IDENTIFICATION OF FEATURES FOR TRAJECTORY SEGMENTATION ACCORDING TO THE TRANSPORT MODE

Martina Erdelić[1*], Tonči Carić[1], Tomislav Erdelić[1] and Nikola Mardešić[1]

[1] Faculty of Transport and Traffic Sciences, University of Zagreb, Croatia

*correspondence E-mail: merdelic@fpz.unizg.hr

Keywords: urban mobility; transport mode; smartphone sensor data; trajectory segmentation; feature selection

1. Introduction

A transport network is a complex system whose analysis requires data from transport network users. The influence of the growing degree of urbanization and population places transport as one of the main factors affecting the quality of life in large cities. Therefore, research fields such as urban mobility, energy consumption, pollution reduction and security are often the subject of scientific research. All these research fields are directly or indirectly related to the mobility of transport network users. Observing the model of human mobility through data collected in different transport environments can significantly contribute to solving transport problems.

An in-depth trajectory analysis, which can be supported by clustering or classification methods, can describe the mobility of users through the transport network. Recognition of everyday human activities during various locomotion and uses of transport modes has several applications, such as educating users to change their behavior pattern or promoting a healthier lifestyle. In addition, applications, which track user mobility, can help with public urban transport planning, smart parking, and vehicle traffic monitoring (Kalašová et al., 2021).

User trajectories often contain several connected transport modes that a user uses, moving from the origin to the destination. Hence, such trajectories need to be segmented before further processing. Therefore, there is a need to develop a trajectory segmentation method that will recognize the Mode Transfer Point (MTP), i.e., the point when one transport mode ends and starts the use of the following transport mode. Furthermore, trajectory segmentation is important for transportation network analysis because a mobility pattern is not the same for the entire trajectory but could be for some of its segments.

The accelerometer sensor is the most widely used among the sensor modalities for transport mode classification and trajectory segmentation (Hemminki et al., 2013). Global Positioning System (GPS) data-based approaches use features derived from speed, GPS coordinates, and traveled distance (Gong et al., 2018). Geographic Information System (GIS) data, such as bus or train stops, are also often included in datasets to avoid the shortage of relevant features. The main advantage of accelerometer sensor data is low power consumption, which enables continuous monitoring of

human behavior with higher sample frequency and does not depend on external signal sources. However, to avoid the shortage of GPS and GIS data, it is necessary to identify the relevant features that will better describe the user trajectory.

In this paper, several features, and their influence on the accuracy of a MTP detection were tested. The time and frequency domain of trajectory data are used for feature extraction. Features are selected using ANOVA F value and different feature sets are tested using Transition State Matrices (TSM) approach described in the research (Erdelić et al., 2022).

2. Materials and methods

In this paper, the benchmark dataset Sussex-Huawei Locomotion-Transportation (SHL) was used (Wang et al., 2019). SHL is a large-scale dataset of smartphone sensor data recorded through seven months in 2017, usually used for multimodal transportation analytics (Gjoreski et al., 2018). The dataset contains 753 h of multimodal data: car (88 h), bus (107 h), train (115 h), subway (89 h), walk (127 h), run (21 h), bike (79 h), and still (127 h). For this research, the original classes were transformed into two new classes: change of transport mode or no change of transport mode. While collecting data, each user had a device in multiple places on the body to track the movement. The mobile device orientation is not necessarily fixed. The dataset contains seven sensors, but only data collected from a 3D accelerometer, gyroscope, magnetometer, and gravity were used in this research. Data are collected from all sensors with a frequency of 100 Hz. The dataset is divided into training and testing datasets containing 70% and 30% of the data, respectively.

Data quality affects the data mining results; therefore, data need to be preprocessed before feature extraction. The paper presents three primary data preprocessing techniques: data cleaning, time window segmentation and data transformation using Discrete Fourier Transform (DFT). Before implementing a methodology for MTP detection, the essential part is extracting valuable and distinct information from time windows. Therefore, data are transformed from the time to the frequency domain for each sensor. Furthermore, a challenging task in the raw signal data analysis is to solve the orientation issue. In the SHL dataset, smartphones can be carried in different orientations, meaning that the same sensor values in various circumstances can represent moving in different directions. In this paper, we computed the magnitude value for all sensors by Equation 1 where $\boldsymbol{M}$ is the magnitude, and $\boldsymbol{S_x}$, $\boldsymbol{S_y}$, and $\boldsymbol{S_z}$ are sensor values along the x, y, and z axes, respectively.

$$\boldsymbol{M = S_x + S_y + S_z} \quad (1)$$

In summary, for every time window, features are computed as a compound of the three axes (x, y, and z) of sensor data with a magnitude vector. For each sensor, all features presented in Table 1 are computed. The calculated features refer to basic statistical features such as mean, median, standard deviation, variance, mean absolute value, interquartile range, maximum and minimum values, and root mean square value. Measures such as the kurtosis, skewness and peak value describe data distribution within a time window. The z value shows the distance of the data in the current time window from the mean, where the distance is measured in standard deviations. The kurtosis, skewness, and shape factor provide insight into the behavior of the data in a single time

window. In contrast, the impulse, crest, and clearness factors focus on peak values within a time window (**Table 1**).

