On Extracting Commuter Information from GPS Motion Data

Dietmar Bauer, Markus Ray, Norbert Brändle, Helmut Schrom-Feiertag
arsenal research
Giefinggasse 2
A-1210 Vienna, Austria
dietmar.bauer@arsenal.ac.at

ABSTRACT

Commuters rely on realistic and real-time information in order to optimize the time spent on commuting between home and work. Delays in (urban) transport and congestion for individual motorized transport are a major issue for unnecessary long travel times. While some of these delays occur randomly, there is also a systematic component. In this paper we describe a data-driven approach to analyze positions of an individual collected using GPS to obtain information on the individual's typical routes, typical schedules and the used mode of transport. Furthermore, we propose an approach to model the probability of an event like missing a train as a function of time. This allows to optimize the expected commuting time based solely on the commuters motion history. Suitability of the approach is demonstrated in a real world application based on a dataset comprising six weeks of GPS tracks.

1. INTRODUCTION

Commuters typically follow a routine in their daily schedule for reaching office from home in the morning and returning home in the evening. Commuting often includes using a combination of regional trains running at fixed schedules, underground or other public transport running frequently but irregularly, individual transport by foot, bicycle or car.

Commuters then have a number of choices such as different routes, different scheduling of the time of leaving home and the office, respectively.

Over the years commuters acquire information on the time it typically takes to reach home from work, the best connections both in terms of reliability and time efficiency using trial and error processes and word-of-mouth from colleagues and friends in the same situation. They also learn to time their journey such as to avoid as much as possible to be caught in congested traffic.

In case of changes in time tables of public transport and changes in the road network, the previously acquired information is obsolete and must be updated with new information contained in the motion history e.g. provided by frequent localization.

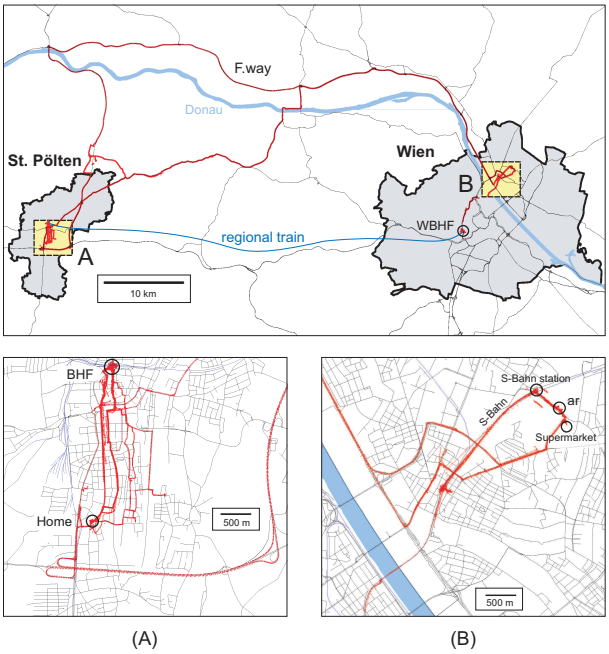


Figure 1: Overview over the collected data set and close up view of two areas: (A) St. Pölten, (B) vicinity of arsenal research.

Addionally, delays caused by unusual disruptions in the transport system (such as accidents) could be prevented by a timely reaction of a commuter. Timely information could be supported by predictions on the route taken based extracting a commuter's habits from the motion history.

This paper proposes a methodology to extract information of a commuter's own motion history measured with GPS tracking. This information can support travel time optimization by 1) providing accurate situation aware predictions of expected travel time and 2) building the basis for personalized pre- and on-trip travel information.

In contrast to other studies using GPS tracking as a means of data collection (like e.g. [1, 2, 3]), our proposed approach uses GPS observations as the single source of information. No time tables for public transport or geometric information is included, thus achieving a fully self-contained data analysis. Moreover, the GPS localizations are assumed to be

MobiQuitous 2008, July 21-25, 2008, Dublin, Ireland

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881

obtained from a mobile device such as a smart phone or a PDA. In contrast to car navigation systems, such increasingly popular mobile devices can also be used in multi-modal situations.
This paper is organized as follows: Section 2 provides an overview of the methodology. Section 3 discusses a real world example covering six weeks of data of a commuter as a demonstration example. Figure 1 provides the measured GPS coordinates of this data sets as red dots. Section 4 draws some conclusions and discusses potential applications.

