Trip-Extraction Method Based on Characteristics of Sensors and Human-Travel Behavior for Sensor-Based Travel Survey

HIROKI OHASHI1,†1,a) PHONG XUAN NGUYEN1,b) TAKAYUKI AKIYAMA1,c) MASAAKI YAMAMOTO1,d) AKIKO SATO2,e)

Received: April 11, 2015, Accepted: October 2, 2015

Abstract: A novel method for extracting “trip periods,” i.e., periods in which a person travels, from continuously collected sensor data, called a “trip-extraction method” hereafter, is proposed to make a sensor-based travel-behavior survey possible. There are mainly two drawbacks in previous studies that detect “stay periods,” i.e., periods in which a person stays within an area, by using the boundary of a “stay area,” i.e., an area in which a person stays and then regard the rest of the periods as trip periods: false positives caused by GPS-positioning errors and false negatives caused by short-distance trips within the boundary. This study solves these problems by using novel features that are effective even in the case where the GPS-positioning error is large and by classifying every single piece of GPS data into either trip periods or stay periods not on the basis of the stay-area boundary but on the newly proposed features. An experimental evaluation showed that the precision of the proposed method was 89.4%, which is much higher than that of conventional methods.

Keywords: intelligent transport systems (ITS), travel behavior survey, smartphone, machine learning, hidden Markov model (HMM)

1. Introduction

Understanding travel demand in cities is very important for efficient transportation planning. Manual methods such as the “household travel survey” and the “person-trip survey” [1] have been conventionally used to grasp the travel demand in cities. However, these surveys are very expensive to conduct since they are based on a manual questionnaire. They have thus typically been conducted only once a decade. GPS-based systems are expected to replace costly manual surveys [2], [3], [4]. GPS-positioning data, which are easily collected with smartphones these days, potentially include the same kinds of “trip information” as those collected in the conventional manual surveys, namely, origin and destination, departure and arrival times, purpose, and transportation mode of each trip. It is troublesome, however, for users if they need to keep a log of trips every time they start and end a trip. Thus, it is necessary to develop a method for automatically extracting trip information. In our previous study [6], [7], a method for automatically classifying modes of transportation on the basis of a smartphone’s sensor data was proposed. In the present study, a novel method for accurately extracting “trip periods,” i.e., periods in which a user travels, is proposed. This method is called a “trip-extraction method,” hereinafter. Origin, destination, departure time, and arrival time are detected by using this method.

Conventional methods [8], [9] tried to extract trip periods with the following two steps. First, they detect “stay periods,” i.e., periods in which a user stays within an area, by using the boundary of “stay area,” i.e., an area in which a subject stays, and then regard the rest of the periods as trip periods. However, there are mainly two drawbacks to these methods: false positives caused by GPS-positioning errors and false negatives caused by short-distance trips within the stay area. The detection of the stay area does not work well when some GPS data are positioned outside a boundary, even though a user stays within the boundary, due to a GPS-positioning error. Short-distance trips within the boundary cannot be extracted since the periods of the trips are regarded as stay periods.

A novel trip-extraction method is proposed to address the aforementioned problems. The proposed method solves the two problems by using novel features that are effective even in the case where GPS-positioning error is large and by classifying every single piece of GPS data into either trip periods or stay periods not on the basis of the stay-area boundary but on the newly proposed features. GPS-positioning accuracy is an example of the features. Most of the time, people are inside a building when they...
are in stay periods. In this case, GPS-positioning accuracy tends to decrease since GPS signals are often blocked or reflected by the walls and ceilings of the building, while the accuracy tends to be high when people travel outside. With this idea of using GPS-positioning accuracy as a feature, the larger a GPS-positioning error is, the more accurately the proposed method can detect a stay period, while the accuracy of conventional methods becomes low in this situation. In addition to the feature selection, the proposed method uses a state-transition model to eliminate improbable transitions between trip and stay periods that can be caused by the above-mentioned data-by-data classification.

The rest of this paper is organized as follows: Related works of the trip-extraction method are reviewed in Section 2. Then a novel trip-extraction method is proposed and explained in detail in Section 3. Section 4 introduces an evaluation result and the discussion on the result. Finally, conclusions are given in Section 5.

