Improving Navigation: Automated Name Extraction for Separately Mapped Pedestrian and Cycle Links

Anita Graser and Markus Straub
AIT Austrian Institute of Technology GmbH, Vienna/Austria · anita.graser@ait.ac.at

Full paper double blind review

Abstract

Navigation instructions in pre- and on-trip routing services are usually based on street names and types, distances, and turn directions. However, in digital street graphs it is common that street names for separately mapped pedestrian and cycle links are missing. This leads to unsatisfactory instructions containing "unknown road" records. Often, these unnamed links run parallel to a named road, and it would be beneficial to use this information to generate instructions similar to "follow the sidewalk along Street A", whereby "Street A" has to be determined by an algorithm. This paper introduces the Unnamed Link Naming Problem (ULNP) and presents a new approach to automatically extract suitable names to describe separately mapped pedestrian and cycle links. The approach has been tested using OpenStreetMap data and manually generated ground truth data for the second district of the city of Vienna, Austria. Results show that our best method achieves 90.7% correct matches in this challenging setting.

1 Introduction

Detailed street network datasets, such as OpenStreetMap (OSM), contain separately mapped pedestrian and cycling infrastructure. In this paper, the term "pedestrian links" refers to all network links, which can be used by pedestrians (including, but not limited to, sidewalks, footpaths, and shared-use paths). Accordingly, "cycle links" refers to all links, which can be used by cyclists (including, but not limited to, cycle lanes, segregated cycle facilities, greenways, and shared-use paths). Separately mapped pedestrian and cycle links have a lot of potential for improved pedestrian and cycle routing and navigation. Currently, navigation instructions focus on street names and types, distances, and turn directions. This approach leads to unsatisfactory "unknown road" directions (for example in Google Maps as shown in figure 1) because separately mapped pedestrian or cycle links are often unnamed in the underlying map dataset.

Fig. 1: Google Maps walking directions example referring to an "Unknown road", Screenshot taken in December 2014 (map data © 2014 Google)

GI_Forum – Journal for Geographic Information Science, 1-2015.

Big online routing providers such as Google Maps or Bing Maps tend to treat pedestrian and cycling infrastructure as attributes of the corresponding street link (rendered as green lines on top of the streets, as shown in Google's bicycling overlay in figure 2a). These maps therefore contain considerably fewer instances of separately mapped pedestrian and cycle links than OSM, where mappers tend to map these links separately (rendered as dotted pink and blue lines, respectively, in figure 2b) when they are separated from the corresponding street by infrastructure such as parking lanes, rows of trees, or grass areas. The practice of mapping all pedestrian and cycling infrastructure separately is controversial within the OSM community (OSM TALK-AT 2014), but it can already be observed in some places, for example, in the city of Linz, Austria. In cases where Google has to deal with the problem of unnamed links, they use compass directions as a fall-back solution, e.g. "Unknown road – Head northwest toward Zelinkag", as shown in the example in figure 1.

(a) (b)

Fig. 2: Pedestrian and cycling infrastructure mapping examples from Google Maps (map data © 2014 Google) (a) and openstreetmap.org (© OpenStreetMap contributors) (b), Screenshots taken in December 2014

While current OSM-based routing services, such as Graphhopper, can take advantage of the separately mapped pedestrian and cycling infrastructure, they do not deal with the issue of missing names, as can be seen in the following example: the cycling directions (figure 3a) are limited to "continue, cycleway" even though there is a turn in the route. This turn is described correctly in the car route (figure 3b), which also contains street information.

Since we want to use OSM to leverage the available detailed data regarding pedestrians and bicycle traffic, we have to develop a new approach to deal with the issue of unnamed separately mapped pedestrian and cycle links. This new approach makes it possible to generate route descriptions such as "follow the bicycle path along Street A" by computing which (if any) named street a pedestrian or cycle link belongs to. To the best of our knowledge, this issue has not been discussed in the literature so far. The closest related topic is line matching, also known as conflation.

(a) (b)

Fig. 3: Graphhopper.com cycling directions based on OSM for cycling (a) and car (b), Screenshots taken in December 2014 (© OpenStreetMap contributors, Lyrk)

2 Literature Review

Line matching or conflation is the process of finding the same real world object (often a street) in different datasets. In the literature, there are several approaches for line matching. The approaches can be roughly categorized into two groups: one group works with line geometries only, while the other group adds additional information such as attributes or metadata such as data quality indicators (COBB et al. 1998) to the matching process.