2. METHODOLOGY

The proposed analysis method is based on the motion history of a commuter represented as a sequence of time step position pairs $(t_i, x_i, y_i), i = 1, \ldots, n$. While this paper focuses on GPS-tracking as a means of data collection, the data measurement technology is of minor importance subject to certain restrictions described below.
The approach is based on the concept of points-of-interest (POIs), defined as places where an individual stops for a certain time span. The most important POIs are the living place and the working place, but also transfer points of transport mode changes. Two POIs are connected by routes. The proposed approach is composed of the following steps:

1. **Detecting POIs and splitting trajectories** by detecting stops and signal losses.
2. **Preprocessing trajectories** by removing outliers due to GPS shadowing effects and closing measurement gaps to POIs.
3. **Obtaining route information** by finding main routes between detected POIs.
4. **Detecting travel mode** based on characteristics of the trajectories.
5. **Obtaining timing information** by analyzing times of entering and leaving POIs.
6. **Estimating probability of catching connection.**

These steps are described in more detail in the following.

2.1 Detecting POIs and Splitting Trajectories

POIs are detected in the raw observation data set provided by conventional GPS devices with the stop detection algorithm published in [4]. A stop is detected in the time-ordered data set henever a segment extending for a time span longer than T minutes is found such that all positions lie within a circle of radius R. The found trajectory segment is then extended until for the last time the smallest circle containing all locations in the segment has radius smaller than R. Then the stop is said to occur at the mean location in the segment. Hence in particular at positions where signal losses over a prolonged time interval are observed (which happens e.g. if a building is entered) POIs are detected.
This results in a typically large set of POIs which is reduced by clustering nearby POIs using hierarchical clustering algorithms. Subsequently, POIs in one cluster are replaced by the corresponding cluster center. The number of POIs obtained in this way is determined by the cutting height in the dendrogram corresponding to the maximal distance between the members of the obtained cluster.
The output of this stage is a set of POIs. Additionally the observations are segmented into trajectories that start or end near POIs.

2.2 Preprocessing Trajectories

The GPS system can provide accurate positioning data only as long as an unobstructed view to at least four satellites can be guaranteed. Shadowing effects caused by obstacles like buildings or trees often lead to inaccurate position measurements during travel. Furthermore, an initialization phase after a longer signal loss (e.g. when switching off the device or entering a building) may last up to five minutes with no location data (depending on the device).
Hence the collected data needs to be preprocessed in order to limit the influence of such problems. Initially, an approach based on individual velocities removes outliers from further processing: Positions which could only be reached with untypically high velocities are removed from the data set. This method only works reliable for high frequent measurements.
In oder to fill the measurement gaps caused by GPS device initialization, linear interpolation to the closest POI is used. Consequently, every trajectory will end at a POI. In many cases the subsequent trajectory will start close to this POI. This information can be used in order to predict the path taken from the POI to the first observation of the next trajectory. Hence, a simple idea in this respect is linear interpolation to the closest POI using the mean speed of the following trajectory.
If the distance to the closest POI reaches a predefined maximum distance, a new POI at the starting point of the trajectory is added. This ensures that all trajectories start and end at POIs.
The resulting set of trajectories is smoothed (to reduce the noise level) and resampled (in order to obtain equal spacing in time).

2.3 Obtaining Route Information

In order to detect the main routes taken by the commuter between different POIs, an automatic path learning algorithm [5] is applied to the trajectory set. This approach uses a vector quantization in order to obtain a set of prototypes. The propotype set is subsequently reduced in order to obtain a minimum distance between the prototypes of 30 meters. The length of 30 meters is chosen to roughly match the accuracy of the position data. The trajectories are then viewed as a sequence of prototypes by identifying data points with the corresponding nearest prototype. Then a clustering algorithm specific to this setting is applied to the sequences of prototypes representing the trajectories in order to obtain sets of prototypes representing a route. Details of the implementation can be found in [6].
After detecting the main routes, the usual behavior can be separated from the unusual one and only routes retained that are 'often' used (the exact meaning of 'often' is a user parameter). Subsequently, only POIs are considere that lie on the main routes.
The output of this stage is a set of routes connecting the various POIs and a collection of trajectories linking the POIs.