2. Related Work and Problem Statement

2.1 Related Work

The most general and frequently used idea for trip-extraction methods contains two steps: detect stay periods and then regard a period between consecutive stay periods as a trip period. Since the second step is almost automatically accomplished if the first step is correctly done, most studies focused on how to correctly detect stay periods. Methods for classifying a mode of transport sometimes include “stop” as their target of classification [10], [11], [12]. However, the “stop” assumed there is, mostly, not a stay at a destination, which is the target of this study (so called “stay periods”), but a temporary stop during travels such as the time spent waiting for a traffic light. While a GPS signal received during a “stop” time is relatively stable, that in stay periods frequently fluctuates a lot. Therefore, the methods for classifying a mode of transport are not applicable to the trip-extraction problem.

Among the studies that focus on extracting stay periods, many of them rely on a rule-based method with fixed thresholds. Quan-nan Li et al. [13] regarded a GPS point to be a “stay point”, which constitutes a stay period if the time difference between consecutive pieces of GPS data is larger than 30 minutes and the distance between the pieces of data is less than 200 meters. Ming Li et al. [8] looked at not only two consecutive data, but also groups of data. A group of data was regarded as a stay period if the data fall within a radius of 120 meters and the time difference between the first piece of data and the last was larger than 10 minutes. Wolf et al. [14] reported that 120 seconds was best for the threshold of time difference. Wang et al. [15] used call detail record (CDR) data for a sensor-based travel behavior survey. In their study, 1 km, which is much longer than in the GPS-based studies, was used as a threshold of distance since CDR data were coarser grained. As seen from these examples, it is difficult to define an appropriate threshold that works generally. Ming Li et al. [8] also suggested that those thresholds must be changed in accordance with regional characteristics.

In addition to the difficulty of defining an appropriate threshold, the methods based on fixed thresholds often suffer from errors because of outliers. Location information obtained by using GPS signals can fluctuate a lot if users are surrounded by things that prevent GPS signals from directly reaching the users. Botone et al. [16] proposed a pre-processing method to get rid of data that have a low accuracy in location calculation. Witayangkurn et al. [17] proposed a method that uses a fixed threshold in combination with an outlier-detection technique. In their study, a group of data was regarded as a stay period if the distance between the first piece of data and the last was less than 196 meters and the time difference between the two pieces of data was larger than 14 minutes. A point whose normal variate on the distance from the centroid is larger than 2.6 is eliminated as an outlier. A good accuracy, namely, a precision of 92.4% and a recall rate of 90.5%, was achieved in their study. However, this method mainly has three drawbacks. First, it cannot, in principle, detect short stay-periods, which are less than 14 minutes. Second, it would wrongly regard a round trip, i.e., the origin and the destination of the trip are the same, as a stay. Last, the outlier elimination would not work well if GPS positioning error is averagely large. In this case, the normal variate becomes smaller even if the positioning error is large (like over 196 meters). An outlier, which should belong to a stay period, will be regarded as the start of a trip period if it is not eliminated.

Ohashi et al. [9] proposed an improvement to these fixed-threshold-based methods. In Ref. [9], a method of using the Mahalanobis distance, instead of the Euclidean distance, was proposed. This was meant to dynamically adjust the shape and the size of a boundary for detecting stay areas in accordance with the distribution of the GPS data collected at that time. By using the Mahalanobis distance, a boundary automatically becomes larger if the GPS positioning error is large. A precision of 81% and a recall rate of 62% were achieved in this study. However, even this method cannot detect short-distance trips within the boundary. In particular, when the GPS positioning error is large, it becomes more difficult to detect short-distance trips since the boundary also becomes larger. This is one of the drawbacks of this method. It is necessary to develop a new method that can correctly detect both stay periods of which GPS positioning error is large and trip periods of which distance is short.

2.2 Problem Statement

To summarize, the boundary-based conventional methods reviewed in Section 2.1 have two main drawbacks: they easily suffer from false positives caused by GPS-positioning errors, and they often fail to detect short-distance trips that are included within stay areas, especially when the boundary is set larger in order to become robust against GPS-positioning error. In this study, we aim to propose a novel trip-extraction method that overcomes these drawbacks without relying on a boundary.\footnote{This paper is an extension of Ref. [18]. A new method for modification based on geographic information system (GIS) is newly proposed, and more experimental data are added for evaluation.}

