Main Entrance Tagging based on Intrinsic and Extrinsic Feature Extracted from OSM

Poster · September 2019

DOI: 10.13140/RG.2.2.25967.18080

Xuke Hu
Universität Heidelberg

Hongchao Fan
Norwegian University of Science and Technology

Alexey Noskov
Philipps University of Marburg

Alexander Zipf
Universität Heidelberg

Main Entrance Tagging based on Intrinsic and Extrinsic Feature Extracted from OSM

Xuke Hu [a, b] , Hongchao Fan [c], Alexey Noskov[a], Alexander Zipf [a], Jianga Shang [d]
[a] GIScience Research Group, Heidelberg University, Heidelberg, Germany
[b] Data Science Institute of DLR, Jena, Germany
[c] Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway
[c] School of Geography and Information Engineering, China University of Geosciences, Wuhan, China

1 Motivation

Knowing the location of the main entrance is significant in accurate navigation. Current LBS providers, such as Google Map often guide users to a wrong location that is far away from the main entrance, as shown in Figure 1. This is an unpleasant experience especially for the mobility-impaired people. Realizing the importance of the entrance information, many buildings on OpenStreetMap (OSM) have been tagged with the entrance. However, the proportion is still small. To mitigate this gap, this work proposes using machine learning to inferring the location of the main entrance of a building based on its spatial contexts (e.g., main road) and its footprint (e.g., centroid).

Figure 1. Misleading instruction from Google Map. Locations tagged by blue and red circle are planned target location by Goolge Map and true main entrance, respectively. Yellow line denotes extra efforts needed to finding true entrance.