2.4 Detecting travel mode

In this step the trajectories linking POIs are labeled according to different modes of transport. The mode detection differentiates between motorized traffic, public transport, bi-

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881

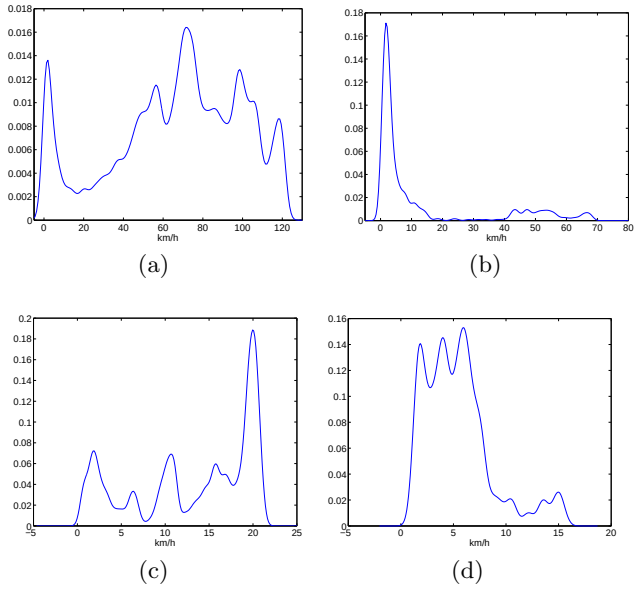


Figure 2: Density plots of velocities for different trajectories: (a) car, (b) regional train, (c) bike and walking, (d) walking. Negative values are artifacts of the chosen kernel density estimator.

ke usage and walking. Each of the transport modes has its own characteristic and its own reasons for time delays. Motorized traffic is subject to congestion and accident related delays, public transport has a fixed time schedule, and hence commuting delays are caused by missing a connection.

Much information on the used mode is contained in the distribution of the velocities in the trajectories as can be seen in Figure 2: It shows the estimates of the probability density function of the distribution of velocity for four manually classified trajectories of the demonstration data set: From these plots it can be seen that the car traffic attains the largest speed with some probability mass at approximately 120 km/h (see Fig. 2 (a)), close to the speed limit on Austrian freeways that lies at 130 km/h. For public transport trajectories, the speed profile (Fig. 2 (b)) contains mass between 40 and 70 km/h and an additional mass around zero due to train stops. The profile for biking shows speed maxima of about 20 km/h. (see Fig. 2 (c)). Walking comprises speeds up to 15 km/h corresponding to running (see (Fig. 2 (d)).

This demonstrates that mode detection can be based on velocity distributions. Note that this requires a correct splitting of the data set into trajectories related to single travel modes in order to avoid missclassification due to mixed velocity distributions. Such a mixing can occur when the mode change does not involve stopping for a longer period. Such changes in travel mode can also be detected based on speed distributions. The above mentioned method certainly is too simplistic and more elaborate mode detection algorithms are left for future research.

The output of this stage is a labeling of the trajectories according to the travel mode chosen.

2.5 Obtaining timing information

The previous processing steps provide for each pair of POIs a number of trajectories leading from one POI to the other are obtained together with a classification into different travel modes.

Hence for each pair of POIs and for each travel mode, the weekdays when this combination is used can be sought and patterns (such as usage of a particular travel mode only on certain weekdays) can be found. This information is useful for situational awareness.

Moreover, for each route and each travel mode, average travel times conditional on the starting time can be obtained from using kernel estimators. This clearly indicates systematic delays due to congestion and provides the basis for optimization.

time tables exist for public transport and regional trains typically run at regular intervals such as e.g. every 15 minutes. Hence for optimization it is important to understand the time schedule. This information is contained in the motion history of commuters for sufficiently long observation period, as the similar entry and exit times at the corresponding POIs are found. Using Gaussian mixture models, the most likely leaving times for the most often used connections can be estimated. Consequently, the personal time schedule can be obtained. This information is useful for optimization as well as for the detection of the current situation of the commuter.

2.6 Estimating probability of catching connection

In some case the optimization of the travel time is not the ultimate goal. Sometimes it is crucial not to miss a connection. In this situation a model for the probability to miss a connection as a function of starting the trip is useful.