Geometry-only approaches described, for example, in DOYTSHER et al (2001) use shape similarity, cumulative distance, and topological similarity between both end-points. Another geometry-only approach for matching road features from different datasets based on the locations of the endpoints of the polylines is presented in SAFRA et al. (2006). A common step in many approaches is to use buffers to find matches (WARE & JONES 1998, GABAY & DOYTSHER 2000). A polyline from one source is augmented with a buffer. When a polyline from the other source is completely contained within that buffer, the two polylines are considered to be a matching pair. Similarly, buffers can be used to filter out impossible matches, as shown in SESTER et al. (1998) and WALTER & FRITSCH (1999).

Matching algorithms for matching linear data in VGI (Volunteered geographic information) research (for example KOUKOLETSOS et al. 2012) use a combination of attribute and geometric constraints to match linear features in different datasets. Geometric constraints include distance, orientation, and length, while attribute constraints focus on road names and types. Other approaches are based on semantic similarity (AL-BAKRI & FAIRBAIRN 2012).

In order to solve the Unnamed Link Naming Problem (ULNP), that is the problem of finding which named street an unnamed pedestrian or cycle link belongs to, we base our approach on existing line matching methods. However, it is necessary to adapt these existing approaches to finding a similar object in the same dataset instead of finding the same real world object in a different dataset. To this end, we build on geometry-only line matching approaches. Line matching methods which use attributes such as name or road class to determine the match are not appropriate for this use case, since neither name nor class of an unnamed pedestrian or cycle link and the best matching named link will be similar.

The following section describes the three developed methods. Section 4 provides an introduction to the data used for method development and evaluation. Section 5 presents and discusses the evaluation results, and section 6 provides an outlook for future work.

3 Methods

This section presents the three developed methods. Conflation approaches using buffers to find matching features served as a starting point for the method development, since they appear most suited to the task of solving the ULNP. Each successive method has been developed to address the shortcomings of the previous methods. All methods and their results (see section 5) are presented to illustrate the improvements gained through more sophisticated matching approaches. All methods use a common preprocessing step, which inserts additional geometry nodes at 1 meter intervals to ensure that distance computations between links will not be affected by long stretches without intermediate nodes.

3.1 Hausdorff Distance Matching

Buffer computations and successive containment operations are computationally expensive. Therefore, we decided to use the Hausdorff distance (HAUSDORFF 1927) instead. This is possible since – for sufficiently densely sampled line geometries – a check for containment within a buffer with a size of x meters leads to the same results as a check for a Hausdorff distance smaller than x meters. The Hausdorff distance is the maximum distance of a set to the nearest point in another set. More formally, the Hausdorff distance from set A to set B is defined as

$$h(A,B) = \max_{a \in A}\{ \min_{b \in B}\{ d(a,b)\}\}, \tag{1}$$

where a and b are points of the sets A and B respectively, and $d(a,b)$ is the Euclidian distance in our approach. Since the Hausdorff distance is asymmetric (in general $h(A,B)$ is not equal to $h(B,A)$), the more general definition of Hausdorff distance, which we use to calculate the matching score, is

$$M(A,B)_H = H(A,B) = max\{ h(A,B), h(B,A) \}. \tag{2}$$

This method therefore finds the most similar line feature in another set of line features, i.e. the set of named links, for each link in the set of unnamed links. Additionally, we used a maximum distance tolerance (δ). If the best match – with the smallest score – exceeds this tolerance value, the link is assigned no name.

3.2 Median Distance Matching

One disadvantage of Hausdorff distance matching is that it is sensitive to outliers because it measures the maximum minimum distance (1). Therefore, we developed a second method, which uses the median minimum distance instead. The matching score is defined as

$$M(A,B)_M = \operatorname{median}_{a \in A}\{ \min_{b \in B}\{ d(a,b)\}\}. \tag{3}$$

Similar to Hausdorff distance matching, the link is assigned no name if the best match exceeds the set distance tolerance (δ). Since this approach does not take link orientation into account, the issue of matching to perpendicular links remains.

