Trip Reconstruction and Transportation Mode Extraction on Low Data Rate GPS Data from Mobile Phone

Apichon Witayangkurn, Teerayut Horanont, Natsumi Ono, Yoshihide Sekimoto and Ryosuke Shibasaki

Institute of Industrial Science, The University of Tokyo, Komaba, Tokyo 153-8505, Japan
{apichon, teerayut}@iis.u-tokyo.ac.jp, {nono, sekimoto, shiba}@csis.u-tokyo.ac.jp

Abstract

With the advancement of mobile phone in functionality, they are increasingly being used as sensors for people mobility analysis in various areas such as a location-based service and urban planning. The information of people's travel trip played an essential role in urban analysis especially in transportation planning. In this paper, a framework based on supervised learning method is proposed to reconstruct user's trip information from low data rate GPS data from mobile phone device. Our approach consisted of four steps including stay point extraction with outlier detection and removal, trip segmentation based on change point detection, transportation mode extraction using inference model and the last step, segment merging to combine multiple small and uncertain segments. Random Forest classifier with Bootstrap aggregating is used for classifying transportation mode including stay, walk, bike, car and train. GIS information such as spatial train network and spatial road network are employed and used for calculating spatial features to improve segmentation and classification effectiveness. We also developed web-based trip visualization with Google map for verification and labeling. We evaluated the framework using the GPS data collected by 100 users over one month period with 5 minutes interval of GPS data. As a result, outlier detection was effectively able to remove

noises and increase the accuracy of stay point extraction and segmentation. Furthermore, our approach with utilizing of spatial features could achieve over 87.80% for inferring transportation modes and more than 97.76% for train mode. Finally, we were able to accurately reconstruct user's trip chains including basic activity, transportation mode and period of each segment.

1. Introduction

The increasing popularity of mobile phones embedded with positioning function such as GPS, is allowing users to acquire their own locations easily and also leading to collect large scale user's trajectories of multiple mobile phone users. In pervasive computing and context awareness, the analysis of human's trajectories and mobility of people is an important research area to understand and discover knowledge about the activities of people such as favorite places, daily activities and transportation they used. By combining the activities of multiple users based on some criteria such as temporal data and geographic location, it provides empirical knowledge of common activities or behaviors shared by a group of people. For example, people in area A usually take a train to area B, and it mostly crowded in the morning time. Moreover, it delivers real dynamic data rather than static data like survey data. Person trip surveys are particularly valuable data for urban analysis and planning. They have been taken manually by telephone interviews and questionnaires for a long period of time which could take up to three years, and it requires lots of budgets. Also, those kinds of survey data are normally collected after the trips happened which then resulting in lower accuracy comparing with real-time GPS trajectories. Hence, to provide an alternative solution, trip reconstruction from GPS data is a suitable option.

For trip reconstruction, it involves with trip segmentation technique and transportation mode extraction. There are existing techniques for trip segmentation and transportation mode extraction reported such as [1], [2] and [3]. The reported techniques are mostly relied on supervised learning models and can be separated into two main approaches. The first approach is that, from GPS data, features are calculated directly at point level such as speed and acceleration of each GPS point and then apply classification model to inferring transportation mode for each point. By the time sequences, the GPS points with same transportation mode are grouped together into a same segment and reconstructed a trip. Instead of processing

at point level, the second approach processes data at segment level. It starts by segmenting GPS point into walk and non-walk segments using speed features, and then for non-walk segment, they are performed classification to infer transportation mode. Those approaches have reported the accuracy of more than 90% for transportation mode extraction. However, the existing approaches are only tested and evaluated on very fine data rate of GPS data such as one point/second or one point/minute at maximum which is rather difficult in the real world application especially on mobile phone which has limitation on battery. To obtain positioning data in every second or minute, it consumes a lot of power and drains battery quickly. In our study, we instead focused on low data rate GPS data such as five minute data rate which is more relatively suitable to the real world application since there is a mobile operator in Japan providing such positioning data from mobile phone at five minute data rate without interrupting users.

We aim to develop a framework for reconstructing user's trip including transportation mode extraction on low data rate GPS data. The trip segmentation approach in our framework is based on a change point-based segmentation method proposed by [2], however; we have adapted it to support low data rate by adding additional steps and also including new features. Stay point extraction together with outlier detection and removal are applied to GPS data to separate user's trajectories into two main types: Stay and Move. The only Move segments are processed for further segmentation steps. Geographic Information System (GIS) is an important part for distinguishing user transportation mode. Train network and road network polygon are used to calculate features to infer train and car mode. In order to validate the results, we developed a web-based visualization displaying user trajectories on Google Map which is understandable easily. Together with that, we provided labeling tool based on segmentation for labeling each segment of each trip. In summary, we proposed a framework trip reconstruction on low data rate GPS data and this paper is the first to address low data rate issues. The contribution of the paper lies in five aspects:

- We proposed a framework for trip construction on low data rate GPS data consisted of four steps: stay point extraction, change point extraction, determining transportation mode and segment merging.
- We introduced an outlier detection and removal technique adding to stay point extraction to remove noise data as well as to increase the extraction accuracy.