The definitions of two words are given to concretely specify the target of this study. “Main trip” is defined as a sequence of movements that correspond to one purpose, such as “go to work,” “go home,” and “shopping,” while “sub trip” is defined as
Movement that corresponds to one mode of transportation, such as "walking," "bicycle," and "train" (see Fig. 1). Sub trips do not end until subjects reach their destination or a place for changing transportation modes, i.e., one sub trip is not divided into two by the time spent waiting for a traffic light, congestion, etc. One main trip can be composed of multiple sub trips. For example, the trip of a person walking from home to the station, taking the train, and walking to the office is regarded as one main trip. In this case, one main trip corresponding to "go to work" contains three sub trips, namely, "walking," "train," and "walking." The proposed method focuses on extracting main trips. Therefore, one trip period corresponds to one main trip including the time spent waiting to change the transportation mode, while one stay period corresponds to the period between two consecutive main trips. Hereinafter, the word "trip" means "main trip" unless another meaning is explicitly explained.

3. Proposed Method

3.1 Overview

In this study, we first classify every single piece of GPS data into either a stay point, which is an element of stay periods, or a trip point, which is an element of trip periods, on the basis of newly proposed features (details are given in Section 3.2) and the previous point (details are given in Section 3.3). Then, trip periods (stay periods) are extracted by combining consecutive trip points (stay points) as shown in Fig. 2. Last, some periods are modified on the basis of GIS information and some rules derived from human-travel behaviors (details are given in Section 3.4).

3.2 Features

This study proposes to use the following five features, which are derived on the basis of the characteristics of sensors [(i) and (ii)] and characteristics of human-travel behavior [(iii) to (v)];

(i) accuracy of GPS positioning
(ii) time difference between the point in question and the previous (and next) point
(iii) similarity of heading direction
(iv) velocity
(v) variance of acceleration data

The features are calculated for each piece of GPS data.

Table 1 summarizes which value each feature is likely to be in stay periods and trip periods, and Fig. 4 represents the actual distribution of the features. The detailed explanation for each feature is given below. The accuracy of GPS positioning (i) becomes lower in stay periods. In stay periods, people are most likely in a building. In this case, the accuracy of GPS positioning tends to decrease due to the presence of large buildings or indoor environments. The velocity (iv) is likely to be low in stay periods because people are typically static or moving at a slow pace within buildings. Conversely, during trips, people move at higher speeds due to the nature of their transportation and activities.

The time difference (ii) is calculated as follows.

\[
time\ \text{difference} = \frac{t_{i+1} - t_i}{2} = \frac{t_{i+1} - t_{i-1}}{2},
\]  

where \( t_i \) denotes the time-stamp of the \( i \)th piece of GPS data. The similarity of heading direction (iii) is calculated as follows by using GPS data.

\[
similarity_i = \frac{1}{e_i - s_i + 1} \sum_{j=s_i}^{e_i} \frac{\Delta_j \cdot \Delta_j'}{\Delta_j ||\Delta_j||},
\]  

where \( s_i \) is the smallest index that satisfies \( t_{s_i} \geq t_i - \delta t \) and \( e_i \) is the largest index that satisfies \( t_{e_i} \leq t_i + \delta t \). \( \delta t \) is set to be 1 minute in this study. The definition of \( \Delta_j \) is shown in Fig. 3. The variance of acceleration data (v) is calculated by using the acceleration data between \( t_i - \delta t \) and \( t_i + \delta t \).
to decrease since the walls and ceilings of the building often reflect GPS signals. Conventional methods often suffer from mis-detection due to this inaccuracy since they basically try to detect stay areas on the basis of this unstable GPS positioning data. This study does not just avoid mis-detection but makes the inaccuracy one of the good features in stay periods. As a result, it can achieve a robust classification. The time difference (ii) becomes larger in stay periods. As mentioned above, people are most likely in a building in stay periods. It becomes difficult to catch a GPS signal if a GPS receiver is inside a building since walls and ceilings can block the signal. Thus, the time difference tends to be larger.

The similarity of heading direction (iii) becomes higher in trip periods because people do not frequently change their moving direction when traveling. This feature significantly contributes to the robust detection of short-distance trips. The velocity (iv) becomes, obviously, higher in trip periods compared with stay periods. The variance of acceleration data (v) becomes larger in trip periods. It becomes larger while people are moving since it is related to the vibration of an acceleration sensor. The vibration tends to be larger when people are walking, running, bicycling, and even when using motorized vehicles.

