Transportation modes behaviour analysis based on raw GPS dataset

Qiuhui Zhu and Min Zhu*

College of Computer Science,
Sichuan University,
Chengdu, China
Email: zhuqiuhui813@gmail.com
Email: zhumin@scu.edu.cn
*Corresponding author

Mingzhao Li

School of Science,
GPO Box 2476 Melbourne VIC 3001, Australia
Email: mingzhao.li@rmit.edu.au

Min Fu

College of Computer Science,
Sichuan University,
Chengdu, China
Email: fumin1412@gmail.com

Zhibiao Huang

Chengdu Institute of Computer Application,
Chinese Academy of Sciences,
No. 11, 4th Section, South Renmin Road,
Wuhou District, Chengdu, China
Email: huangzb007@gmail.com

Qihong Gan

Modern Education Technology Center,
Sichuan University,
Chengdu, China
Email: gqh@scu.edu.cn

Zhenghao Zhou

High School No. 7,
No. 1 Linyin Middle Street,
Wuhou District, Chengdu, China
Email: vikingross70@yeah.net

Abstract: Significant information has been existed in the global positioning system (GPS) data for understanding behaviours and transport planning. However, fine-grained identification of transportation modes is still required in the literature. In this paper, we present a robust framework to identify different means of transportation modes from raw GPS dataset. We make the following contributions. 1) We design an effective trajectory segmentation algorithm to divide raw GPS trajectory into single mode segments based on logical assumptions; 2) we propose several modern features, which are more discriminating than traditional features; 3) we adopt an additional segments expansion procedure by considering the wholeness of trajectory. Experiments prove that our framework achieves a promising accuracy for identifying transportation modes.

Keywords: global positioning system; GPS; transition point; transportation mode; random forest classifier.

Biographical notes: Qiuhui Zhu is a Master degree Candidate in College of Computer Science, Sichuan University, Chengdu, China. His research interests are related to urban computing and data analysis.

Min Zhu received her PhD in Computer Application from Sichuan University of China in 1996. She joined the College of Computer Science of Sichuan University as an Assistant in 1996, and currently she is a Professor and Vice Dean, as well as the Director of the Lab for Vision-Computing at College of Computer Science. She has wide research interests, mainly including visualization, image processing, intelligent information processing and wireless network. She is a member of IEEE and China Computer Federation (CCF). She is an evaluation expert of the National of Science and Technology Awards (NOSTA), the Ministry of Science and Technology (MOST), the Ministry of Education (MOE), and National Science Foundation of Beijing and Zhejiang province.

Mingzhao Li is currently a PhD candidate in the School of Computer Science, RMIT University, Melbourne, Australia. He received his Master degree in Computer Technology and Bachelor degree in Software Engineering both from Sichuan University in 2014 and 2011. His research interests are mainly in the areas of information visualisation and visual analytics, including multidimensional visualisation, geographical visualisation, text and document visualisation; computer-human interactions and composite visualisation design, etc.

Min Fu is a Master degree Candidate in College of Computer Science, Sichuan University, Chengdu, China. Her research interests are related to data analysis and expert prediction.

Zhibiao Huang is a Master degree Candidate in Chengdu Institute of Computer Application, Chinese Academy of Sciences, China. His research interests are related to data analysis and image processing.

Qihong Gan is working for Modern Education Technology Center, Sichuan University, Chengdu, China. Her research interests are related to information visualisation and educational technology. She received her Master degree in College of Computer Science, Sichuan University, Chengdu, China. She has published articles in international journals and conferences.

Zhenghao Zhou is a senior high school student in No. 7 High School, Chengdu, China. He is good at doing simple researches in the field of mapping and navigation. He is also the Team Leader of No. 7 High School VEX Robotics Team 7627A and during the latest 2016 VEX World Competition, he and his teammates made the best record since 7627A has been established a few years ago.

This paper is a revised and expanded version of a paper entitled ‘Identifying transportation modes from raw GPS data’ presented at The Second International Conference of Young Computer Scientists, Engineers and Educators (ICYCSEE) 2016, Harbin, China, 20–21 August 2016.