- We used GIS data including Train network and Road network and proposed two spatial features for segmentation and extracting transportation mode: a percentage of points in train line and a percentage of points in road.
- We introduced buffer technique and spatial index for fast searching GIS polygon during features calculation.
- We developed web-based trip visualization with Google map for labeling and verification.

2. Related work

Mining the trajectories of people from GPS data both from positioning device and mobile phone has become an attractive research area during the past years. Most of studies have been focused on extracting meaningful place of people [4][5], understanding people moving pattern [6], and predicting movement of people [4]. Trip reconstruction from GPS data is another important topic for urban and transportation planning. Chung et al. [1] reported the work on a trip reconstruction tool using GPS portable device for personal survey. One point/second data rate was used for collecting positioning data and users were required to carry a dedicated GPS device all the time which would be very difficult for large scale implementation. Our work is different in that we focus on using mobile phone as positioning device instead of using GPS portable device. We also emphasize on low data rate such as one data per five minutes which consume less battery. Moreover, it is rather possible for deployment in the real world application without interrupting user usages. In [2,8], the authors proposed a change point-based segmentation method for determining transportation mode using a novel set of machine learning features and classification; however, it did not consider train as a transport mode and also the defined features did not well perform classification in low data rate GPS data. For our approach, we add a stay point extraction as the first level segmentation before processing a change point detection to split trajectory data into two types: STAY and MOVE. The only MOVE segments are forwarded to the next step.

For stay point extraction, it used maximum distance and minimum time duration as criteria for detecting stay points [4][5]. In addition to that, we incorporate an outlier detection and removal to standard process to remove noises or far-distance point comparing to their neighbors. For determining transportation mode, there are many features proposed by [1,2,3]; howev-

er, they did not effectively infer transportation mode on low data rate GPS. Hence, we introduced two new features: a percentage of points in train line and a percentage of point in road in order to increase the accuracy of mode classification. Even though, Stenneth et al. [3] was the first to propose of using spatial train network as a classification feature, they used closeness distance to closest rail line for each GPS points as a feature which required a lot of calculation especially when there are a large number of train lines. Instead, we used a buffer technique together with spatial index for searching points in polygon which was very fast and then calculated the percentage of points located in train network and road network. Those two features become very important for identifying train mode and car mode. Inspired by visual analytic tools for analysis of movement data by Andrienko et al. [7], we developed a web-based trip visualization to validate the results by displaying trips on Google map. Additionally, we provided labeling function for adjusting label data as well as labeling new data. In summary, the work in this paper aims to address issues on low data rate GPS to effectively reconstruct trip and determine transportation mode for each trip segment. Table 1 summarizes the related works that uses GPS.

Table 1. Related work with GPS data

	Class	**Source**	**Data Rate**	**Duration**	**Users**
[1]	Car, Bus, Bike, Walk	GPS Devices	1/sec	1 months	1
[2,8]	Car, Bus, Bike, Walk	GPS Devices	1/2sec	10 months	65
[9]	Car, Bus, Bike, Walk	GPS Devices	1/sec	6 months	45
[3]	Car, Bus, Train, Bike, Walk, Stay	Mobile Phone	1/15sec	3 weeks	6
Our	Car, Train, Bike, Walk, Stay	Mobile Phone	1/5min	1 months	100

3. Overall framework

We described an overall framework for trip reconstruction. As shown in Figure 1, it consisted of four main steps: stay point extraction with outlier detection, change point detection, determining transportation mode and

merging segments. The right side diagram showed input and output passing through each step. In the first step, GPS logs are used as an input for extracting stay points in order to create a trip with two types of segments: Stay segment and Move segment. For Move segment, it contains only commuting points and can have multiple segments if involved with multiple stay points. For the next step, change point detection takes Move segments of the previous step as an input to split a segment into Walk and Non-Walk segments. For Non-Walk segment, transportation mode classification is applied to determine the mode of each Non-Walk segment. In the last step, small segments in term of distance and time period are considered being merged with previous or next segment. Moreover, for consecutive segments that have same transportation mode, they are also considered being merged into one segment. The detail descriptions of each step are expressed in section 3.1, 3.2, 3.3 and 3.4 respectively.