2 Method

2.1 Workflow

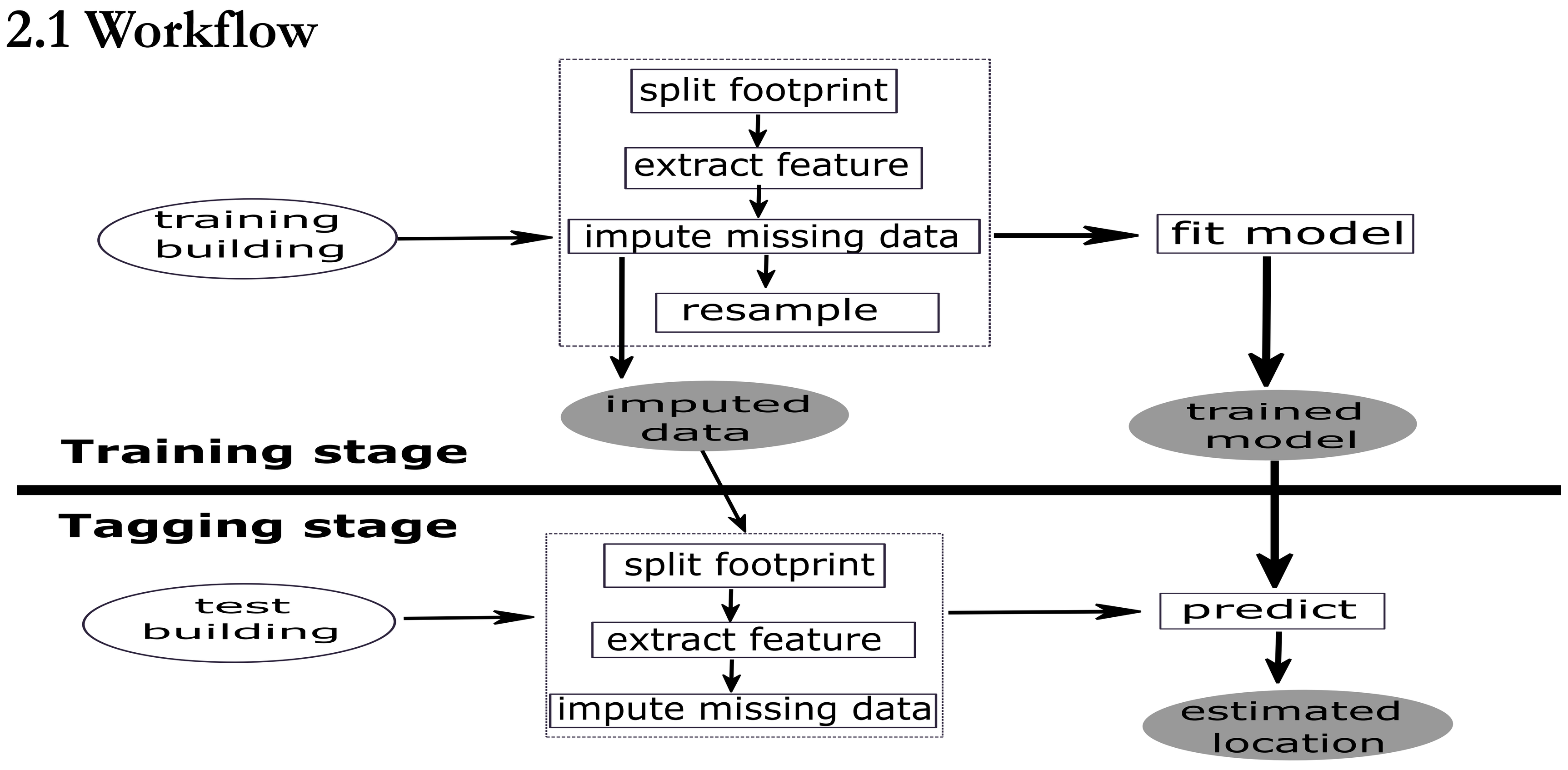
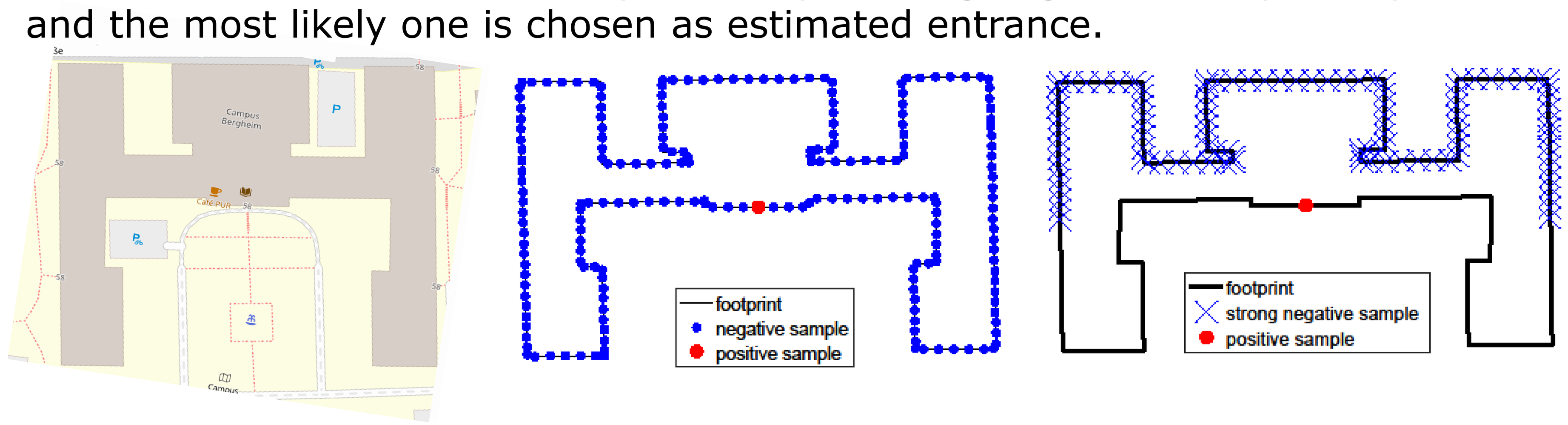


Figure 2. Workflow of proposed method

During training stage, each building is split into points (samples). Then, for each sample, the corresponding features can be extracted by measuring the relationship between the sample and the footprint and surrounding spatial contexts. To mitigate the interference of close negative samples on positive ones (true entrance), the negative ones that have a close physical or feature distance to the positive sample are removed. After collecting the samples of all the buildings, the missing value is filled out by using a strawman imputation strategy [1]. Finally, SmoteBoost [2], Balanced Random Forest [3], and Weighted Random Forest [4] models are fitted. During tagging stage, a building is split into samples in the same way as training stage. Then, the fitted model is used to calculate the probability of assigning each sample as positive and the most likely one is chosen as estimated entrance.



a) Footprint of a building on OSM (b) Split footprint into points tagged with positive and negative, respectively (c) Select **strong** negative sample

Figure 3. Footprint split and strong negative sample selection

2.2 Features

a) Entrance is close to centroid b) Entrance is close to axis c) Entrance is at convex edge

Figure 4. Correlation between entrance location and shape of footprint

Intrinsic features: **Distance to centroid (*), Distance to axis (*) , Proportion on edge, Shape of opposite edge, et al.** Symbol '*' means the sorting result among all the samples is also used.