1 Introduction

Human activities and associated life trajectories become more complex and intensive due to ever-growing traffic congestion. User behaviour extraction, trajectory analysis and traffic pattern recognition are particularly significant for service provider and decision making. In recent years, researchers focus more attention on user behaviour analysis based on mobile big data. Through the detail analysis of mobile big data, Liao et al. (2015) introduced a mobile data infrastructure and a mobile data lifetime management model to reveal hidden information and potential user behaviour. Zheng et al. (2008a), Witayangkurn et al. (2013) and Shin et al. (2015) mined user’s traffic patterns by designing different framework. In addition to the application of mobile big data, Hung et al. (2011) also utilised quantifiable emotional data for travelling schedule recommendation. Liao et al. (2006) and Biljecki et al. (2013) obtained geographic data as supplementary information for transportation mode identification. As an important kind of human behaviour, people’s transportation mode behaviour, such as walking, driving, etc., can endow their mobility with more significance and provide the pervasive computing systems with rich context information.

In the past several years, researchers collected the data information of transportation modes through questionnaires and telephone interviews recorded by participants. This way often results in inaccurate and incomplete data under easy overlooked or short trips, because users are reluctant to undertake such heavy burden after a long trip. It is not feasible to require them to tag corresponding patterns to
their trajectory data clearly, let alone the accurate time, especially in the hybrid transportation modes trip (McGowen and McNally, 2007). Nowadays, global positioning system (GPS) devices have become the mainstream sensor devices as the positioning technologies advances, and can generate extensive volumes of trajectory data detailing tourist trajectories. Inexpensive and straightforward acquisition of trajectory data makes sequential time-stamped positions more convenient and popular.

Normally, transportation modes are classified as road ones (car, bus, bike and walk), rail ones (subway and train) and aircraft ones (plane). In this paper, we present a robust framework to identify different means of transportation modes including car, bus, bike, walk, train and plane from raw GPS dataset. The contributions of the paper lie in three aspects:

- A trajectory segmentation algorithm is designed based on logical assumptions, which can find almost 90% single mode segments.
- Several modern features are defined such as acceleration change rate (ACR), timeslice type (TS), 85% percentile velocity and acceleration, which were more discriminating than traditional features.
- Relying on the wholeness of trajectory, the segments expansion procedure is developed to further improve the precision of mode identification without geographic information.

The rest of the paper is structured as follows. We first discuss the related work in Section 2, followed by the architecture of our framework. Section 4 introduces the data preprocessing. We then give a presentation of methodologies in Section 5. Next, we conduct experiments and present the result in Section 6. Finally, conclusion and future works are discussed.

2 Related work

Identifying hybrid transportation modes from context information of GPS dataset is still a relatively popular study. Table 1 gives an overview of the main methods, with their main characteristics. The published methods classify between four and five transportation modes, and the accuracies of the studied methods are mostly between 76% and 82%. All methods use a trajectory segmentation method to divide trajectory into single mode segments, then related features are extracted as input of classification models.

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. (2015) 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 cannot be significantly improved. In addition, many researchers explored transition points relying on geographic data instead of acceleration data. For instance, Liao et al. (2006) segmented multi-modal trajectories by analysing the proximity to potential transition locations such as bus stops. Biljecki et al. (2013) 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 (2001) 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. (2008b) 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.