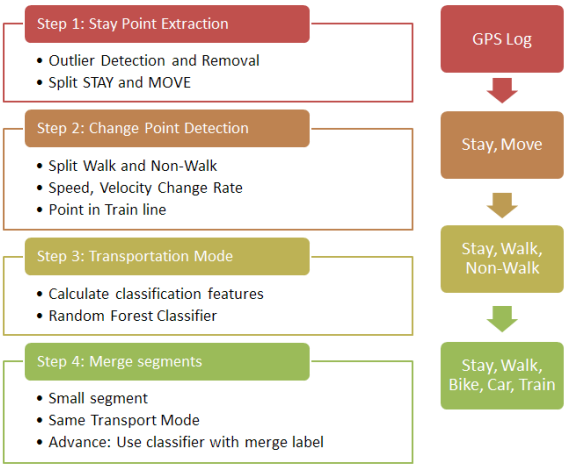

Fig. 1. Overall Framework of Trip Reconstruction

3.1. Step 1: stay point extraction

GPS points represent the spatio-temporal location of people defined by $P = (p_1,p_2,\ldots p_n)$ where p = (id, time, lat, lon) and n = a total number of points. Connecting consecutive points of a user in a day according to their time series, user trajectory can be obtained. In order to split trip segments, we applied stay point extraction algorithm [4][5] based on the spatial and temporal values of points. In the algorithm, a stay point represents a geographic region which a user stays for a while. Space distance and time difference between observed points as shown in following constraints were applied as criteria to detect stay points.

$$\text{Distance}(p_{start}, p_{end}) < D_{threh} \text{ and } \text{TimeDiff}(p_{start}, p_{end}) > T_{threh}$$

Where D_{threh} and T_{threh} are adjustable parameters. D_{threh} is the maximum distance covering a place considered as a stay point. T_{threh} is the minimum time that users spend in the same places. In the experiment, a stay point was detected if T_{threh} > 14 minutes and D_{threh} <= 196 meters. The Haversine formula was used to calculate the great-circle distance between two points instead of Euclidean distance to increase distance accuracy. The stay point extraction was applied to extract stay points of user in each day and kept as a list of stay points defined by $SP = (sp_1,sp_2,\ldots sp_m)$ where sp = (lat, lon, start-time, end-time) and m = a total number of stay points. The latitude and longitude of a stay point is the centroid of all points in the stay points. In addition to a normal stay point extraction method, we used outlier detection and removal technique to automatic remove noise points. Since we focused on using GPS data from mobile phone, GPS location may be shifted due to cell site switching or changing from using GPS to Cell site for identifying the location. Cell site switch usually happen when user move from one location to another location which require mobile phone to contact with nearest Cell site. In some areas such as market places, it has high density of Cell sites. In such case, Cell site switch is likely happened even user did not move along. We defined three types of outlier as following and the example of three types of outlier is shown in Figure 2.

- **First point outlier**: a point that is a last point of the previous move segment but detected as first point of stay point.
- **Inner point outlier**: a point that distances to a previous point and a next point are rather far than other neighbors.
- **Last point outlier**: a point that is a start point of the next move segment but detected as a last point of stay point.

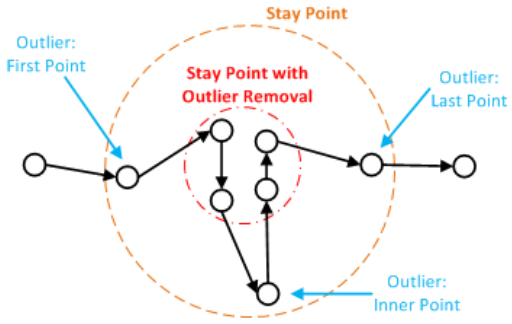

Fig. 2. Stay Point with outlier detection

In order to detect outliers, we assumed that the data is from a normal distribution. Then, we calculate the mean (μ) and standard deviation (σ) of the observed data using the following formula.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Where $x_i = \{x_1, x_2, x_3, \ldots x_N\}$ and N = a total number of observed data. The observed data (x) are the great-circle distances between each two consecutive points. The outliers are defined as a set of points (P) where $(x_i - \mu)/\sigma > 2.6$. If the outlier is a first point or a last point of a stay point, it is removed from stay point and the removed point is merged into move segment. For Inner Point outlier, it is not used for calculating a centroid of a stay point to increase accuracy.