Table 1. Extrinsic feature extraction by measuring relationship bewtween samples and spatial contexts

	address street	main road	pedestrian way	service way	railway	bicycle parking	landmark	postbox
shortest path distance (*)	√	√	√	√	√	√		
accessible	√	√	√	√	√	√		
degree of visibility (*)	√	√		√	√			
visible	√	√		√	√	√	√	√
Euclidean distance (*)							√	√

3 Experiment

3.1 Tagging accuracy

320 public buildings with the average perimeter at 350 meters are tested based on the five-fold cross-validation. They are collected from seven German cities. For weighted RF approach, the mean error between the true and estimated location in linear distance along the footprint and in shortest path distance are both around **20 m**, and in **80%** of the cases below **30 m**. This can greatly reduce users' effort to finding the entrance.

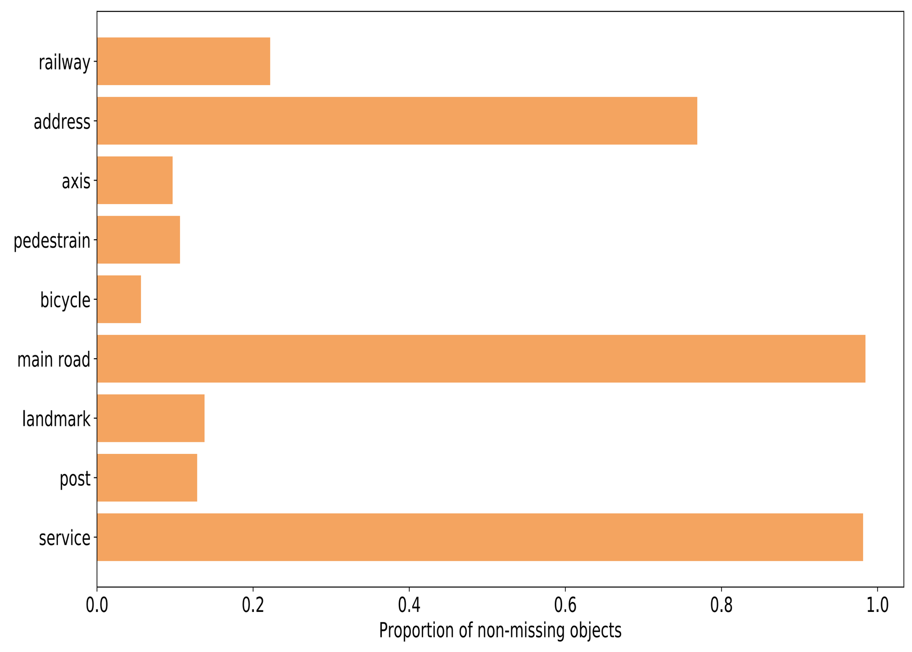


Figure 4. Occurrence frequency of spatial context and symmetric building

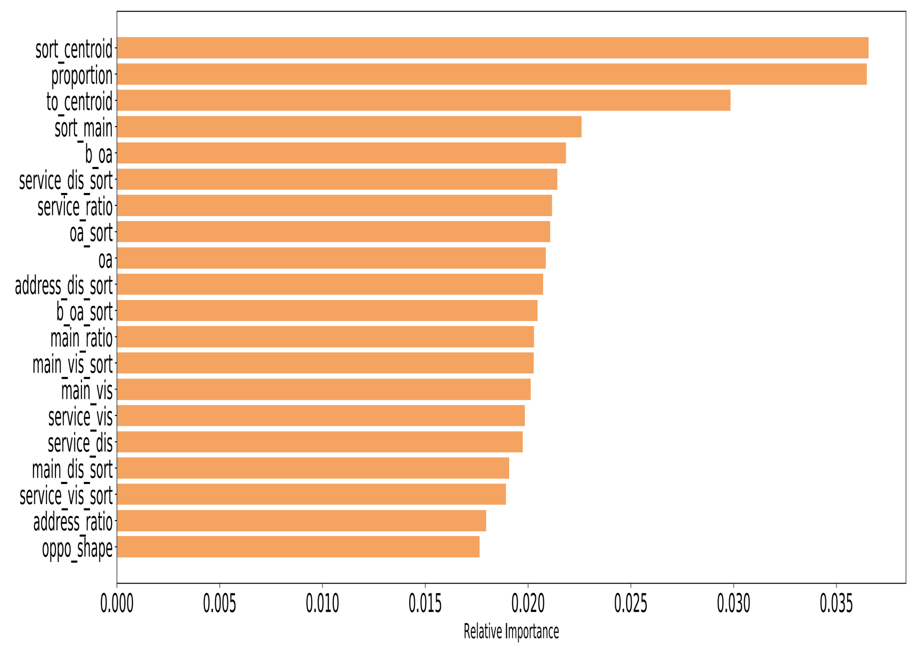


Figure 5. Importance of top 20 features

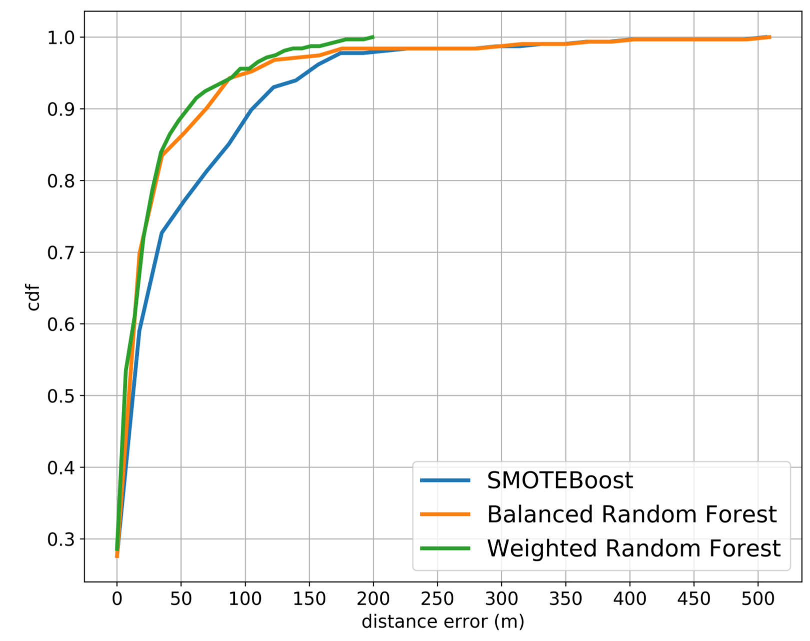


Figure 6. Linear distance error of three approaches

