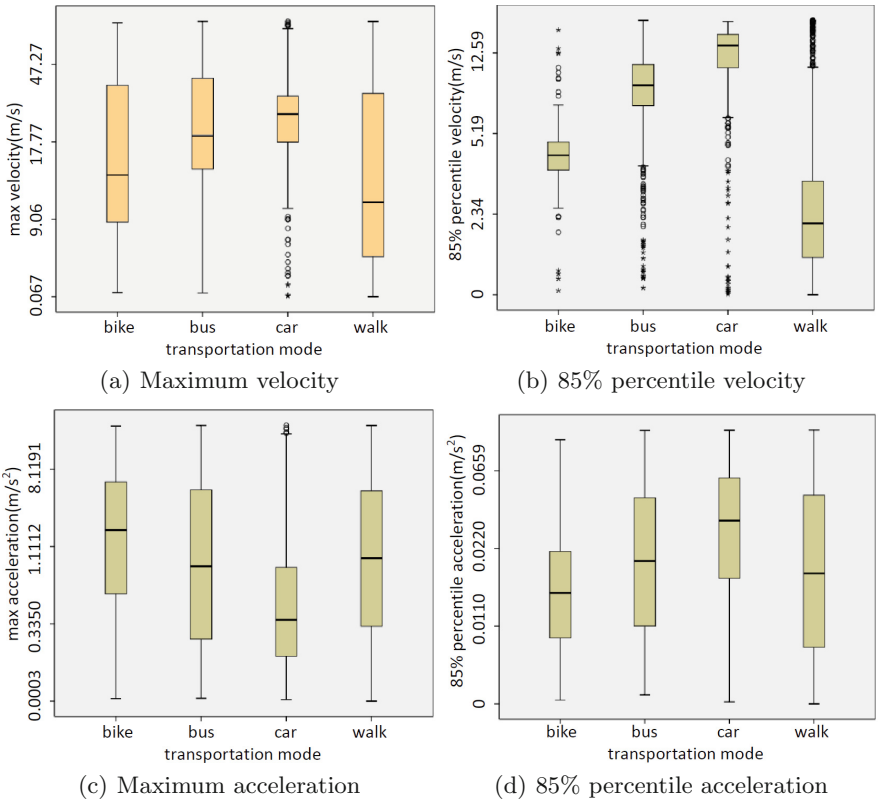
Acceleration change rate (ACR). Nowadays, a majority of modern cities have built the private passageways for buses in order to economize the time wasted on the roadway, especially in the heavy traffic city of Beijing. In rush hours, transportation modes always line up together under the red light. Even though, bus drivers can drive straight in special bus lane regardless of the states of cars or pedestrians in the arterial road. Also, taxi drivers always shift down or speed up frequently according to the drivers' personal behaviors, skills and preferences. For example, under the tempt of profit, a taxi driver would continually change velocity in a very small time slot to keep high speed, slow down or speed up suddenly. Therefore, there are many swings in the acceleration distribution of single car mode. However, the bus drivers or pedestrians are prone to keeping a small acceleration change. This phenomenon implies the potential mode difference among bus, car and walk. ACR modeling this principle is defined as Eqs. (1) and (2). First we can calculate the $ARate$ of each GPS point based on Eq. (1), in which A_i is the acceleration of point i. Then we can get the statistics of the number of GPS points whose $ARate$ are greater than a certain threshold A_r, and calculate ACR based on Eq. (2), in which $Distance$ is the total distance of single segment.

$$p_1 : ARate = |A_2 - A_1|/A_1; \tag{1}$$

$$ACR = |P_v|/Distance; \tag{2}$$

Where $P_v = \{p_i | p_i \in P, p_i.ARate > A_r\}$. Generally speaking, ACR makes it clear that the change frequency of acceleration in different transportation modes, which can be identified from each other.

85th percentile of velocity and acceleration (85thV, 85thA). As the primary variables for mode identification, velocity and acceleration play important roles in transportation mode identification. Due to sensor noise and data drift, the points always deviate from their original trend in the raw GPS trajectory. In Fig. 2, we use box plot to present the distribution of velocity and acceleration. The box plot uses the median, the approximate quartiles, and the lowest and highest data points to convey the level, spread and symmetry of a distribution of data values. Figure 2(a) and (b) make a comparison between 85 % percentile velocity and the maximum velocity. The comparison indicates the robustness of 85th percentile velocity, which is different from the maximum velocity that is prone to being disturbed by positioning errors. Also, as shown in Fig. 2(c) and (d), Fig. 2(d) describes the actual distribution of acceleration compared to Fig. 2(c). Noticeable mode shift of car reflects the potential characteristic when used in distinguishing car from other modes. These two features will get supports from later experiment results.