3.2. Step 2: change point detection

We use a change point detection approach based on a method presented by Zheng et al. [2]. However, we adapted some parameters to effectively use with low data rate GPS. Before going further, there are two terms needed to be clarified: segment and change point. A segment is a set of continuous GPS points that belongs to the same transportation mode. A walk segment is a segment where people walk. A non-walk segment can be either a bike, car or train segment. A change point is a GPS point at which user changes transportation mode. Segmentation is done by searching change points based on features in each point and criteria. If a GPS point matches criteria of a previous segment, then it is grouped to the pre-

vious segment. Otherwise, that point is marked as a change point and a new segment is created with a type of walk or non-walk based on criteria. The overall processing step is described in Figure 3.

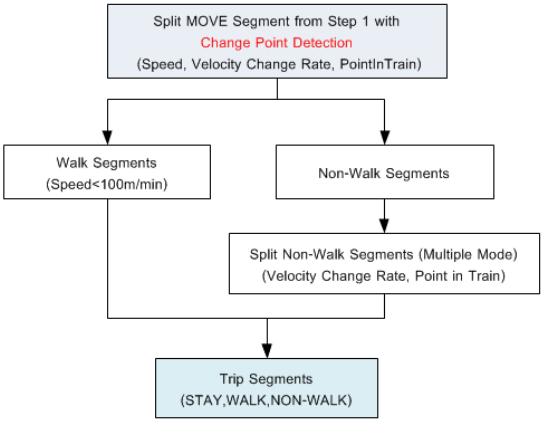

Fig. 3. Change Point Detection for separating segments

In order to create walk and non-walk segment, Zheng et al. [2] used speed and acceleration as features for detecting change points. However, for low data rate such as 5 minute interval, acceleration did not perform well and stable enough because within five minutes, people may change multiple transportation modes. That is also resulted in a possibility of having multiple transportation modes in the first level non-walk segment. To overcome such problem, we introduced two new features which are velocity change rate and point in train line to detect if there are multiple transportation modes in the non-walk segment. If it is detected, a non-walk segment will be split to two non-walk segments. A velocity change rate (VCR) is an average speed of current segment ($S.Speed_{average)}$ comparing with speed of the current observed point (P.Speed). If VCR is over the threshold, the current observed point is marked as a change point.

$$VCR = |S.Speed_{average} - P.Speed| / S.Speed_{average}$$

For point in train line (PiT), we calculate current segment (S.PiT) for whether the points in the segment are located in Train network or not and also do the same with current observed point (P.PiT). Spatial processing is used to search for points in train line polygon. If the current segment did

not in train line but the current observed point is, the current observed point is marked as a change point. For the current segment, S.PiT = a number of points in train line/total points in a segment. We assumed that if S.PiT >0.5, the segment belongs to train line. For current observed point, P.PiT = (p_i .PiT + p_{i+1} .PiT)/2. If P.PiT=1, then it belongs to train line. To avoid spatial resolution problem, we created 50 meter buffer on each train line to increase the possibility of detecting points on the train network.

Fig. 4. Point in train line

3.3. Step 3: transportation mode classification

In this step, we used a supervised learning method for determining transportation mode of users in each segment. As shown in Figure 5, the process is separated into two stages: learning stage and classifying stage. In the learning stage, the non-walk segments from the previous step are labeled transportation mode with ground truth data. Web-based trip visualization and labeling is provided for validating segment and label information. This data is used to create classification features that are used for training classification model. For the classifier, we decided to employ Random Forest as a model because the works reported by [2,8,10] showed that Decision Tree based classifier outperformed other models for inferring mode such as Naïve Bayes (NB), Bayesian Network (BN), Support Vector Machines (SVM). Moreover, Stenneth et al. [3] reported that performance of Random Forest is better than Decision Tree in transportation mode extraction. Basically, Random forest is an ensemble classifier that composes of various decision trees and the result is combined from the outputs of those tree [11,12]. In classifying stage, for each segment, we extract the same features as in the learning stage and then input those features to the

classification model to predict transportation mode of each segment. Hence, differentiate between each mode is done only through the classification process. For ground truth for training the classification process, it has been done with 100 mobile phone users for one month period. Each mobile phone user is required to input their activities, trip purpose, location and transport used for the trip. See more detail in section 4.1.