Such models can be obtained using the logistic function (see e.g. [7]): Consider the situation where a particular POI is identified to correspond to the junction of public transport, and where for a longer time period the times leaving the previous POI as well as the times of arriving at the junction and leaving it are observed. Based on this information it can be decided whether the commuter missed her connection resulting in a unusually long waiting time. Let $y_j = 1$ if the commuter reached the train according to the j-th data point and $y_j = 0$ else. Let x_j denote the time of leaving the previous POI. The logistic function then parameterizes the probability of catching the train according to

$$\mathbb{P}(y_j = 1|x_j) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_j)}$$

Here $\beta_0, \beta_1 \in \mathbb{R}$ are parameters to be estimated. Typically, the maximum likelihood paradigm is used under the assumption of conditionally binary distribution of the dependent variable for the estimation supplying also information on the accuracy of the obtained estimates.

3. DEMONSTRATION EXAMPLE

Here we demonstrate the proposed approach using data collected with a GlobalPoint Emtac Trine II GPS receiver equipped with the SiRF Star IIe LP technology (see www.sirf.com for details). The receiver also contains a data logger such that the collected data can be accessed ex post via a bluetooth connection. The data has been collected on twenty seven days within seven weeks and covers the route to and from work for one individual. The individual did not change his commuting habits during that time span and no

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881

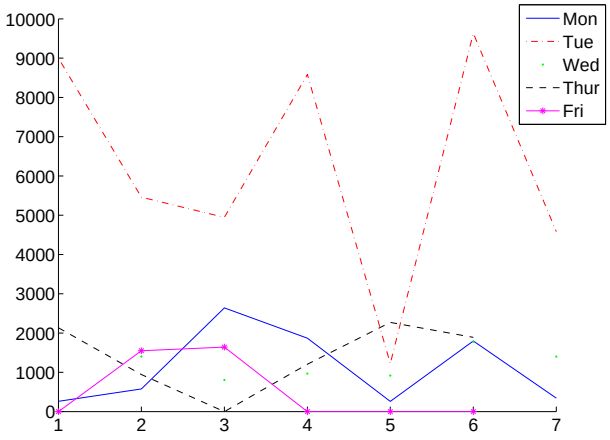


Figure 3: Number of data points collected per trip.

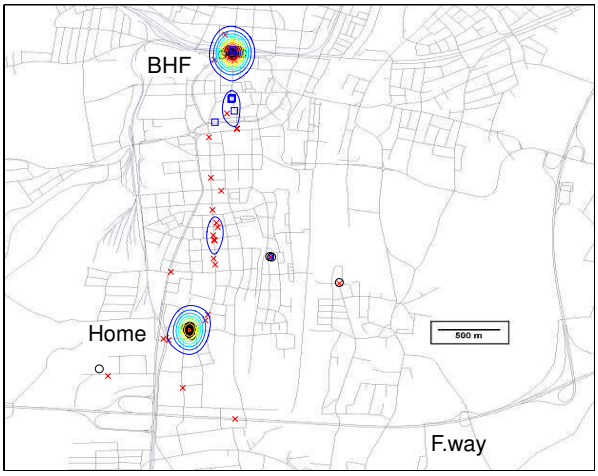


Figure 4: Detected stops in the map region Fig.1 (A).

time table changes for the used public transport occurred. Fig. 1 shows the collected data points.

A total of 70076 data points have been collected at a fixed sampling rate of 2 seconds in the region of Lower Austria and Vienna. This high sampling rate has been chosen to make a trade-off between gathering as much information as possible and saving memory for longer observations. While the GPS receiver also provides velocity information in it's own right, we judge it as being unreliable and instead prefer processing of the raw position data only. The tracked individual lives in St. Pölten, Lower Austria and commutes to work either using a car via a freeway ('F.way' in Fig. 1) or public transport (via the Westbahnhof train station, 'WBHF'). The workplace at arsenal research is labeled as 'ar' ((B of Fig. 1)). The car is used on Tuesdays since the individual drives to a sports ground in Herzogenburg in the evening. At all other days public transport is chosen. In this case the sequence of travel modes is bike (home to BHF), regional train (BHF to WBHF), subway (WBHF to S-Bahn station) and walking (S-Bahn station to ar).

It can be noticed that no GPS tracks are available between WBHF and the train station in St. Pölten, denoted as 'BHF' in part (A) of Fig. 1.