3.3 Classification Method
In this study, each GPS point is classified into either a trip point or a stay point to achieve a fine-grained trip-extraction. As a result, it becomes possible to extract trip periods no matter how short they are. However, the point-by-point classification sometimes causes mis-classification due to outliers.

The ergodic hidden Markov model (HMM) is used to deal with this problem. HMM is effective for reducing mis-classification in the case where some amount of outliers is expected. By appropriately adjusting state-transition probabilities, it can suppress unlikely transitions. The ergodic HMM is a type of HMM, and it deals with a recurrent state-transition model. Since GPS data are supposed to have outliers and states are recurrent, the ergodic HMM is suitable for this study’s setting.

Feature vectors that consist of the five features explained in the previous section are regarded as “observed signals,” and states that are either a trip period or a stay period are regarded as “hidden states.” In this formulation, the goal is to estimate a sequence of hidden states by using observed signals.

The state-transition model used in this study is shown in Fig. 5.

\[
A = \begin{bmatrix}
q_{11} & q_{12} \\
q_{21} & q_{22}
\end{bmatrix}
\]

The values in the transition matrix \(A\) are empirically determined.

Fig. 5 State-transition model. The values in the transition matrix \(A\) are empirically determined.
difficult to collect training data from all of the persons (e.g., travel behavior survey for city planning), it is better to empirically determine the values in transition matrix $A$.

The posterior probability of hidden states is calculated by using the following equation.

$$p(s_i|x_i) = \sum_{s \in S} p(x_i|s)p(s).$$ (3)

$$p(x_i|s_i) = \frac{\sum_{m=1}^{M} N(x_i|\mu_{s_i,m}, \Sigma_{s_i,m})}{\sum_{s_i \in S} \sum_{m=1}^{M} N(x_i|\mu_{s_i,m}, \Sigma_{s_i,m})},$$ (4)

$$p(s_i) = \sum_{s \in S} p(s_i|s)p(s).$$ (5)

$$S = \{\text{trip, stay}\},$$ (6)

where $x_i$ denotes a $5$-dimensional feature vector at time $t_i$ and $s_i \in \{\text{trip, stay}\}$ denotes a state at time $t_i$. As is shown in Eq. (4), the conditional probability $p(x_i|s_i)$ is calculated using a Gaussian mixture distribution. $M$ in Eq. (4) denotes the number of Gaussian distributions used in the Gaussian Mixture Model (GMM). The parameters $\mu_{s_i,m}$ and $\Sigma_{s_i,m}$ denote a mean vector and a covariance matrix of $m$th Gaussian distribution of state $s_i$ respectively. They are preliminarily learned using Expectation-Maximization (EM) algorithm. In this study, a part of the experimental data (about 20% of the total data) that are randomly chosen were used to determine these parameters. The conditional probability $p(s_i|x_{i-1})$ is given by the transition matrix $A$.

Each piece of GPS data is classified into a trip(stay) point if the state “trip (stay)” has a higher probability in Eq. (3). Then, consecutive trip (stay) points are combined together and corresponding time periods are extracted as trip (stay) periods (Fig. 2).

### 3.4 Modification

Since extremely short periods are likely to be a false detection of trips, the classified periods are modified accordingly. Different criteria are used to modify trip periods and stay periods.

If a trip period is extremely short (less than 60 seconds, and the number of pieces of data is less than 20), it is regarded as a mis-detection caused by outliers and combined with the previous and the next periods. Sixty seconds corresponds to about 67 meters if a person walks at 4 km/h, which is the average speed of walking. It is very rare to have a main trip shorter than 67 meters. Even if a person has such a short trip, the number of pieces of data usually becomes larger if it is really a trip. Therefore, this procedure does not disable detecting short-distance trips.

If a stay-period is extremely short (less than 120 seconds), it is regarded as a mis-detection or a short-time stay during trips such as the time spent waiting for a traffic light or time spent changing trains, and it is combined with the previous and the next period.

In addition to these modifications, another modification for dealing with the time spent waiting for a train is applied in order to overcome the problem reported in the previous research[18].

It was reported that since the time spent waiting for a train sometimes becomes longer than 120 seconds, one main trip is wrongly divided into two in this case. This study avoids this error by using GIS information. If a stay period is relatively short (less than 10 minutes) and the location is close to a station, it is regarded as the time spent waiting for a train and combined with the previous and the next periods. The modification is processed as follows.