Table 1. Computed features in the time and frequency domain

Feature	Symbol	Time domain	Frequency domain
Mean	$\overline{x}$	✓	✓
Variance	σ	✓	✓
Standard deviation	s	✓	✓
Median	X	✓	✓
Skewness	g	✓	
Kurtosis	k	✓	
Z score	z_i	✓	
Interquartile range	IQR	✓	
Peak to peak	PTP	✓	
Maximum	x_{max}	✓	
Minimum	x_{min}	✓	
Energy	E	✓	
Average absolute value	$\overline{x}_{abs}$	✓	
Root mean square	RMS	✓	
Variance of frequency data	RVS		✓
Frequency center	F_c		✓
Impulse Factor	F_i	✓	
Crest Factor	F_p	✓	
Clearance Factor	F_r	✓	
Shape Factor	F_s	✓	✓
Kurtosis Factor	F_k	✓	
Skewness Factor	F_g	✓	

All features are calculated for accelerometer, gyroscope, gravity, and magnetometer data resulting in 432 features (27 features * (3 axes + 1 magnitude) * 4 sensors= 432). To identify relevant features, feature importance analysis is performed. Another reason for feature analysis is to determine which features are more important for MTP identification, primarily due to the imbalance between the number of time windows in which MTP has occurred and those representing time windows in which there is no MTP.

The process of identifying relevant features is divided into two parts. First, the features are sorted according to the ANOVA F value, and secondly, features with a significant impact on the detection of false positives are removed (classification of time windows in which no change occurred as those in which a change occurred).

The ANOVA test is a statistical method used to determine significant differences between two or more classes by testing differences in mean values using the variance of included features (Sawilowsky, 2002). In other words, the value obtained by the ANOVA test shows how well a feature distinguishes two classes. Two measures are used to calculate the ANOVA test: variability within and between classes. Both measures use the sum of squares to determine the dispersion in the data. Variability within classes SS_u shows the dispersion of data in the entire dataset, where data belonging to different classes are considered separately, given by Equation 2. Variability

between classes SS_i shows the variability of a subset of data belonging to the same class in relation to the entire dataset, given by Equation 3. In both equations, Equation 2 and Equation 3, k represents the number of classes, x_{ij} is the i value of the observed feature belonging to the class j in a dataset of size N. $\overline{x}$ is the mean value of the observed feature, and $\overline{x_j}$ is the mean value of all records belonging to class j. The total number of records of a class is denoted by n_j. By setting the obtained values for SS_u and SS_i in a ratio, the ANOVA F value is obtained, given by Equation 4, (Sawilowsky, 2002).

$$SS_u = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k}(x_{ij}-\overline{x_j})^2}{N-k} \quad (2)$$

$$SS_i = \frac{\sum_{j=1}^{k} n_j(\overline{x_j}-\overline{x})^2}{k-1} \quad (3)$$

$$F = \frac{SS_u}{SS_i} \quad (4)$$

3. Results and discussion

The dataset contains 23544 records and 547 time windows with MTP. The influence of individual feature and sensor on the trajectory segmentation result was analyzed to find features that better describe the original data. ANOVA F values are assigned to the features indicating their relevance, and the features that provide the most information during the transition between transport modes are highlighted. **Figure 4** shows feature importance, represented as filled circles with two attributes: size, which represents the importance of the feature; the greater the importance is, the larger the circle is and color, from dark blue to dark red, represents the ANOVA F value. The y-axis represents sensors, and the x-axis represents feature symbols. The best ANOVA F score is achieved for the standard deviation and the gravity along the z-axis. Of all sensors, the magnetometer shows the lowest ANOVA F values, while features with a minor influence are the median, skewness, and kurtosis.