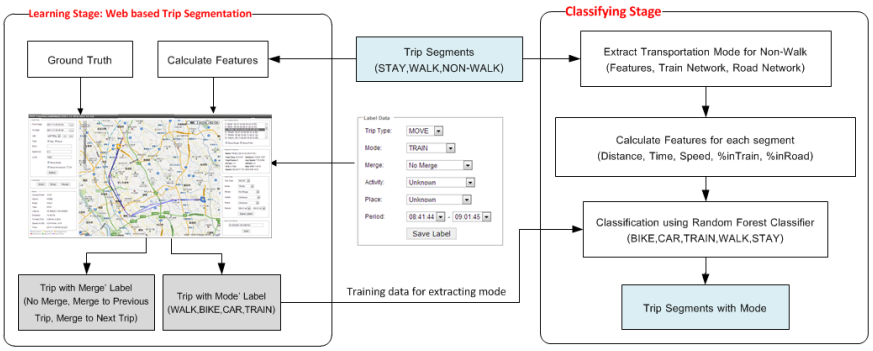

Fig. 5. Transportation mode classifications

3.3.1. Classification features

This section deliberates the classification features used in the proposed method including distance, time, speed and percentage in train line and percentage in road. We calculated all features at segment level.

Total distance (meters)

For each segment, we calculated total distance of a segment by computing a distance of each consecutive point and making a sum of all distances.

Time duration (minutes)

We calculated time duration by calculating time difference between a first point and a last point of each segment.

Speed features

It consisted of six features associated with speed which are minimum speed, maximum speed, average speed, overall average speed, maximum acceleration, velocity change rate. For speed of each point, it is calculated from consecutive GPS points. Average speed is average of speed of each point in a segment. For overall average speed, it is calculated from total distance divided by time duration in the segment.

Percentage of points in train line and road network

Spatial train network and spatial road network are used to calculate features. With an assumption that if people use train or car as transportation, the GPS point should be in the train line or road respectively, we then used a spatial query for finding whether a point is in network polygon or not. We also utilized GIS Buffering function to increase the bounding region of the network because GPS positions are always had some small error and it resulted in located outside the network. Basically, a buffer is an area defined by the bounding region determined by a set of points at a specified maximum distance from all nodes along segments of an object. We used a buffer at 50meters for train network and 100 meters for the road network.

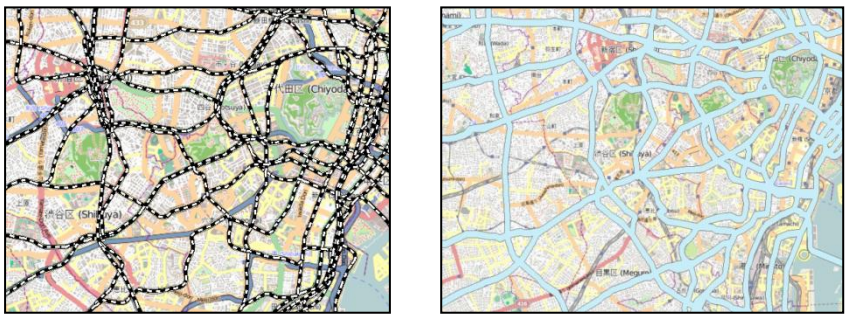

Fig. 6. Train network (left) and Road network polygon (right)

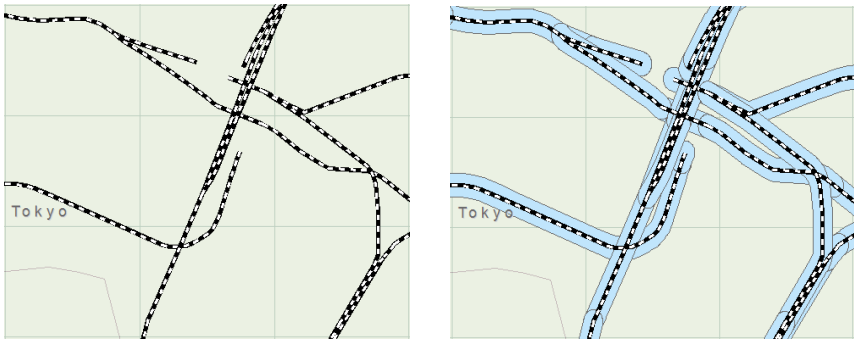

Fig. 7. Example of Buffering techniques on train network

3.4. Step 4: merging segments

To avoid trivial and uncertain segments such as short distance segment or short time segment, we defined a set of policies and thresholds to merge such segments to another segment. We used two thresholds which are min-

imum points in a segment (3 points) and minimum time of a segment (6 minutes). Even though GPS sampling rate is 5 minutes, timestamp of each consecutive point actually fluctuates between 5-6 minutes.