(1) Locations of stations are extracted from map information and stored in a database in advance. OpenStreetMap[19], which is a free editable map of the world, is used in this study.

(2) The geometric median of a stay periods is calculated by using the following formula

$$gm_k = \arg \min_{p \in \mathbb{C}} \sum_{i \in C_k} ||p_i - p||,$$ (7)

where $p_i$ denotes a position of the $i$th piece of GPS data, $||p_i - p||$ denotes the distance between $p_i$ and $p$, and $c_k$ denotes a set of GPS data that belongs to the $k$th stay period. The geometric median, instead of a simple centroid, is used since it is more robust against outliers.

(3) $d_{\text{min}}^k = \arg \min_{s \in S \mid \sigma_k} ||gm_k - st||$, which is the distance between the geometric median and the closest station, is calculated. $S$ represents a set of stations stored in a database in step (1).

(4) The stay period is regarded as the time spent waiting for a train and combined with the previous and the next periods if $d_{\text{min}}^k$ is less than a threshold and the period is shorter than a threshold.

An image of the modification is shown in Fig. 6.

### 4. Evaluation

#### 4.1 Experimental Data

Table 2 shows the conditions in the evaluation. The data col-
lection experiment was not conducted throughout the whole experimental period. Each subject collected data only for a certain period, averagely one week, within the experimental period; the first subject started the experiment on Jan. 17th, 2014 and finished it about a week later, while the last subject started the experiment around the end of Mar., 2015 and finished it on Apr. 6th, 2015. This is why the total number of days during which data were actually collected was 50. The average number of trips per person was 3.24 (=162/50). The experimental data contained GPS data and acceleration data. The sampling rate for collecting data was 1 Hz for GPS data and 50 Hz for acceleration data. The data were collected by using an Android smartphone in this experiment, but data collected with any other devices could be used as long as they were collected at a similar sampling rate.

In the experimental period, the subjects always carried a smartphone during their daily routine. The data-collection app was started before the first trip of the day and terminated after the final trip of the day. The app was kept running even when the subjects were not moving if more trips in the day were planned. Notes on all of the trips during the experiment were taken to obtain the ground truth (see Fig. 7 for an example of the notes). Notes on transportation modes and purposes were also taken for the further development of the method in the future. Since it was found that one of the biggest problems in the previous method [9] is a “false negative” in detecting short-distance trips, more data including short-distance-trip data were used for this evaluation.

4.2 How to Evaluate the Proposed Method

The accuracy of separating main trips was evaluated in terms of precision and recall rate. The data that contained no GPS co-ordinates were not evaluated. Precision indicates the preciseness of detecting trip starts and ends. It is therefore lower if a certain time period that does not correspond to a trip is detected as one corresponding to a trip. Recall rate indicates the completeness of the detection. It will therefore be lower if a time period that corresponds to a trip is not detected as one corresponding to a trip.

To apply the proposed method to a travel-behavior survey, which is the goal of this study, precision is more important than recall rate. This is because even if the recall rate is low, it is still possible to conduct a travel-behavior survey if a large amount of sample data can be acquired. Precision is critical since it affects the reliability of the survey.

The trip separation is regarded as successful if the detected trip satisfies two conditions: the difference between the estimated arrival time and the arrival time written in the notes is less than 10 minutes, and the difference between the estimated departure time and the departure time written in the notes is less than 10 minutes. It is reasonable to accept the difference of 10 minutes for appropriate evaluation because acquiring precise ground truth data is very difficult. Since the notes were manually taken, sometimes the subject may forget to take notes or be in a too big hurry to take notes, and may result in filling out the notes afterwards. In this case, the number tends to be a round number (e.g., if a subject actually arrived at 8:52, the arrival time written in a note may be 9:00). In addition, there are some cases where it is difficult to clearly define when a new trip is started or ended. For example, if a subject commutes by car, arrival time to an office can be the time when the subject passed through an entrance gate of the office site, the time when the subject parked the car at the office’s parking lot, the time when the subject entered a building, or the time when the subject sat down at a desk. Because of the above-mentioned difficulties, ground truth data themselves can sometimes contain error of up to around 10 minutes. In order to evaluate the method appropriately under this situation, it is appropriate to accept the difference of 10 minutes.