In the vicinity of the workplace (see Fig. 1 (B)), the most used station of the regional train is visible ('S-Bahn station'), the workplace itself ('ar') and also a supermarket, which was used a couple of times ('Supermarket'). The plot of St. Pölten (see Fig. 1 (A)) shows a number of different paths connecting home ('Home') to the railway station ('BHF') as well as the entry and exit of the freeway frequented with the car.

Data collection is not fully reliable. Every two seconds a position should be available. Fig. 3 provides some quantitative information. It can be seen that the amount of data obtained at the various weekdays differs vastly. One reason for these differences can be seen in the mode of transport: in the car, the GPS receiver records position data frequently, whereas during the train trip no data is collected at all. Data collection in the subway and the regional train is partly possible. However, also sometimes data collection failed. This leads to a situation where not for all evaluations to be discussed below the same data basis can be used.

The first step in the analysis of the data sets is to obtain the POIs via detection of stops. Here $T = 5$ minutes and $R = 25$ meters were used as parameters for the stop detection algorithm. Fig. 4 provides a plot showing the results in the town of St. Pölten. In the plot the starts of the trajectories are plotted as red crosses, the ends as black circles and the stops as blue squares. Furthermore the spatial density of all these points is represented using a contour plot. The density is estimated using a standard spherical Gaussian kernel with manually adjusted bandwidth. The plot comprises 223 marked locations originating in 90 starts, 90 ends of trajectories and 43 detected stops (not all within the plotted region).

After preprocessing according to Sect. 2.2, the main routes were computed leading to 26 paths, where only eight are used for more than two times. The found eight paths correspond to the typical route taken with the car (one path for each direction), the two bike routes (differing for the direction to the station BHF and home respectively) and four paths representing the subway and regional train routes connecting arsenal research and the Westbahnhof. Fig. 5 shows the segment of four of these paths appearing in the region (A) (see Fig. 1). The figure also shows that the quality of the found paths - in terms of position accuracy - is very high, approximately in the range of the map accuracy.

After omitting POIs not on the main routes, the number of POIs is reduced to eight points: In St. Pölten only two POIs remain: One is located at the living place of the tracked individual, the other one at the railway station (BHF). The other stopping points found (not shown in the figure) are located at the workplace, the supermarket close to the workplace, the station of the local train close to the workplace and at the Westbahnhof where the author changes trains. The last two clusters of detected stopping points are located at Herzogenburg where the tracked individual uses to play basketball at Tuesday nights. Thus all expected points are correctly found with this simple algorithm. The position accuracy of the found POIs mainly depends on the quality of the GPS and hence lies in the range of a few ten meters.

The next stage in the proposed approach is mode detection performed for each trajectory as described in Sect.2.4. In this particular data set, mode detection effectively separates the different modes of transport: In all but 4 (out of 90)

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881

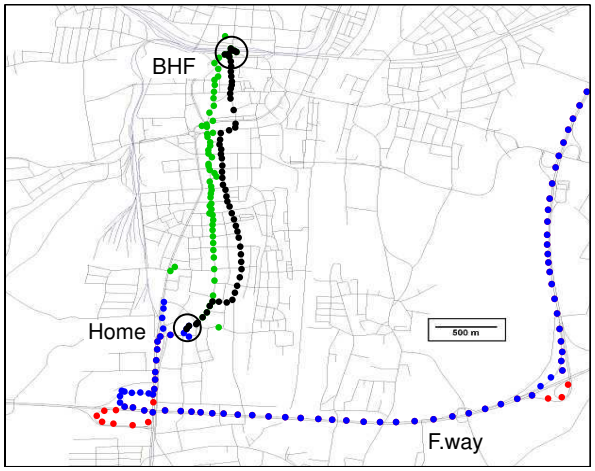


Figure 5: Detected main paths in the St. Pölten area (see Fig. 1).