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensor</th>
<th>Geographic information</th>
<th>Modes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng et al. (2008a)</td>
<td>GPS</td>
<td>No</td>
<td>4</td>
<td>76.2%</td>
</tr>
<tr>
<td>Witayangkurn et al. (2013)</td>
<td>GPS</td>
<td>No</td>
<td>5</td>
<td>77.4%</td>
</tr>
<tr>
<td>Lin et al. (2013)</td>
<td>GPS</td>
<td>Yes</td>
<td>4</td>
<td>76.3%</td>
</tr>
<tr>
<td>Shin et al. (2015)</td>
<td>GPS, accelerometer</td>
<td>No</td>
<td>4</td>
<td>82.0%</td>
</tr>
</tbody>
</table>

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. (2012) 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 (Schuessler and Axhausen, 2009; Stopher et al., 2008). Zheng et al. (2008a) proposed the method which is still robust for noise using top two maximum values of velocity. Besides, Stenneth et al. (2011) derived features related to transportation network to improve classification effectiveness. In spite of high accuracy, their work needs great calculation consumptions. Zheng et al. (2008a) considered features that characterise 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 behaviour. Our work extracts more powerful features to achieve high quality transportation mode.
identification, which can also fix and enhance traditional methods.

3 Architecture of our framework

As shown in Figure 1, the architecture of our approach includes two parts, offline training and online prediction. In the offline training part, we first divide GPS trajectory into single mode segments based on preprocessing procedure and extract detail features from these segments. Then, the features and corresponding ground truth are employed to train a classification model for online prediction. In the online prediction part, when users upload their GPS logs to our system, same as aforementioned process, we first divide GPS trajectory into single mode segments and extract the same features from each segment. Secondly, take these features as input; inference model will output the predicted transportation modes. Finally, the segments expansion procedure is applied to revise the predicted result.

Figure 1 Architecture of our framework

4 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.

4.1 Data collection

The GPS trajectory dataset used in this paper was collected in Geolife project (Zheng et al., 2008a, 2009, 2010) 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. In the data collection program, a portion of users have carried a GPS logger for years, while some of the others only have a trajectory dataset of a few weeks. Figure 2 shows GPS log and the distribution of the dataset. 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. A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude. Considering the similar behaviours, we regard the label of both taxi and car as car, subway and train as train.

Figure 2 GPS data log and distribution of data

(see online version for colours)

4.2 Travel survey definitions

In order to de-construct a GPS trajectory, some definitions have been standardised to be used for the description of different fragments. 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.). The route between any two consecutive GPS points is called a segment. A transition point is the point whose previous and posterior points belong to different segment.
4.3 Data preprocessing procedure

Just as Mountain and Raper (2001) 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. Once the transition point is detected, using transition points as a separator, a bundle of segments can be created from the GPS dataset. However, Shin et al. (2015) stated that a long sampling period can result in a small number of segments with missing an important segmentation, and a short sampling period is more accurate but can produce unnecessary segments. Thus, our method chooses a sampling period of 5 s to compute the feature value as Shin et al. (2015) discuss. Moreover, we need to address problems with data such as occasional gaps caused by signal shortages and noise. The positions of GPS point among gaps are predicted according to the method proposed by Li et al. (2016). Our trajectory segmentation approach is comprised of two portions, label specification and segmentation procedure. We describe the detecting procedure as follows:

Step 1 Label specification. Label GPS point as walk-point if its velocity lower than velocity threshold \( v_{thd} \) and its acceleration lower than acceleration threshold \( a_{thd} \), or label as non-walk-point.

Step 2 Label revision. For example, if there are at least \( M \) in \( N \) of both adjacent previous points and posterior points of \( p_i \) is walk-point, then we label \( p_i \) as walk-point, or opposite.

Step 3 Transition point detection. When a GPS point’s previous and posterior segment belong to different segment, we regard it as a transition point.

Step 4 Segment combination. If the length of segment composed by two transition point less than a threshold \( d_{thd} \), we merge this segment with its previous and posterior segments into one segment.

4.3.1 Label specification

Firstly, we specify the points as walk-point [white points in Figure 3(a)] if its velocity and acceleration is under the appointed threshold values, or non-walk-point [black points in Figure 3(a)]. However, as depicted in Figure 3(b), sometimes some abnormal points may appear in the trajectory because of measurement error or other situations. A few non-walk-point are specified as walk-point mistakenly. So we put forward label revision method which adopted the greedy theory to eliminate the abnormal points specification. The state of point depends on the state of its previous and posterior points.