1. If a segment is walk mode and does not exceed the threshold limits, we set it as an uncertain segment and process for merging.
2. If two consecutive segments have same transportation, the segments are grouped into one segment.
3. For uncertain segment, if a mode of the previous segment is STAY and the next segment is not STAY, it will be merged with next segment (STAY-> Uncertain ->BIKE). In opposite, if a mode of the next segment is STAY and the previous segment is not STAY, it will be merged with the previous segment (BIKE-> Uncertain -> STAY).
4. For uncertain segment, if previous and next segments have same transportation mode, all three segments are grouped into one segment (TRAIN-> Uncertain ->TRAIN).
5. For other complex cases (optional), we used a classifier with training data from merging label given from a web-based trip visualization and labeling.

4. Evaluation

In this section, we discuss our data collection method and the experimental results. In order to evaluate and validate the proposed framework, we evaluated the framework by analyzing performance and accuracy of stay point extraction with outlier, transportation mode extraction and trip reconstruction. We show the results of with outlier detection and without for stay point extraction. We also present the transportation accuracy results comparing between common features and common features with spatial features. Finally, we demonstrate the results of trip reconstruction.

4.1. Data collection

The GPS data we used in the experiments is collected by 100 users over a period of one month. Android-based mobile phone is used as a data collection device. The GPS coordinates are collected in every five minutes. With an application installed in the platform, users are able to label trip

purpose and transportation which are used as ground truth data for our classifier. GPS data of each user are processed for segmentations and visualization on our web-based application as shown in Figure 8. We merged ground truth with the created segments and then refined again on our application. In order to give some clue for refining label data, we calculated all classification features and display once segment is selected. We also used Google Map [13] to virtualize a segment as a connecting line for clearly interpretation. For example, the segment located on train lines which can be assumed that user used train for transportation.

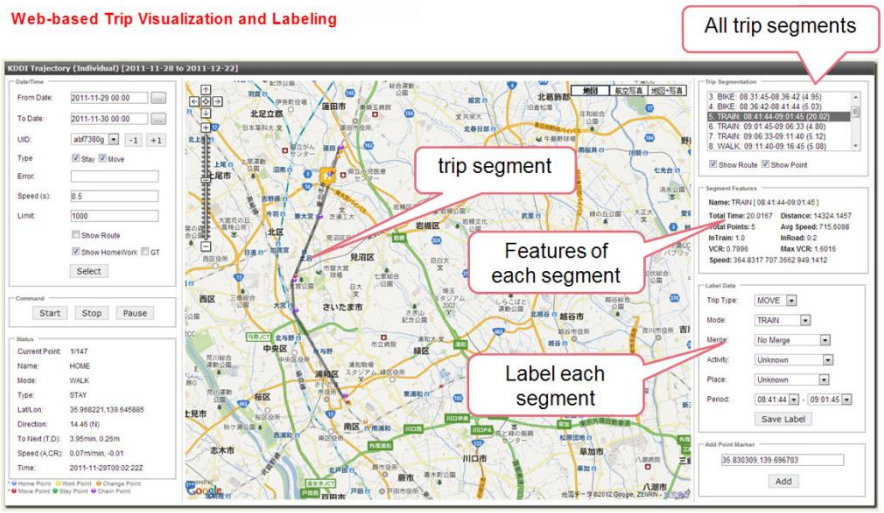

Fig. 8. Web-based trip visualization and labeling tool

4.2. Implementation

For implementation, we used Java language for development. Java Topology Suite (JTS) [14], which was a java-based spatial library, is used for supporting spatial calculation such as find points in geometry and spatial index for fast searching geometry. For data mining techniques, we used Java Machine Learning Library (Java-ML) [15] for clustering, feature selection and classification. PostgreSQL with PostGIS was used as a database system to store GPS data and label data. We used Java Servlet and Google Map [13] for developing of web-based visualization tool.

4.3. Experimental results

For stay point extraction, we first did an experiment on ground truth data to find optimum parameters for stay point extraction which were maximum distance and minimum time duration. As shown in Figure 9, the optimum threshold of maximum distance was 196 meters since it obtained the best result.

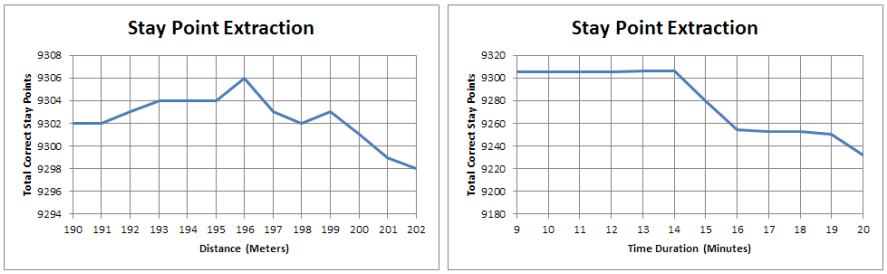

Fig. 9. Maximum distance and minimum time duration for stay point

For the minimum time duration, we decided to use 14 minutes as an optimum threshold because it was an interval that had large changed after a stable period as illustrated in Figure 9. For accuracy evaluation of stay points, precision and recall are used as measurement value. With a distance of 196 meters and time of 14 minutes, it obtained 90.47% for Precision and 85.83% for Recall.