(a) Maximum velocity (b) 85% percentile velocity

(c) Maximum acceleration (d) 85% percentile acceleration

Fig. 2. Distribution of velocity and acceleration

Besides the features described above, we also extract other prominent features about velocity and acceleration including features between second maximum velocity and acceleration ($MaxV_2$, $MaxA_2$), mean velocity and acceleration (MeanV, MeanA), median velocity and acceleration (MedianV, MedianA), minimum velocity and acceleration (MinV, MinA), expectation of velocity and acceleration (Ev, Ea), covariance of velocity and acceleration (Dv, Da). Considering the details of GPS trajectory, features introduced by Zheng et al. [18] such as Heading Change Rate (HCR), Stop Rate (SR) and Velocity Change Rate (VCR), are extracted from the raw GPS trajectory.

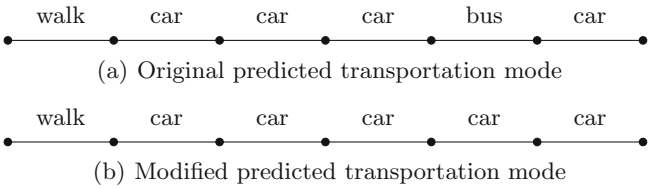
4.2 Classification Model

The GPS trajectory dataset used in this paper was collected in Geolife project from Microsoft Research Asia. From this type of dataset, experimental results have demonstrated that the segmentation method based on transition points followed by a Decision Tree algorithm showed the highest identification accuracy of the transportation modes [18,19]. In our work, we decided to employ Random Forest as a model because the works by Stenneth et al. [15] showed that performance of Random Forest is better than Decision Tree in transportation mode identification. Random Forest Classifier, developed by Breiman et al. [6], is an ensemble classification and regression method that constructs a number of decision trees at the training level, predicts the class using each tree and outputs the final class as the mode of the individually predicted classes. One of the major advantages of RF model is that it can handle high dimension data, obtain important features automatically. It is obvious to improve the training speed by extracting a few attributes from original features set every time. It is more stable and less prone to prediction errors as a result of data perturbations [7]. Therefore, RF model is viewed as one of the most accurate general-purpose learning techniques available [1].

5 Postprocessing Procedure

After applying the former inference model, we can obtain the predicted transportation modes of segment divided by transition points. Considering the wholeness of trajectory and walk mode logical assumption [19], and the analysis that car, bike and bus mode have similar velocity in the heavy traffic. Therefore, we view the whole trajectory as a chain of predicted modes, and modify the predicted modes as the high probability modes which follow the general trend. This practice can solve the above issues to some extent. Figure 3(a) gives the predicted mode sequence after the first classification. According to experience, the segment classified as bus mode in the line is likely the segment with improper classification. So we change the predicted bus mode to car mode, reasoning that the person is impossible to switch car mode to bus mode, or bike mode directly. Moreover, this segment is surrounded with predicted car mode and has similar characteristics with non-walk mode. After processing, we 'repair' the original

predicted mode sequence as the available final-result mode sequence. Figure 3(b) shows the modified transportation mode line.



(a) Original predicted transportation mode

(b) Modified predicted transportation mode

Fig. 3. General case of segment postprocessing

6 Results and Discussion

In this section, we firstly describe how we select the parameters for each procedure. Secondly, we verify the efficiency of presented features and get the corresponding results about our overall inference model.

6.1 Parameter Setting

In the preprocessing step, two consecutive GPS points are divided into two different segments if the time gap is more than 20 min. When labelling the points in Algorithm 1, the value of velocity and acceleration threshold v_{thd}, a_{thd} is 1.8 m/s and 0.6 m/s^2 [19]. From Fig. 4, when variable N and $scale$ is set to be 10 and 0.8, we can get highest recall of transition points. Referring to the situation that most of pedestrians achieve 50 m with normal walk speed in 90 s, then interval distance d_{thd} is set to be 50 m. About the features, we set the threshold value for HCR, SR and VCR of 15, 3.2 and 0.36 respectively [18]. Figure 5 shows the inference accuracy changing over the threshold value A_r when ACR is used alone to identify transportation modes. Obviously, when A_r equals to 0.25, ACR shows its greatest advantages in identifying transportation modes.