4.3 Result and Discussion

Table 3 shows a comparison of accuracy between the proposed method and the two latest conventional methods on the same dataset. One hundred and sixty two main trips were col-

<table>
<thead>
<tr>
<th>Item</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of actual data collection day</td>
<td>50</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>8</td>
</tr>
<tr>
<td>Number of trips</td>
<td>162</td>
</tr>
<tr>
<td>Place</td>
<td>Tokyo, Japan</td>
</tr>
<tr>
<td>Device</td>
<td>Samsung Galaxy SIII (Android OS 4.1.2)</td>
</tr>
<tr>
<td>Sensor</td>
<td>GPS (Sampling rate : about 1 Hz) Accelerometer (Sampling rate : about 50 Hz)</td>
</tr>
</tbody>
</table>

Fig. 7 Example of notes concerning the ground truth. One row corresponds to one sub trip. Notes were originally taken in Japanese but were translated (with no modifications to content) for easy understanding in English.
Table 3 Evaluation result.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>89.4% (126/141)</td>
<td>77.8% (126/162)</td>
</tr>
<tr>
<td>Conventional (outlier detection) [17]</td>
<td>60.9% (78/128)</td>
<td>48.1% (78/162)</td>
</tr>
<tr>
<td>Conventional (dynamic boundary) [9]</td>
<td>60.6% (94/155)</td>
<td>58.0% (94/162)</td>
</tr>
</tbody>
</table>

(a) Small fluctuation in GPS positioning  
(b) Large fluctuation in GPS positioning

One of the main causes of the false positives, i.e., errors in precision, was that stay periods were mistaken for trip periods because the features of the stay periods were similar to those of trip periods (7 out of 15). Although a subject was staying in a building, the accuracy of GPS positioning was high, the time difference between consecutive points was large, and the similarity of heading direction was relatively high (See Fig. 9 (a)). These false positives can be corrected by integrating the conventional idea of using a stay-area boundary into the proposed method when the accuracy of GPS positioning was high. One of the conventional methods [17] could actually detect the stay period shown in Fig. 9 (a) correctly. Another major cause was an error in the modification step (3 out of 15). This error occurred especially when a subject did a short-time shopping near a station. As explained in Section 3.4, a short-time stay-period near a station is not regarded as a stay at the destination but as the time spent waiting for a train. They are therefore modified to trip periods. An idea to avoid this error in modification is to use acceleration data in the modification step as well. While the vibration calculated by using acceleration data is expected to be smaller when a subject is waiting for a train, it is expected to be larger when a subject is doing shopping. The modification step needs to modify only the former case. The last major cause was that a subject walked around during a stay period (2 out of 15). Figure 9 (b) shows an example of this error. In this case, a subject actually walked around in a park. However, the subject did not regard this walk as a trip because it was not a travel from one place to another with some purpose but he stayed in one location, namely the park. One possible modification is to use GIS information. If subjects were found to be walking within a special area such as a park or a play ground,
the period during which the subjects stayed in the area can be regarded as a stay period. Note that, however, such periods should not always be regarded as stay periods but sometimes be regarded as trip periods. It is important to decide referring to the purpose of an analysis.

To clarify the contribution of each feature, an accuracy was calculated after getting rid of each feature one by one. Figure 10 shows the result. As shown in the figure, the contribution of the velocity and the variance of acceleration data, which are commonly used in previous studies, is relatively low since the accuracy does not decrease very much by getting rid of these features. On the contrary, the contribution of the similarity of heading direction, time difference, and GPS-positioning accuracy is high. If any one of the features are gotten rid of, a precision decreases to below 80%. This result indicates that the proposed features in this study are properly selected.