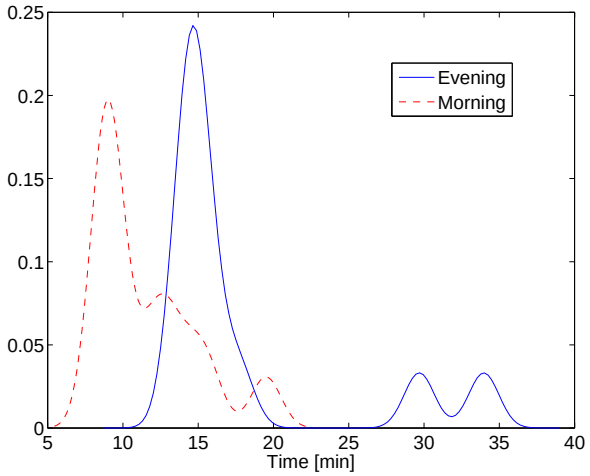


Figure 6: Measured bike travel times.

trajectories the correct mode is detected. Car is misclassified twice as public transport, since in these trajectories no freeway was used. Public transport has been classified once as car traffic. Finally also walking has been misclassified as bike riding due to noise in the initialization phase. Note however, that this situation seems to be special for the characteristics of this particular commuting paths. The data set distinguishing car use from public transport relies on the assumption that a freeway is used with the car. Also mode changes within the trajectories are not detected. In more complex situations also more elaborate mode detection schemes need to be devised.

By inspecting the days of car usage it could be observed that the car is used only on tuesdays. The reason for this is that the commuter has one stop on the way home on Tuesdays at the sports ground in Herzogenburg. The path followed on the way home is significantly different from the path to work.

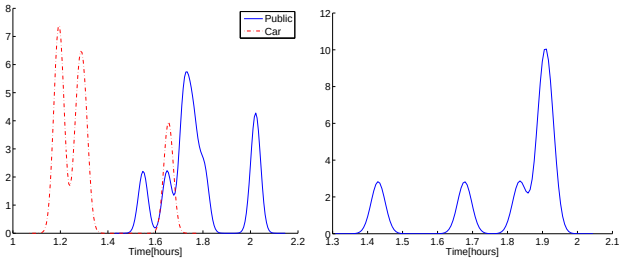
With respect to the scheduling of the travel times it can be observed that the first position in the morning when using the car vary substantially (ranging from 6:58 to 7:31). Next consider the times of leaving home in the morning in case public transport is chosen: In this case the data lists as first observations 6:50, 6:53, 6:54 (three times), 6:55, 6:56, 6:58, 7:03, 7:05 (twice), 7:07 and 7:10. Again it is observed that initialization of the GPS device leads to a time lag in recording the first position. This is corrected for by using a rough estimate of the time leaving home obtained by assuming the mean speed to be taken from home to the position of the first collected data point. This leads to corrected times of 6:43, 6:46, 6:47, 6:49 (three times), 6:51, 6:54, 6:56, 7:01, 7:04 (twice) and 7:06 indicating the usage of two different trains one leaving at approximately 6:50 while the other one leaves at approximately 7:05. The times recorded for the start of sixteen bike trips from BHF to home in the evening are 19:02 (twice), 19:04 (twice), 19:05, 19:06, 19:10, 19:20 (twice), 19:21 (four times), 19:22, 19:28 and 20:05. This indicates the usage of three different trains arriving around 19:04, 19:21 and at 20:05

Fig. 6 shows a comparison of the duration of the bike ride. The plot shows the estimated density of the distribution of the bike travel times for the morning (red dashed curve; based on 13 observations) and the evening (blue curve; 12 observations). It can be seen that the ride to the train station takes on average 11.5 minutes with 3.4 minutes standard deviation, whereas the ride back takes on average 18 minutes with more than 6 minutes standard deviation. Also visible are two large observations which can be traced to stops on route. Note, however, that the sample size is quite small as a basis for nonparametric density estimates and hence single observations influence the shape of the estimate much.

Finally the total travel time was calculated separately for the mornings and the evenings. Only those trips which directly connect home and work place were considered and trips having stops in between (such as the car trips with stops for sports activities) are discarded for this part of the analysis. Fig. 7(a) provides a plot of the estimated density of the travel time (separated into car and public transport) in the morning. It shows that mean travel time is 1.6 hours with a peak at 1.25h representing the travel time using car (red dash dotted line) and a second peak at 1.75h representing typical public transport travel times (blue solid line). One can notice car travel time at 1.6h. The explanation for this delay lies in bad road conditions as a consequence of heavy snowing leading to a significant slowdown. Also the two long travel times with public transport (over 2 hours) can be attributed to delays due to bad weather. Fig. 7 (b) shows the direction from work to home for public transport. For car travel only one observation (1.2h) is contained in the data set due to the frequent stops at Herzogenburg on the way home. As a difference to the morning travel the public transport travel time in the evening shows two fast trips while the bulk is located around 1.9 hours reflecting the longer bike trips. The two fast trips occur in combination with the train at 18:20 while all longer trips use the 18:34 train.