4.3.2 Segmentation

Usually, transition point exists in the situation of rapid and sustained change in direction or speed, such as users travel from walk points to non-walk points or non-walk points to walk points. So we get candidate transition points collection firstly. Meanwhile, we find many transition points which is more than real transition points. In other word, we cannot get high precision of transition points. As illustrated in Figure 3(c), some segments composed by consecutive walk points appear in the bus segment. This phenomenon can be explained as:

1. Bus arrives at station and restart for the next station frequently. Because 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 \( \rightarrow \) Walk \( \rightarrow \) Bus \( \rightarrow \) Walk \( \rightarrow \) Bus.

2. A bus stops for a few seconds in red light, then goes on the trip with high velocity, and the interval usually does not exceed 90 seconds.

This statistic time is consistent with the maximum tolerable time of pedestrians in red light. Most of pedestrians achieve 50 metres with normal walk speed in 90 seconds. In order to eliminate unnecessary transition points, the segments with a length under the specified value \( d_{thd} \) should be merged with its nearby segments.

Figure 3 An example of detecting transition points

5 Methodologies

This section is organised 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.

5.1 Feature selection

5.1.1 Timeslice type

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 TSs, as $T_{busy}$ and $T_{idle}$. Specifically, we denote the TS value of segment as $T_{busy}$ if its timeslice falls into the specified time slots described above, otherwise, the type will be set to $T_{idle}$.

5.1.2 Acceleration change rate

Nowadays, a majority of modern cities have built the private passageways for buses in order to economise 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 behaviours, 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 difference among different modes. $ACR$ modelling this principle is defined as equations (1) and (2). First, we can calculate the $ARate$ of each GPS point based on equation (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 equation (2), in which $Distance$ is the total distance of single segment.

$$p_i = ARate = \frac{|A_i - A_r|}{A_r};$$

(1)

$$ACR = \frac{\sum_{p_i \in P} |A_i - A_r|}{Distance};$$

(2)

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

Figure 4  Distribution of velocity and acceleration, (a) maximum velocity (b) 85% percentile velocity (c) maximum acceleration (d) 85% percentile acceleration (see online version for colours)
5.1.3 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 high velocity between train and plane, we focus on the analysis of urban road transportation modes including car, bus, bike and walk. Figures 4(a) and 4(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 Figure 4(c) and Figure 4(d), Figure 4(d) describes the actual distribution of acceleration compared to Figure 4(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.

Besides the features described above, we also extract other prominent features about velocity and acceleration. Table 2 presents the overall features we explored in this experiment.

<table>
<thead>
<tr>
<th>Features</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>85thV</td>
<td>85% percentile velocity</td>
</tr>
<tr>
<td>MaxV2</td>
<td>Second maximum velocity</td>
</tr>
<tr>
<td>MedianV</td>
<td>Median velocity</td>
</tr>
<tr>
<td>MinV</td>
<td>Minimum velocity</td>
</tr>
<tr>
<td>MeanV</td>
<td>Average velocity</td>
</tr>
<tr>
<td>Ev</td>
<td>The expectation of velocity</td>
</tr>
<tr>
<td>Dv</td>
<td>The covariance of velocity</td>
</tr>
<tr>
<td>85thA</td>
<td>85% percentile acceleration</td>
</tr>
<tr>
<td>MaxA2</td>
<td>Second maximum acceleration</td>
</tr>
<tr>
<td>MedianA</td>
<td>Median acceleration</td>
</tr>
<tr>
<td>MinA</td>
<td>Minimum acceleration</td>
</tr>
<tr>
<td>MeanA</td>
<td>Average acceleration</td>
</tr>
<tr>
<td>Ea</td>
<td>The expectation of acceleration</td>
</tr>
<tr>
<td>Da</td>
<td>The covariance of acceleration</td>
</tr>
<tr>
<td>SR</td>
<td>Stop rate</td>
</tr>
<tr>
<td>HCR</td>
<td>Head change rate</td>
</tr>
<tr>
<td>VCR</td>
<td>Velocity change rate</td>
</tr>
<tr>
<td>ACR</td>
<td>Acceleration change rate</td>
</tr>
<tr>
<td>TS</td>
<td>Timeslice type</td>
</tr>
</tbody>
</table>