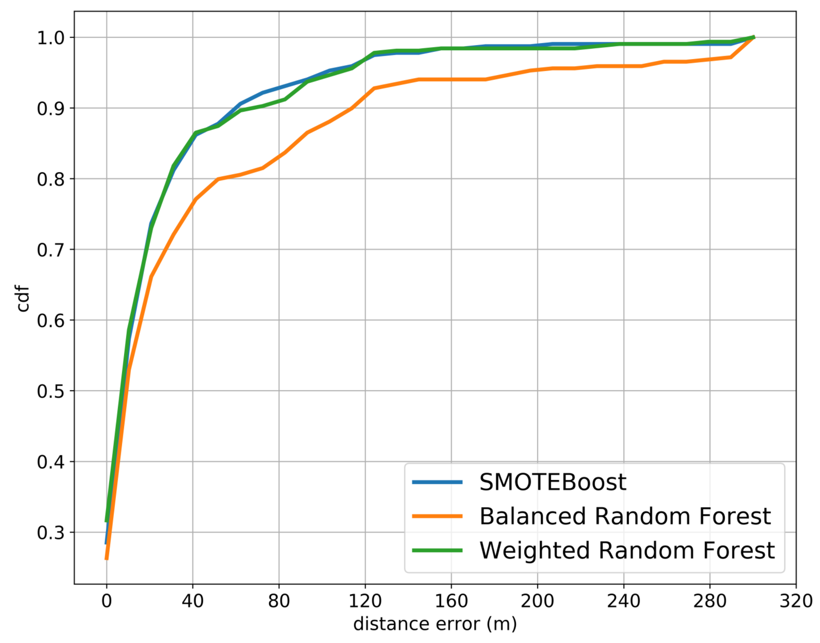
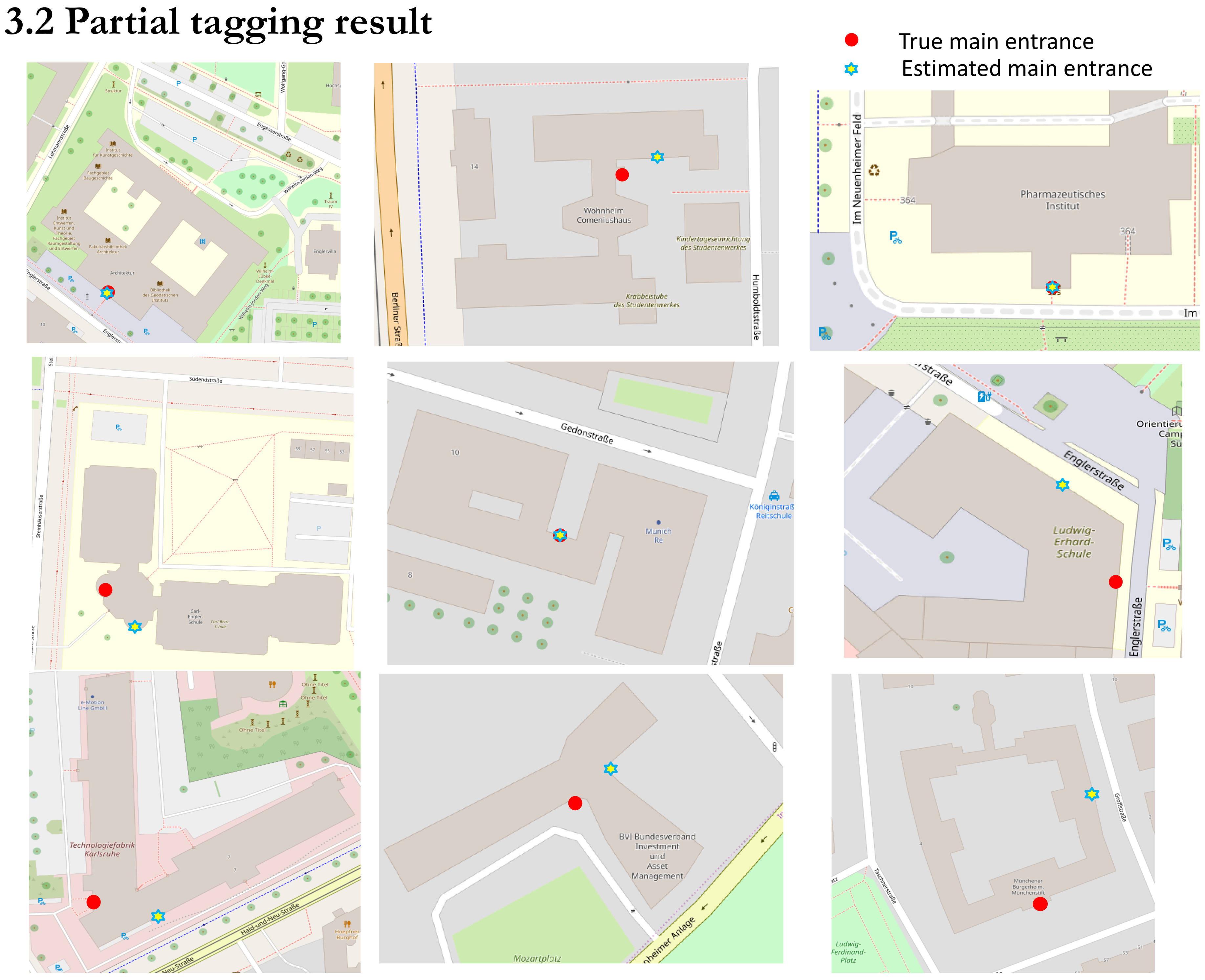


Figure 7. Path distance error of three approaches

3.2 Partial tagging result



References

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
[2] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 51.
[3] Effendy, V., & Baizal, Z. A. (2014, May). Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In *2014 2nd International Conference on Information and Communication Technology (ICoICT)* (pp. 325-330). IEEE.
[4] Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363-377.