For standard deviation (SD) value which is used as a threshold for detecting outliers, we chose 2.6 as optimum value for SD because a number of incorrect stay points are a bit stable before start to increase again as illustrated in Figure 10.

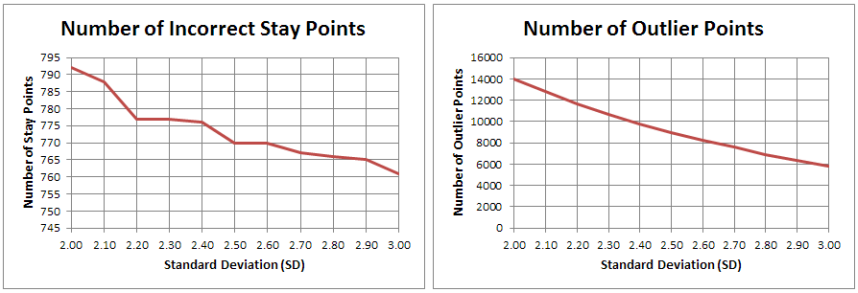

Fig. 10. Comparing Standard Deviation for outlier detection

Moreover, we found that with using outlier detection, it was able to detect all three types of outliers. From our dataset and all stay points, First point outliers were detected about 10.47% of stay points and 15.59% for Last point outliers. We also found that more than 50.89% of stay points contained Inner point outliers. For more clarification, we showed an example of stay point extraction by comparing between using outlier detection and without as seen in Figure 11. The outlier detection showed significant improvement of stay point extraction by making the centroid of stay point more accurate as well as collecting start point of Move segment.

Fig. 11. Comparing results of stay point with and without outlier detection

Figure 12 showed a performance comparison between ordinary stay point extraction technique and stay point extraction with outlier detection and removal. The one with outlier detection had a better result by obtained 92.40% of precision and 90.50% of recall.

Oridinary Stay Point Extaction		With Outlier Detection	
Precision	Recall	Precision	Recall
90.47%	85.83%	92.40%	90.50%

Fig. 12. Performance comparisons of stay point extraction techniques

For transportation mode classification, we evaluated the classification by using two well-known performance measures: Precision and Recall. As described in the previous section, we used Random Forest as a classifier and used a 10-fold cross-validation technique for performance measurement. We first compared the classification results using only common features and using common features with spatial features which are a percentage of points in train line and a percentage of points in road. Then, we tried improving the performance by applying Bootstrap Aggre-

gating (Bagging). As shown in Figure 13, when apply spatial features, the overall performances were significantly improved from 77.41% to 86.89% for Precision and from 61.62% to 84.17% for Recall. Especially for train and car mode, the performances were improved 16.64% and 11.79% respectively. For Bagging, the performances were slightly increased.

Mode	Random Forest (RF)				RF with Bagging	
	Common Features		With Spatial Features			
	Precision	Recall	Precision	Recall	Precision	Recall
TRAIN	83.41%	70.08%	97.30%	88.52%	97.76%	89.34%
CAR	75.36%	83.81%	84.24%	86.54%	83.98%	89.90%
BIKE	59.66%	41.28%	59.15%	48.84%	63.33%	44.19%
WALK	93.59%	97.13%	93.76%	98.46%	93.92%	98.06%
STAY	75.00%	15.79%	100.00%	98.50%	100.00%	98.50%
Overall	77.41%	61.62%	86.89%	84.17%	87.80%	84.00%

Fig. 13. Transportation mode classification results

Figure 14 showed an example of the final result of trip reconstruction. For each segment, it included with necessary such as total time, total distance, transportation mode and trip type.