3.3 Composite Matching: Distance and Orientation

The composite matching method matches features based on the smallest matching score consisting of weighted median distance and orientation difference. It is defined as

$$M(A,B)_C = \frac{\text{median}_{a \in A}\{ \min_{b \in B}\{ d(a,b)\}\}}{\delta} * w_d + \frac{orientation\ difference(A,B)}{180} * w_o, \quad (4)$$

where w_d is the weight of the distance term, w_o is the weight of the orientation term, and δ is the distance tolerance. The orientation difference is a value between 0 and 180 in degrees, which represents the difference in azimuth measures between the first and last point of A and B respectively. For the purpose of this comparison, a line digitized in east-west direction is equal to a line digitized in west-east direction. Therefore, link directions are not considered in azimuth computations.

This method checks for both a distance tolerance (δ) and an angular tolerance (φ). The distance tolerance is used to remove links that are too far from any named link. The angular tolerance is used to remove "close to perpendicular" links. In composite matching, δ is applied to the minimum distance in order to handle cases where the unnamed link extends or partially overlaps the corresponding named link. The unnamed link is assigned no name if it is not possible to find a named link that fits both requirements.

4 Input Data and Data Processing

For the development and evaluation of the developed methods, we used the cycle network of the second district of Vienna, Austria. (OSM data was downloaded on Feb. 14th 2014.) The second district was chosen since it provides a challenging setting with different street network designs, as well as unnamed links, which should not be matched to any named link. Data preprocessing is necessary to turn OSM data into a routable network by splitting links at the appropriate intersections. Furthermore, driving permissions were evaluated to create the network of all links available for cycling.

For the validation, a ground truth dataset of 804 separately mapped cycle links was generated manually by visually matching unnamed cycling links to the corresponding named link. This matching can be described as "assigning a name which a human would use to describe the route". Overall, 778 (96.8%) of the 804 separately mapped cycle links are unnamed. The remaining cycle links already have a name assigned to them in OSM.

The relationship between cycle link and corresponding named link can take any of the four types of spatial relationships between polylines described by SAFRA et al. (2006): (1) complete overlap (two pairs of corresponding endpoints), (2) extension (one pair of corresponding endpoints and the other endpoint is an intermediate point), (3) containment (both endpoints of one line are intermediate points of the other), and (4) partial overlap (each line has an intermediate point in the other). The network contains some complex situations where a human observer would identify several potential matches. Since we decided not to introduce additional links by splitting existing links in ambiguous situations (for example where

two thirds of the cycle link run parallel to street A and one third runs parallel to street B), the links were assigned the name that best describes them.

In those cases where the unnamed cycle link does not have a matching named link because there is no appropriate link nearby, we recorded "none" into the name column of the ground truth dataset. In total, 400 (49.8%) of the 804 cycle links do not have a matching named link. The optimal automatic matching method should avoid matching these links.

5 Results and Discussion

All three methods were evaluated using the ground truth dataset of manually matched cycle links and different values of distance and orientation tolerance. However, the methods can be applied to pedestrian links as well. A summary of the evaluation results is shown in table 1. The highest percentage of correct matches for each method is highlighted in bold font. Darker background colours mark worse results. The baseline for this evaluation, which can be achieved by not matching any cycle links, is 53% (since 49.8% should not be matched and 3.2% are already named).

Table 1: Evaluation results: percentage of correct matches for the three presented methods with varying parameter settings

Distance tolerance [m]	Hausdorff dist.	Median dist.	Composite matching depending on orientation tolerance							
			5°	10°	15°	20°	25°	30°	35°	40°
5	54,0	53,2	58,1	59,1	59,2	59,0	58,8	58,5	58,1	58,2
10	66,8	67,5	75,7	79,5	80,3	79,9	79,7	79,2	79,0	79,0
15	75,5	**76,9**	86,2	89,4	90,0	89,9	89,8	88,9	88,4	87,8
20	**76,6**	**76,9**	87,4	90,5	**90,7**	89,7	89,4	88,6	88,1	87,7
25	76,1	75,2	87,6	**90,7**	90,3	89,4	89,1	88,2	87,7	86,9
30	75,5	74,0	87,1	90,0	89,7	88,4	88,1	87,2	86,7	85,9
35	75,5	71,4	86,6	88,6	88,1	87,1	86,8	85,6	85,3	84,5
40	74,6	69,0	85,8	87,8	86,6	85,3	84,7	83,3	82,8	82,0

Up to 76.6% of cycle links in the test dataset are matched correctly using **Hausdorff distance matching**. The best results are achieved for a distance tolerance δ of 20 meters. The detailed evaluation in table 2 shows that 66.7% of the links that should be matched were matched correctly. The errors are distributed equally between missing matches (17.7%) and wrong matches (15.6%). Additionally, we can also see that the algorithm did match 15.5% of the links that should not be matched.