Besides, our Random Forest classifier is the combination of 100 randomized decision trees. At each node in decision tree, a subset of features is randomly selected. Typically the size of every subset is $\sqrt{n}$, where n is the total number of features. In our experiment, the number of feature set used in the inferring is 19, i.e., n equals to 19. Thus, we set the number of subset features k is 4. With regard to the toolkit we used in the experiments, Weka (Waikato Environment for Knowledge Analysis) 3.7 toolkit [5] is selected to implement Decision Tree and Random Forest. About 70 % of all the segments are trained, and the remaining are used for testing.

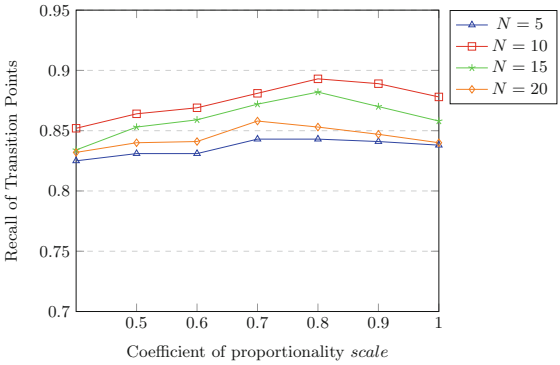


Fig. 4. Selecting value for variable N and *scale* through recall of transition points

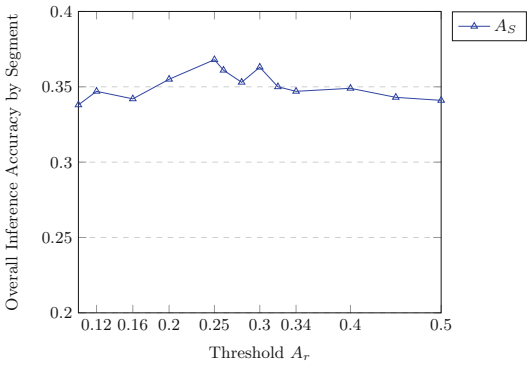


Fig. 5. Selecting threshold (A_r) for ACR. ACR is the only feature used in the inference model

6.2 Effectiveness of Preprocessing and Postprocessing Step

The preprocessing step divides the GPS trajectory into single mode segments based on transition points collection. In the first part, we evaluate the effectiveness of our label specification procedure by the precision of points specification. According to the statistics of the number of walk points and non-walk points, the precision of points specification rises to 78.79 % from 76.42 % after the procedure of label specification, demonstrating 2.37 % improvment in labelling the state of points. Then the second part of preprocessing step is measured in terms of the recall of transition points. While the recall of transition points has higher priority over their precision mainly because we hope to obtain all the transition points. Therefore, if the distance between an inferred transition points and its ground truth is within 150 m, we regard the transition point as a correct inference. As a result, we retrieved 89.3 percent of the actual transition points from the corresponding GPS data.

Meanwhile, in the postprocessing procedure, we consider the wholeness of trajectory to further improve the precision of mode identification. As it turns out, the postprocessing procedure has achieved an accuracy by distance of 82.85 % based on the inference model using features we explored in the experiment, while 81.13 % without postprocessing. This indicates that we can make almost 2 percent improvement in accuracy by distance for transportation mode identification. Zheng et al. [18] put forward the graph-based postprocessing which bring 3.4 % promotion over the preliminary inference result. This comparison makes clearly that our postprocessing procedure accuracy is 1.7 % lower than the performance of graph-based postprocessing. Nevertheless, the graph-based postprocessing mentioned above requires spatial knowledge as input, we have to know most geographic information of urban region. For the regions which are not covered by trajectory data, it can not perform better. What is more, it needs a lot of statistical calculation. But our postprocessing procedure not only doesn't rely on spatial knowledge, but also can handle whole trajectory of individuals anywhere.