5.2 Segments expansion

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 (Zheng et al., 2008a, 2008b). In our work, we decided to employ random forest as a model because the works by Stenneth et al. (2011) showed that performance of random forest is better than decision tree in transportation mode identification. After applying the former inference model, we can obtain the predicted transportation modes of segment divided by transition points. In order to understand users’ behaviour better and get high precision of transition points, we put forward two segments expansions to reprocessing predicted segments.

5.2.1 Same segments expansion

Several consecutive segments which were identified as a same transportation mode should be merged into one segment. As illustrated in Figure 5, Figure 5(a) gives the predicted mode sequence and Figure 5(b) gives the result of same segment expansion processing. Precision identification of transportation mode is attributed to the high quality recognition of transition points. This expansion can eliminate extra transition points and apply for the recognition of transition points in future.

5.2.2 Neighbour segments expansion

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. Figure 6(a) gives the predicted mode sequence after the first classification. According to experience, the segment classified as bus mode in this 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 other non-walk 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 6(b) shows the processing result of neighbour segments expansion.

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 preprocessing procedure, the value of velocity and acceleration threshold \( v_{thd}, a_{thd} \) is set to be 1.8 m/s and 0.6 m/s \(^2\) (Zheng et al., 2008b). The variable \( N \) and \( scale \) is set to be 10 and 0.8, respectively. Referring to the situation that most of pedestrians achieve 50 m with normal walk speed in 90 seconds, 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 Zheng et al. (2008a). Figure 7 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 randomised decision trees. Other parameters are default parameters. With regard to the toolkit we used in the experiments, Waikato Environment for Knowledge Analysis (Weka) 3.7 toolkit (Bouckaert et al., 2010) is selected to implement random forest. About 70% of all the segments are trained and the remaining are used for testing.

6.2 Effectiveness of preprocessing step

The preprocessing step divides the GPS trajectory into single mode segments based on transition points collection. We evaluate the effectiveness of our preprocessing procedure by 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 metres, we regard the transition point as a correct inference. As a result, we retrieved almost 90% of the actual transition points from the corresponding GPS dataset.

In our experiment, we compare our trajectory segmentation method with three different segmentation methods including uniform length segmentation, uniform duration segmentation and trajectory segmentation method proposed by Zheng et al. (2008b). For the convenience of description, we abbreviated our trajectory segmentation procedure to TS, trajectory segmentation method of Zheng et al. (2008b) to ZTS, uniform duration segmentation to UDS (every 120 second a segment) and uniform length segmentation to ULS (every 100 metre a segment). Table 3 gives the comparison of different segmentation methods about transition point recognition. The performance of TS outperforms others in both precision and recall of transition points recognition. Despite of similar recall of transition points, we still to want to get high precision of transition points. It is commonsense that the longer a segment is the richer features of its transportation mode a segment will express.

Figure 7 Selecting threshold \( (A_r) \) for ACR (see online version for colours)

Note: ACR is the only feature used in the inference model.

Table 3 Comparison of different segmentation methods about transition point recognition

<table>
<thead>
<tr>
<th></th>
<th>TS</th>
<th>ZTS</th>
<th>UDS</th>
<th>ULS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.48</td>
<td>0.41</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>Recall</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
<td>0.87</td>
</tr>
</tbody>
</table>

6.3 Feature evaluation

Considering the out-of-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 4, 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.

Another, we evaluate entire features in three different combination ways. ACR, 85thV, 85thA and TS make up the combine features. The traditional features 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 new features is the feature set which we explored in this experiment. The results are shown in Table 5, from which the overall accuracy by segment of combine features 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 new features rises from 58.4% to 60% compared the traditional features.