Table 2: Hausdorff distance matching results for $\boldsymbol{\delta = 20}$: number of links per category (matching errors are marked with a red background)

	correct name	no match	wrong name	sum
already named	26 (100%)			26
should be matched	252 (66.7%)	67 (17.7%)	59 (15.6%)	378
should not be matched		338 (84.5%)	62 (15.5%)	400

Some notable errors include situations such as the one depicted in figure 4: Hausdorff distance matching fails to find a suitable match for the cycle lane along Untere Donaustraße (depicted by the red dotted line in figure 4a), because the cycle lane is modelled as one continuous link, while Untere Donaustraße is divided into multiple links. This modelling difference results in a Hausdorff distance that exceeds the distance tolerance.

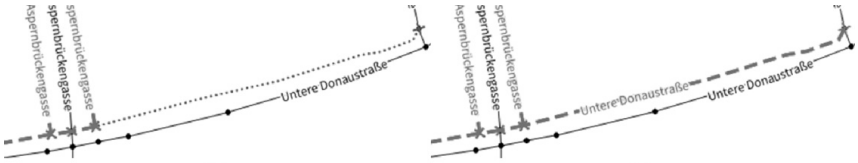
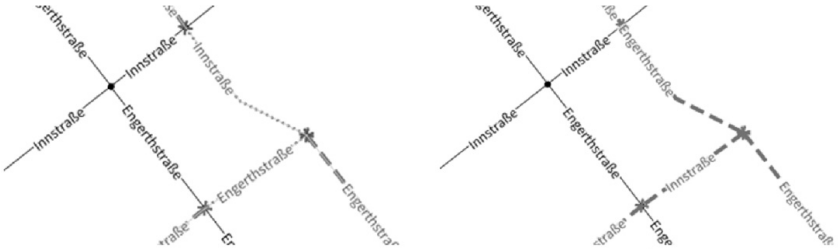


Fig. 4: Hausdorff distance (a) and Median distance (b) matching results (map data © OpenStreetMap contributors)

Median distance matching succeeds in matching most cases of link extension, containment, and partial overlap. For example, the cycle lane along Untere Donaustraße is matched correctly, as illustrated in Figure 4b. Overall, up to 76.9% of the cycle links are matched correctly using Median distance matching. The best results are achieved for a distance tolerance δ of 15 or 20 meters. For links that should be matched, the detailed evaluation in table 3 shows that this approach increases the share of links that were matched correctly to 75.1% and decreases the share of unmatched ones to 0.5%. At the same time, the share of incorrectly matched links increases by around 8% for both links that should be matched, and those that should not. Since Median distance matching does not take link orientation into account, this approach leads to incorrect matches when the cycle links are matched to perpendicular named links due to the smaller median distance (see figure 5a).

Table 3: Median distance matching results for $\boldsymbol{\delta = 20}$: number of links per category

	correct name	no match	wrong name	sum
already named	26 (100%)			26
should be matched	284 (75.1%)	2 (0.5%)	92 (24.3%)	378
should not be matched		308 (77.0%)	92 (23.0%)	400



(a) (b)

Fig. 5: Issues with perpendicular links using Median distance matching (a) and improved results using Composite matching (b) (map data © OpenStreetMap contributors)

For the evaluation of the **Composite matching** method, distance and orientation weights w_d and w_o were both set to 1 to assign both terms equal weight. The evaluation shows that up to 90.7% of the cycle links are matched correctly using this method. The best combination of tested parameters are $\delta = 20$, $\varphi = 15$ and $\delta = 25$, $\varphi = 10$. The detailed evaluation in table 4 shows that Composite matching increases the share of correct matches, and decreases the number of wrong matches. It increases the share of correctly matched links that should be matched to 92.6%, and decreases the share of incorrect matches to 2.1%. Additionally, it also correctly avoids matching 88.3% of links that should not be matched. Future work should test different w_d and w_o values to evaluate if further improvements can be achieved.