6.3 Feature Evaluation

Considering the out-off-balance caused by the distance of each segment with its characteristics, we focus on the accuracy by segment (AS), which means the accuracy of the number of segment classified correctly. In order to evaluate the efficiency of the features, we ranked features by information gain and single feature classification in our work. From Table 2, we can observe that two ranking methods keep top 11 identical features, 85thV shows obvious advantage over other features, ACR performs well in identifying transportation modes and 85thA outperforms other features related to acceleration.

Table 2. Classification features ranking

(a) Information gain ranking

Rank	Features
1	85thV
2	MedianV
3	MeanV
4	Ev
5	Dv
6	SR
7	HCR
8	ACR
9	VCR
10	$MaxV_2$
11	85thA

(b) Single feature for segment accuracy ranking

Rank	Feature	AS
1	85thV	45.70%
2	MeanV	43.23%
3	MedianV	41.57%
4	Ev	41.52%
5	HCR	40.87%
6	SR	39.74%
7	VCR	38.83%
8	Dv	37.53%
9	ACR	36.80%
10	$MaxV_2$	36.66%
11	85thA	29.78%

Table 3. Feature comparison

Combine features	Traditional features	New features
52 %	58.4 %	60 %

Another, we evaluate entire features in three different combination ways. ACR, 85thV, 85thA and TS make up the *combinefeatures*. The *traditionalfeatures* include top two maximum velocity ($MaxV_1$, $MaxV_2$), top two maximum acceleration ($MaxA_1$, $MaxA_2$), MedianV, MedianA, MinV, MinA, MeanV, MeanA, Ev, Dv, Ea, Da, SR, HCR and VCR, while the *newfeatures* is the feature set which we explored in this experiment. The results are shown in Table 3, from which the overall accuracy by segment of *combinefeatures* indicates that the new features are enough to identify transportation mode in the proposed work. It means that our approach will not lose too much performance when applying our feature combination only. Meanwhile, considering the whole features, the overall accuracy by segment of *newfeatures* rises from 58.4 % to 60 % compared the *traditionalfeatures*.

Finally, when Decision Tree is selected to perform the inference, the overall accuracy by segment can achieve 58.9 %, about 1.1 % lower than Random Forest which can perform 60 %. To summarise, Random Forest outperforms other classification model while considering the *newfeatures*.

6.4 Mode Identification

For transportation mode identification, we evaluated the classification by using two well-known performance measures: Precision and Recall. As described in the previous section, we used Random Forest classifier as the mode identification classifier. We could achieve an overall accuracy by distance of 82.85 %, demonstrating a better discrimination about transportation modes than previous researches [9,14,17,18]. As shown in Table 4, walk mode identification achieves about 66.11 % accuracy while 91.23 % of recall. Very few actual walk segments are classified as other modes because of label revision in data preprocessing step. In addition, both accuracy and recall of car mode identification in matrix remain high scores. Among many reasons, the most important one is that car segments hold very long distance with the characteristic of high velocity. Also, the overall accuracy by distance of bike mode can reach about 81.82 %.

Table 4. Matching matrix of the detecting results in terms of distance and percentage

Detected results (KM and percentage)					
Mode	Walk	Car	Bus	Bike	Recall
Walk	1665.0 (91.23 %)	50.7 (2.78 %)	72.2 (3.96 %)	37.2 (2.03 %)	91.23 %
Car	400.5 (3.22 %)	11547.3 (92.91 %)	469.6 (3.78 %)	11.7 (0.9 %)	92.91 %
Bus	192.0 (6.25 %)	626.8 (20.40 %)	1931.7 (62.89 %)	321.3 (10.46 %)	62.89 %
Bike	261.0 (10.00 %)	403.8 (15.48 %)	276.8 (10.63 %)	1666.3 (63.89 %)	63.89 %
Precision	66.11 %	91.44 %	70.24 %	81.82 %	

From the misclassification between car and bus, 3.78 % length of car mode are misclassified as bus mode and 20.4 % length of bus mode are misclassified as car mode. For the dataset we explored in this paper, it was created in Beijing, China. Being the capital city of China, it has the complex road network. During the daytime, bus and car are more likely to encounter traffic congestions and perform the similar behavior. In spite of this phenomenon, the overall accuracy by distance of bus mode can perform 70.24 percent, which is better than similar researches [9,18].