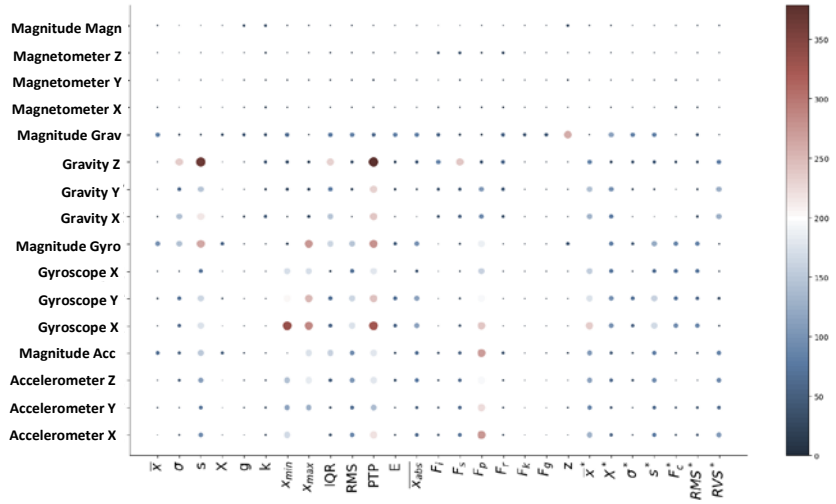

Figure 4. Feature importance using ANOVA F score

In the next step, the required number of features that will be used to predict MTP is determined using the TSM method. The testing process included 200 features with the highest ANOVA F value. Prediction accuracy was observed through measures of the model's overall accuracy, recall

and F score. **Figure 5** shows the testing result for different number of features. The best result is achieved with 80 features (the column marked with a bold black line) where the recall takes one of the highest values, and the decrease in overall accuracy is the smallest compared to the shown examples.

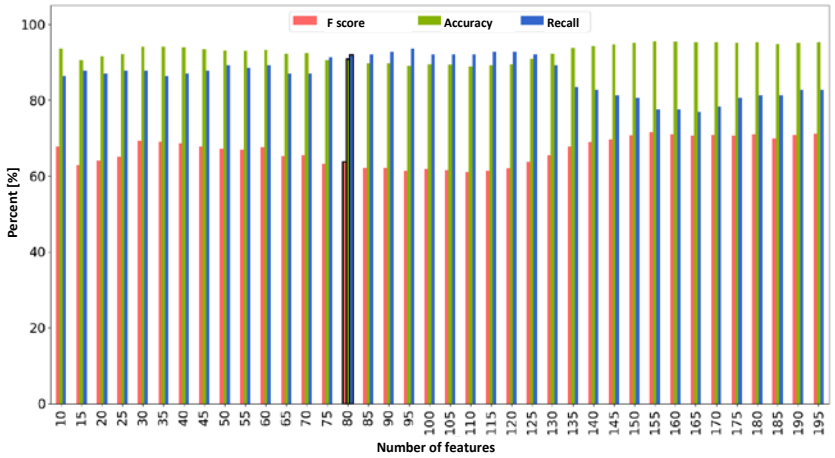

Figure 5. TSM test result for different number of features

Based on the results of the trajectory segmentation method using the identified set of features, it can be concluded that the computed and selected features can describe the user behavior pattern when changing the transport mode. However, to verify the pattern of user behavior and the applicability on other traffic networks, it is necessary to expand the dataset using open datasets on which the validation process can be performed.

Acknowledgements: The authors acknowledge the financial support from TODO project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857592.

References

Erdelić, M., Carić, T., Erdelić, T., & Tišljarić, L. (2022). Transition State Matrices Approach for Trajectory Segmentation Based on Transport Mode Change Criteria. Sustainability 14: 2756. https://doi.org/10.3390/su14052756

Gjoreski, H., Ciliberto, M., Wang, L., Morales, O., Javier, F., Mekki, S., Roggen, D. (2018). The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices. IEEE 42592-42604.

Gong, L., Kanamori, R., Yamamoto, T. (2018). Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. Travel Behaviour and Society 11: 131-140. https://doi.org/10.1016/j.tbs.2017.03.004

Hemminki, S., Nurmi, P., Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. Roma, Italy.

Kalašová, A., Čulík, K., Poliak, M., Otahálová, Z. (2021). Smart Parking Applications and Its Efficiency. Sustainability 13: 6031. https://doi.org/10.3390/su13116031

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: the probable difference between two means when s1 2= s2 2. Journal of Modern Applied Statistical Methods 1(2): 461-472.

Wang, L., Gjoreski, H., Ciliberto, M., Mekki, S., Valentin, S., Roggen, D. (2019). Enabling Reproducible Research in Sensor-Based Transportation Mode Recognition With the Sussex-Huawei Dataset. IEEE 10870-10891.