Table 4: Composite matching results for $\boldsymbol{\delta = 20}$, $\boldsymbol{\varphi = 15}$: number of links per category

	correct name	**no match**	**wrong name**	**sum**
already named	26 (100%)			26
should be matched	350 (92.6%)	20 (5.3%)	8 (2.1%)	378
should not be matched		353 (88.3%)	47 (11.8%)	400

One disadvantage of Composite matching is that it tends to match cycle links to named links, which end where the cycle link starts. Figure 6a shows an example of this issue: the cycle link is matched to Rotundenplatz even though it extends in the opposite direction. This is caused by the decision to apply the distance tolerance to the minimum distance. The minimum distance was chosen instead of, for example, the mean distance, because we would otherwise run into issues with extension and partial overlap cases where the cycle link extends past the corresponding named link. One approach to solve this issue would be to dissolve named links based on their name in order to create a single long link with the same name. Unfortunately, it is not possible to guarantee that the dissolve result will be one continuous line. For example in figure 6a, it is easy to imagine that dissolving all Rotundenplatz links would result in a complex MultiLinestring with multiple forks and loops. This makes it impossible to compute meaningful orientation values for Composite matching.

Figure 6b shows a different situation at Mortaraplatz, which cannot be solved using the current approaches. A single continuous cycle link runs along Leystraße and continues along Mortaraplatz. In order to compute meaningful matches, it would be necessary to split the cycle link so that one part can be matched to Leystraße and the other part to Mortaraplatz. Of course, it can be argued that the best instructions in this case would be "follow the cycle lane along Leystraße" since Leystraße continues after Mortaraplatz.

The majority of the 75 incorrectly matched links are very short, with a median length of 14.8 meters (30 links under 10 meters), while the median length of all links in the sample is 46.0 meters. The majority of these short links is found close to intersections, and their orientation tends to differ too much from the orientation of the named link in the ground truth dataset. At the roundabout at Praterstern, 16 (21.6%) of 74 links are matched incurrectly. The cycling infrastructure at Praterstern is characterized by multiple intersections, where the cycle lane crosses roads leading up to the roundabout. While a human observer would match the whole circle of cycle links around the roundabout to Praterstern, the matching algorithm fails at multiple occasions. The matching errors are mostly due to orientation differences that exceed the angular tolerance. In the Prater park area, 16 (8.0%) of 201 links are matched incorrectly. All but one of the errors in this area are due to the algorithm assigning names to links that should not be matched. In the majority of cases, the difference between human and algorithm matches can be traced back to extensions (as illustrated in figure 6a), and instances where a human observer would take into account that cyclists will have difficulty seeing the named link through the trees separating cycle link and named link.

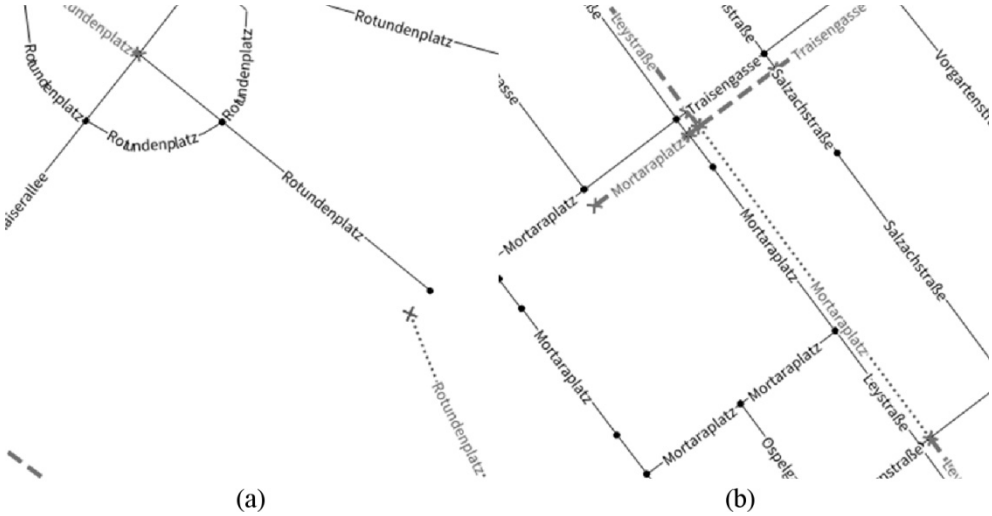


Fig. 6: Challenging network situations at Rotundenplatz (a) and Mortaraplatz (b) (map data © OpenStreetMap contributors)