7 Conclusions

In this paper, we presented a new robust framework for identifying road transportation modes focusing on raw GPS data. Firstly, without geographic information, we design a trajectory segmentation algorithm which can find almost all the transition points. Secondly, we propose some features which are more discriminating in transportation mode identification than the features which existing works [4,13,18] used. Additionally, the natural flow of whole GPS trajectory is considered when processing segments after classification. As a result, our work maintains relatively high precision when comparing with previous work [9,14,17,18], especially for car mode detection. The overall accuracy by distance of our framework can perform 82.85 % in transportation mode identification.

The application of our framework may be useful for user behavior analysis. However, many issues remain open and they are worthy of further study. First, the wholeness of trajectory maybe play a more significant role in identifying transportation modes instead of each segment individually. We can not ensure the natural flow of the travel pattern of every participant. Secondly, segmentation method generates many little segments segmented by transition points and influences the effectiveness of new introduced features. So combing different segmentation method is also potential work to do.

Acknowledgments. The research was supported by the project of Key Technology R&D program of Sichuan province (2013GZ0015). Special thanks to Zhaoyang Xie, Binbin Lu, Ruoyu Jia and Lei Gong for their inspiring discussions on the design of framework. Furthermore, we would also like to thank all of the reviewers for their valuable and constructive comments, which greatly improved the quality of this paper.

References

1. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. **13**(1), 1063–1095 (2012)
2. Biljecki, F., Ledoux, H., Van Oosterom, P.: Transportation mode-based segmentation and classification of movement trajectories. Int. J. Geogr. Inf. Sci. **27**(2), 385–407 (2013)
3. Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the netherlands. Transp. Res. Part C Emerg. Technol. **17**(3), 285–297 (2009)

4. Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J.: Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. Comput. Environ. Urban Syst. **36**(6), 526–537 (2012)
5. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: Weka–experiences with a Java open-source project. J. Mach. Learn. Res. **11**, 2533–2541 (2010)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R.: Random forests for land cover classification. Pattern Recogn. Lett. **27**(4), 294–300 (2006)
8. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Building personal maps from GPS data. Ann. N. Y. Acad. Sci. **1093**(1), 249–265 (2006)
9. Lin, M., Hsu, W.J., Lee, Z.Q.: Detecting modes of transport from unlabelled positioning sensor data. J. Location Based Serv. **7**(4), 272–290 (2013)
10. McGowen, P., McNally, M.: Evaluating the potential to predict activity types from GPS and GIS data. In: Transportation Research Board 86th Annual Meeting, Washington (2007)
11. Mountain, D., Raper, J.: Modelling human spatio-temporal behaviour: a challenge for location-based services. In: GeoComputation, Brisbane (2001)
12. Sankaran, K., Zhu, M., Guo, X.F., Ananda, A.L., Chan, M.C., Peh, L.S.: Using mobile phone barometer for low-power transportation context detection. In: Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, pp. 191–205. ACM (2014)
13. Schuessler, N., Axhausen, K.: Processing raw data from global positioning systems without additional information. Transp. Res. Rec. J. Transp. Res. Board **2105**, 28–36 (2009)
14. Shin, D., Aliaga, D., Tunçer, B., Arisona, S.M., Kim, S., Zünd, D., Schmitt, G.: Urban sensing: using smartphones for transportation mode classification. Comput. Environ. Urban Syst. **53**, 76–86 (2015)
15. Stenneth, L., Wolfson, O., Yu, P.S., Xu, B.: Transportation mode detection using mobile phones and GIS information. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 54–63. ACM (2011)
16. Stopher, P., FitzGerald, C., Zhang, J.: Search for a global positioning system device to measure person travel. Transp. Res. Part C Emerg. Technol. **16**(3), 350–369 (2008)
17. Witayangkurn, A., Horanont, T., Ono, N., Sekimoto, Y., Shibasaki, R.: Trip reconstruction and transportation mode extraction on low data rate GPS data from mobile phone. In: Proceedings of the International Conference on Computers in Urban Planning and Urban Management (CUPUM 2013), pp. 1–19 (2013)
18. Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y.: Understanding mobility based on GPS data. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 312–321. ACM (2008)
19. Zheng, Y., Liu, L., Wang, L., Xie, X.: 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)
20. Zheng, Y., Xie, X., Ma, W.Y.: Geolife: a collaborative social networking service among user, location and trajectory. IEEE Data Eng. Bull. **33**(2), 32–39 (2010)
21. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th International Conference on World Wide Web, pp. 791–800. ACM (2009)