As a consequence two facts stick out: First the commuter has a strong preference for public transport since he accepts an additional approximate half hour. Secondly, the duration of the trip back home depends crucially on the fact whether the 18:20 train is reached or not. The penalty for missing this train is an additional 15 minutes of travel time. Thus the timing of the trip in order to grant a certain probability

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881



(a) (b)

Figure 7: Travel times during the morning (a) and in the evening (b).

to reach the train is of importance. The small sample size in this respect does not allow for an accurate estimation of the probability catching the train which is hence left for future research.

4. CONCLUSIONS

We have proposed a methodology to extract information from the motion history of a commuter. Information on the chosen mode of transport, places of changing transport mode, route choice as well as timing information is obtained. This information is retrieved based solely on the collected data without the need of interaction with the commuter. The main results in this respect are the main routes taken, the habits with respect to the usage of car and of public transport, usual times for leaving home and leaving work.
The methods are demonstrated to be suitable using a small data set obtained using GPS localization over a period of seven weeks. The discussion showed that the subject of study prefers public transport (for whatever reason) since he accepts longer travel times during most of the work week while on tuesdays the car is chosen relating to an obvious activity on the return trip. For the train trip it has been found that in the evening mainly two trains connecting Westbahnhof and St. Pölten Hauptbahnhof are chosen where the second train only results if the first one is missed. This provides a detailed description of the habits of the commuter extracted solely from his own motion history using the proposed methodology.
During the investigation a weak point of the methodology has been found to be the mode detection scheme. The current implementation is definitely too simple in order to work in more complex settings. Two extensions immediately come to mind: Either geocoded information is consulted including the location of public transport facilities or a more elaborate statistical approach is to be chosen, such as the one followed in [8]. This is left for future research.
Finally also the mode of data collection might be questioned. Future research will examine the possibility to use the more widely available cell-ID localization techniques in combination with mobile phone usage.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the help of Rainer Rehberger in the data collection phase.

6. REFERENCES

[1] S. Schönfelder, K. Axhausen, N. Antille, and M. Bierlaire, "Exploring the potentials of automatically collected GPS data for travel behaviour analysis - a swedish data source," in *GI-Technologien Für Verkehr und Logistik, IfGIprints 13*, J. Möltgen and A. Wytzisk, Eds. Institut für Geoinformatik, Universität Münster, Münster, 2002, ch. 13, pp. 155–179.

[2] K. W. Axhausen, S. Schönfelder, J. Wolf, M. Oliveira, and U. Samaga, "80 weeks of GPS-traces: Approaches to enriching the trip information," *Transportation Research Record*, vol. 1870, pp. 46–54, 2004.

[3] J. Wolf, R. Guensler, and W. Bachmann, "Elimination of the travel diary: An experiment to derive trip purpose from GPS travel data," *Transportation Research Record*, vol. 1768, pp. 125–134, 2001.

[4] R. Hariharan and K. Toyama, "Project lachesis: Parsing and modeling location histories." in *Proceedings of the Third International Conference on GIScience*, Adelphi, MD, USA, 2004.

[5] R. Pflugfelder, "Visual Traffic Surveillance Using Real-Time Tracking," TU Wien, Austria, TU Wien, Tech. Rep. PRIP-TR-071, 2002.

[6] D. Bauer, N. Brändle, R. Pflugfelder, and S. Seer, "Finding highly frequented paths in video sequences," 2006, submitted to ICPR'06.

[7] M. Ben-Akiva and S. Lerman, *Discrete Choice Analysis*. MIT Press, 1985.

[8] D. Patterson, L. Liao, D. Fox, and H. Kautz, "Inferring high-level behavior from low-level sensors," in *UBICOMP 2003: The Fifth International Conference on Ubiquitous Computing (October 12-15, 2003)*, 2003.

Digital Object Identifier: 10.4108/ICST.MOBIQUITOUS2008.3881
http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3881