6 Conclusion and Outlook

To be able to generate useful navigation instructions for routes containing separately mapped unnamed pedestrian or cycle links, it is necessary to develop methods to automa-

tically extract suitable names from available street network data. We presented three methods for finding which named street a pedestrian or cycle link belongs to. The most successful method, Composite matching – combining median distance and orientation difference – succeeded in matching 90.7% of all cycle links in the second district of Vienna to the correct named street. Based on the detailed evaluation of error sources, it is expected that the algorithm will perform better for networks where streets run at right angles to each other, forming a grid, than for networks with organically grown street patterns. For unnamed links at roundabouts, a local relaxation of the angular tolerance might lead to better results.

Future developments should address the issues of unnamed links being matched to named links, which end where the unnamed link starts, as well as identifying those situations where unnamed links have to be split in order to be able to compute appropriate matches. Furthermore, the approaches should be tested in other cities to evaluate their transferability and avoid overfitting of parameter values to a specific situation.

Another potential application of the developed methods is street graph generalization. The methods could be used to enrich the generalized link with attributes from all matching links. This could enable the automatic inference of street cross-section characteristics such as the number of carriageways, or the presence of pedestrian and cycling infrastructure.

Acknowledgements

This work is partially funded by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the programme “Mobilität der Zukunft” under grant 844434 “PERRON” as well as the Vienna Business Agency within the call “From Science To Products 2013” under the grant for project “sproute”.

References

AL-BAKRI, M. & FAIRBAIRN, D. (2012), Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. International Journal of Geographical Information Science, 26 (8), 1437-1456.

COBB, M. A., CHUNG, M. J., FOLEY, H., PETRY, F. E. & SHOW K. B. (1998), A rule-based approach for conflation of attribute vector data. GeoInformatica, 2 (1), 7-33.

DOYTSHER, Y., FILIN, S. & EZRA, E. (2001), Transformation of datasets in a linear-based map conflation framework. Surveying and Land Information Systems, 61 (3), 165-176.

GABAY, Y. & DOYTSHER, Y. (2000), An approach to matching lines in partly similar engineering maps. Geomatica, 54 (3), 297-310.

HAUSDORFF, F. (1927), Grundzüge der Mengenlehre. Walter de Gruyter, Berlin.

KOUKOLETSOS, T., HAKLAY, M. & ELLUL, C. (2012), Assessing data completeness of VGI through an automated matching procedure for linear data. Transactions in GIS, 16 (4), 477-498.

OSM TALK-AT (2014), Austrian OpenStreetMap mailing list, Mappen von Gehsteigen. http://lists.openstreetmap.org/pipermail/talk-at/2014-February/006402.html.

ROSEN, B. & SAALFELD, A. (1985), Match criteria for automatic alignment. In: Proceedings of 7th International Symposium on Computer-Assisted Cartography, 1-20.

SAALFELD, A. (1988), Conflation-automated map compilation. International Journal of Geographical Information Systems, 2 (3), 217-228.

SAFRA, E., KANZA, Y., SAGIV, Y. & DOYTSHER, Y. (2006), Efficient integration of road maps. In: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, 59-66.

SESTER, M., ANDERS, K. H. & WALTER, V. (1998), Linking objects of different spatial data sets by integration and aggregation. GeoInformatica, 2 (4), 335-358.

WALTER, V. & FRITSCH, D. (1999), Matching spatial data sets: a statistical approach. International Journal of Geographical Information Science, 13 (5), 445-473.

WARE, J. M. & JONES, C. B. (1998), Matching and aligning features in overlayed coverages. In: Proceedings of the 6th ACM International Symposium on Advances in Geographic Information Systems, 28-33.