Table 4 Classification features ranking

<table>
<thead>
<tr>
<th>Information gain rank</th>
<th>Single feature rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>Feature</td>
</tr>
<tr>
<td>1</td>
<td>85thV</td>
</tr>
<tr>
<td>2</td>
<td>MedianV</td>
</tr>
<tr>
<td>3</td>
<td>MedianV</td>
</tr>
<tr>
<td>4</td>
<td>Ev</td>
</tr>
<tr>
<td>5</td>
<td>Dv</td>
</tr>
<tr>
<td>6</td>
<td>SR</td>
</tr>
<tr>
<td>7</td>
<td>HCR</td>
</tr>
<tr>
<td>8</td>
<td>ACR</td>
</tr>
<tr>
<td>9</td>
<td>VCR</td>
</tr>
<tr>
<td>10</td>
<td>MaxV_2</td>
</tr>
<tr>
<td>11</td>
<td>85thA</td>
</tr>
</tbody>
</table>

Table 5 Feature comparison

<table>
<thead>
<tr>
<th>Combine features</th>
<th>Traditional features</th>
<th>New features</th>
</tr>
</thead>
<tbody>
<tr>
<td>52%</td>
<td>58.4%</td>
<td>60%</td>
</tr>
</tbody>
</table>

6.4 Mode identification

For transportation mode identification, we evaluate the classification by using two well-known performance measures: Precision and Recall. As described in the previous section, we use random forest classifier as the mode identification classifier. We can achieve an overall accuracy by distance of 89.31%. The accuracy by distance in identifying car, bus, bike and walk is 82.85%, demonstrating a better discrimination about transportation modes than previous researches (Lin et al., 2013; Shin et al., 2015; Witayangkurn et al., 2013; Zheng et al., 2008a).

As shown in Table 6, walk mode identification achieves about 65.90% accuracy while 90.53% 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, train and plane mode identification in matrix remain high scores. Among many reasons, the most important one is that car, train and plane segments hold very long distance with the characteristic of high velocity. However, only a small proportion of car mode is classified as bike mode. For this uncommon result, statistical analysis of classification results indicates that the velocity value of bike and car mode concentrate in the region of 3–6.5 and 14–25 m/s, respectively. Figure 8 gives the distribution of 85thV about car mode. The black curve is the theoretical trend of Gauss distribution curve. In the value slot of 4–14, the total number of car segments cannot reach the edge of black curve. This unoccupied blank implies the small probability of classifying car as bike mode mistakenly. Also, the overall accuracy by distance of bike mode can reach about 80%.

Figure 8 Velocity distribution histogram of driving in Beijing (see online version for colours)
From the misclassification between car and bus, about 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 behaviour. In spite of this phenomenon, the overall accuracy by distance of bus mode can perform 68.08%, which is better than similar researches (Lin et al., 2013; Zheng et al., 2008a).

7 Conclusions

In this paper, learning from raw GPS dataset, we presented a new robust framework for identifying different means of transportation modes including urban road mode (car, bus, bike and walk), rail mode (subway and train) and aircraft mode (plane). 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 (Bolbol et al., 2012; Schuessler and Axhausen, 2009; Zheng et al., 2008a) used. Additionally, the segments expansion procedure is considered after classification. As a result, our work maintains relatively high precision when comparing with previous work (Lin et al., 2013; Shin et al., 2015; Witayangkurn et al., 2013; Zheng et al., 2008a). The overall accuracy by distance of our framework can perform 89.31% in transportation mode identification.

The application of our framework is useful for user behaviour analysis. However, many issues remain open and they are worthy of further study. First, combining the extra transition points is still a challenge work in transportation modes behaviour analysis. We would like to integrate some detail semantic notions into the classification scheme. For instance, providing home and work location, and knowing the route that is most likely taken each day is helpful for further analysis. Secondly, the wholeness of trajectory maybe plays a more significant role in identifying transportation modes instead of each segment individually. We cannot ensure the natural flow of the travel pattern of every participant. So collecting abundant trajectory data is also potential work to do.

Acknowledgements

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