To clarify the effect of the GPS sampling rate on the accuracy, a simulation was conducted. Before the result is shown, the characteristics of GPS data collection using Android API are explained. It is not possible to directly specify the sampling rate of GPS data via the Android API. The only thing that can be done for controlling the frequency of collecting GPS data is to specify the maximum interval of activating a GPS receiver. A detailed explanation is given referring to Fig. 11. It is possible to set the maximum interval between \( t_2 \) and \( t_3 \) via the Android API. If the maximum interval is set to T, the length of \( (t_1-t_2) \), which corresponds to the time periods when a GPS receiver is inactive is always shorter than T. However, it does not necessarily mean that a GPS datum is actually collected every T. Even though a GPS receiver is active, it might fail to catch a GPS signal. The length of periods when a GPS receiver is active (i.e., \( (t_2-t_1) \) and \( (t_1-t_3) \)) is not fixed, but according to our experience, the length tends to be about 30 seconds in our experimental setting. While a GPS receiver is active, GPS data were usually collected every second (, which means that the sampling rate of GPS data was always around 1 Hz when a GPS receiver is active). Considering above-mentioned characteristics, the simulation was designed to simulate the cases where the maximum interval is set to be different values. In each of the cases, the GPS data that belonged to the periods when a GPS receiver was inactive were eliminated and not used for trip separation. The length of periods when a GPS receiver is active was fixed to 30 seconds in the simulation. Figure 12 shows the result. As is expected, both precision and recall rate decrease as the interval becomes larger. This is mainly because trips are mistakenly separated by the period when a GPS receiver is inactive. Among the five features used for trip separation, the time difference is greatly influenced by this period. Since the contribution of this feature to the overall accuracy is large as shown in Fig. 10, the influence has a large impact on the accuracies. It may be possible to install a modification process to address the problem of accuracy decrease when the maximum interval is set to be larger. For example, since the value of the maximum interval is preliminarily known, the trips separated by the stay periods of that duration can be combined.

5. Conclusion

A novel method for automatically extracting trips on the basis of continuously collected sensor data was proposed. While conventional methods based on detecting stay areas with a boundary often suffer from false negatives caused by short-distance trips within the boundary and false positives caused by GPS-positioning errors, the proposed method was able to correctly extract the trips by using robust features derived from the characteristics of sensors and human-travel behavior. Since large GPS-positioning error, which is the main cause of the false positives in conventional methods, is used as one of the good features of stay periods, the larger the error is, the more accurately the proposed method can distinguish stay periods. A problem caused by outliers in classifying each GPS point into either a stay point or trip point was suppressed by using a state-transition model based on a characteristic of the travel-behavior of people. The model is mathematically formulated by using the ergodic HMM. This formulation has enabled a fine-grained trip-extraction.

The evaluation of the proposed method showed a promising result of 89.4% in precision. The proposed method achieved a much higher accuracy compared with conventional methods. In particular, the proposed method makes it possible to correctly classify short-distance trips that conventional methods have failed to do
Future work includes overcoming the problem of mistaking stay periods for trip periods when the features of the stay periods are relatively similar to those of trip periods by integrating the conventional idea of using a stay-area boundary into the proposed method. The next step toward a sensor-based travel survey is integrating a transport-mode-classification technique and evaluating the accuracy of detecting sub trips. In addition, the development of a trip-purpose estimation method is the last biggest missing piece.

References


Hiroki Ohashi received his bachelor’s and master’s degrees from Kyoto University in 2009 and 2011, respectively. He joined central research laboratory of Hitachi Ltd. in 2011 and has worked for Hitachi’s European R&D group since 2015. His main research interests are applied machine learning and artificial intelligence. He is a member of the institute of electronics, information and communication engineers.

Phong Xuan Nguyen was born in Vietnam and has received numerous awards in national information technology competitions since 2000. He received his bachelor’s degree in business administration from Coventry University in 2010. After a year of working, he continued his education in Carnegie Mellon University and received a master of science degree in information technology in 2013. He then joined Hitachi R&D group, and has worked for Center for Technology Innovation – Systems Engineering since 2015. He is currently interested in making machine learning practical solutions in different industries.

Takayuki Akiyama received his bache lor’s and master’s degrees from the University of Tokyo in 2006 and 2008, respectively. He joined central research laboratory of Hitachi Ltd. in 2008 and has worked for Center for Technology Innovation – Systems Engineering since 2015. He is engaged in a research of human flow, traffic flow and traffic simulation using sensor data.

Masaaki Yamamoto received his B.E. and M.E. degrees from Mie University, Japan in 2003 and 2005, respectively. In 2005, he joined Central Research Laboratory, Hitachi, Ltd. He is now a Ph.D. student at Keio University. He has been engaged in a research on wireless communication systems. He received the Niwatakayanagi Award 2006 of ITE and the Best Paper Award of IEICE Transactions on Communications (Japanese Edition) in 2013.
Akiko Sato received her bachelor’s and master’s degrees from the University of Tokyo in 1996 and 1998, respectively. She joined central research laboratory of Hitachi Ltd. in 1998 and has worded for R&D Center of Hitachi Asia, Ltd. Since 2015. She is engaged in a research on sustainable urban mobility.