1,TripPoint [name=STAY, mode=STAY, Time=00:00:15-08:03:50
2,TripPoint [name=MOVE, mode=WALK, Time=08:03:50-08:18:30
3,TripPoint [name=MOVE, mode=TRAIN, Time=08:18:30-08:48:38
4,TripPoint [name=MOVE, mode=WALK, Time=08:48:38-10:56:15
5,TripPoint [name=STAY, mode=STAY, Time=10:56:15-21:16:29
6,TripPoint [name=MOVE, mode=WALK, Time=21:16:29-23:20:25
7,TripPoint [name=MOVE, mode=TRAIN, Time=23:20:25-23:35:57
8,TripPoint [name=MOVE, mode=WALK, Time=23:35:57-23:45:58
9,TripPoint [name=STAY, mode=STAY, Time=23:45:58-23:55:25

Fig. 14. An example of trip reconstruction

5. Conclusion and future work

In this paper, we proposed a detailed design framework to reconstructing user trips including transportation modes focusing on low data rate of GPS data such as at five minutes data rate or one point per five minutes. The framework consisted of four steps: stay point extraction, change point detection, transportation mode classification and segment merging. In stay point extraction, we utilized outlier detection method to remove outliers or error points. We also defined three types of outliers based on its location:

first point outlier, last point outlier and inner point outlier. Moreover, in our dataset, all three types of outliers are detected at 10.47%, 15.59% and 50.89% of total stay points which indicated that outlier detection is a necessary process for stay point extraction to increase the accuracy. For trip segmentation with change point detection, we introduced new features including velocity change rate and point in train for assisting segmentation of non-walk segments contained multiple transportation modes. Since our test data were rather sparse or low density of GPS points than other previous works which resulted in ineffectiveness of many features proposed by those works, we then considered using spatial features in addition to existing features to improve the accuracy of transportation mode inferring. Spatial train network and spatial road network are employed as spatial features and the result showed that overall accuracy increased to 87.80% and 84% for precision and recall respectively. For merging segments, we applied our defined policies to group small and uncertain segments. Additionally, we developed web-based trip visualization for validating the results and also be used for labeling propose. Finally, we reconstructed trips of users and kept in segment structure. In each segment, we also attached it all essential features for further analysis. It provided empirical intermediate data for further analysis particularly on urban analysis and intelligent transportation system and especially when implemented on a large scale.

In the future, with the promising results so far, we have planned to determine significant places of users and merge the result with reconstructed trips. Moreover, we are planning to focus more on using other spatial data for assisting and providing some novel label data for the trips.

Acknowledgements

The work described in this paper was conducted with an agreement from KDDI to use mobile phone datasets of personal navigation service users. This work was supported by GRENE (Environmental Information) project of MEXT (Ministry of Education, Culture, Sports, Science and Technology).

References

Chung E, Shalaby A (2005) A Trip Reconstruction Tool for GPS-based Personal Travel Surveys, Transportation Planning and Technology, 28:5, 381-401. doi: 10.1080/03081060500322599

Zheng Y, Chen Y, Li Q, et al. (2010) Understanding transportation modes based on GPS data for web applications. ACM Transactions on the Web 4:1–36. doi: 10.1145/1658373.1658374

Stenneth L, Wolfson O, Yu P, Xu B (2011) Transportation mode detection using mobile phones and GIS information. ACM SIGSPATIAL, ACM Press: 54-63, doi: 10.1145/2093973.2093982

Ashbrook D, Starner T (2003) Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7:275–286. doi: 10.1007/s00779-003-0240-0

Zhou C, Frankowski D, Ludford P, et al. (2007) Discovering personally meaningful places. ACM Transactions on Information Systems 25:12–es. doi: 10.1145/1247715.1247718

Liao L, Patterson DJ, Fox D, Kautz H (2006) Building personal maps from GPS data. Annals of the New York Academy of Sciences 1093:249–65. doi: 10.1196/annals.1382.017

Andrienko G, Andrienko N, Wrobel S (2007) Visual analytics tools for analysis of movement data. ACM SIGKDD 9:38–46. Doi: 10.1145/1345448.1345455

Zheng Y, Li Q, Chen Y, et al. (2008) Understanding mobility based on GPS data. In Ubiquitous Computing, ACM New York, pp. 312-32.

Zheng Y, Liu L, Wang L, Xie X (2008) Learning transportation mode from raw gps data for geographic applications on the web. World Wide Web, pp. 247–256.

Reddy S, Burke J, Estrin D, et al. (2008) Determining transportation mode on mobile phones. 12th IEEE International Symposium on Wearable Computers 25–28. doi: 10.1109/ISWC.2008.4911579

I. Witten, E. Frank (2005) Data Mining: Practical machine learning tools and techniques. Morgan and Kaufmann, San Francisco

R. Duda, P. Hart, D. Stork, (2000) Pattern Classification. Wiley, New York

Google Map http://maps.google.com/

Java Topology Suite http://tsusiatsoftware.net/jts/main.html

Java Machine Learning Library http://java-ml.sourceforge.net