Do You Need Instructions Again? Predicting Wayfinding Instruction Demand

Do You Need Instructions Again? Predicting Wayfinding Instruction Demand

Negar Alinaghi
Geoinformation, TU Wien, Austria

Tiffany C. K. Kwok
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Peter Kiefer
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Ioannis Giannopoulos
Geoinformation, TU Wien, Austria
Institute of Advanced Research in Artificial Intelligence (IARAI), Austria

Abstract

The demand for instructions during wayfinding, defined as the frequency of requesting instructions for each decision point, can be considered as an important indicator of the internal cognitive processes during wayfinding. This demand can be a consequence of the mental state of feeling lost, being uncertain, mind wandering, having difficulty following the route, etc. Therefore, it can be of great importance for theoretical cognitive studies on human perception of the environment. From an application perspective, this demand can be used as a measure of the effectiveness of the navigation assistance system. It is therefore worthwhile to be able to predict this demand and also to know what factors trigger it. This paper takes a step in this direction by reporting a successful prediction of instruction demand (accuracy of 78.4%) in a real-world wayfinding experiment with 45 participants, and interpreting the environmental, user, instructional, and gaze-related features that caused it.

2012 ACM Subject Classification Computing methodologies → Activity recognition and understanding; Computing methodologies → Supervised learning by classification

Keywords and phrases Wayfinding, Navigation Instructions, Urban Computing, Gaze Analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.1

Supplementary Material *Dataset*: `https://geoinfo.geo.tuwien.ac.at/resources/`

Acknowledgements We would like to thank our colleagues from Vienna University of Technology, Dr. Markus Kattenbeck, for suggesting the use of the Big Five Personality Traits test, and Antonia Golab for collecting the valuable data used for this work.

1 Introduction

Human-computer interaction (HCI) in wayfinding and pedestrian navigation has attracted much attention in recent years [27]. Reducing cognitive load in a complex task such as wayfinding is an important goal in this domain. Efforts in this area range from working on the structure, content, and presentation of navigation information to better adapting it to the needs of users. The current research trend in navigation assistance systems provides instructions in various modalities, from conventional turn-by-turn instructions with map visualization to auditory instructions based on landmarks and, more recently, visual instructions with augmented reality (e.g., [44, 10]). The performance of these different modalities is evaluated using various metrics including travel time, number of errors, deviation from shortest/fastest path, cognitive demand, subjective ratings, etc. (see Section 2). One possible indicator which is less explored is how often the user asks for the instruction after first receiving it, i.e., the instruction demand. The behavior of requesting more information

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 1; pp. 1:1–1:16
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

or the same information again can also be considered as an indicator of how cognitively demanding the processing of this information is. However, we still do not know whether such behavior is caused by the complexity of the environment, the user's personal characteristics or spatial abilities, the content of the instructions and how they are conveyed, or a combination of all these factors.

Being able to predict the overall demand for instructions shortly after receiving a navigation instruction, can be very useful in developing more customized navigation aids. On the other hand, knowing why people need more instructions and what factors trigger this demand in users is of great importance for cognitive studies on human perception of the environment. For instance, the need for repetition of navigation instructions can be seen as a prototypical behavior of feeling lost or needing reassurance. Therefore, predicting this need can as well help us predict these cognitive states during wayfinding. Theoretical studies and empirical evidence suggest that cognitive demand in wayfinding is strongly influenced by user-, environment-, and assistance-related factors [39, 15]. Requesting instructions is one of the many activities performed during this process and can be considered a proxy for cognitive demand. Machine learning (ML) has shown promising results, not only in predicting such activities but also in partially explaining the predictions. The latter is crucial when it comes to interpreting the results and extending knowledge about the causality of actions, here e.g., instructional needs.

In this paper, we show that instruction demand in a pedestrian wayfinding experiment can be predicted with reasonably high accuracy using ML techniques. This prediction is not a black box; rather, our results suggest that it can be interpreted by environmental features, user- and instruction-related features, and gaze features. These findings are the result of an outdoor wayfinding experiment conducted in the city with 45 participants navigating to two different known and unknown destinations using an audio-assisted landmark-based navigation system. Participants were allowed to ask for auditory instructions at any time and as often as they wished. While walking, their behavioral data in the form of eye movements and trajectory were recorded by eye-tracking glasses and a high-precision GNSS antenna (see Section 3.1).

We trained several classifiers, namely Support Vector Machines (SVM), RandomForest, and XGBoost on a variety of gaze features, environmental features, user- and instruction-related features. Our analysis shows that instruction demand can be predicted shortly after the first instruction (within two seconds), mainly based on the complexity of the environment and user characteristics. Through several experimental setups, we found that a minimal subset of 21 features (a combination of the above factors) leads to an accuracy of 78.9% on unseen data, making our prediction approach beneficial for real-time applications and giving us a better understanding of why people may need more informative assistance in wayfinding.

2 Related Work

With the goal of predicting instruction needs in an auditory-aided wayfinding task, we examined the existing literature from several perspectives: the evaluation and perception of navigation instructions, the processing of instructional information using gaze analysis, and exploring the use of machine learning in predicting wayfinding activities and states.

2.1 Research on Navigational Instructions

Efforts have been made to optimize wayfinding support systems through the structure of instructions [25, 41, 35], use of landmarks [34, 31, 12], and modality of information presentation [20, 10]. One of the first papers to address the conceptualization of route instructions is by Klippel et al. [24], who proposed a set of wayfinding choremes as mental conceptualizations of route guidance elements used to simplify visualization of turn-by-turn information. Another method to reduce instruction complexity is the Spatial Chunking method, where unnecessary instructions are chunked together to reduce complexity [23]. A second aspect besides simplifying instructions is how much spatial information they convey. Krukar et al. [26], addressed this matter by introducing orientation instructions that combine local and global route information. In a study with 84 participants, they evaluated the performance of these instructions by measuring the memorability of the instructions and showed that they conveyed survey information without interfering with the retrieval of route information. A systematic review of navigation systems for people with dementia was conducted by Pillette et al. [32] to compare common evaluation standards. They reviewed 23 papers, including indoor, outdoor, and VR-based experimental designs, in terms of presentation modality, navigation content, and timing of presentation. Most objective measures introduced for evaluation were the number of errors, time to complete the navigation task, arrival at destination, and the number of times participants asked questions or received outside assistance. Most subjective measures were obtained either by experimenters observing participants' behavior, hesitation, or difficulty in completing the task, or by interviews and questionnaires.

The work most closely related to ours is that of Golab et al. [17]. The authors used survival analysis to model the times at which participants needed navigation instruction, accounting for personal, environmental, and route-related variables. They reported that "participants request a route instruction later as a function of their age, on segments longer than 120m and in unfamiliar conditions if they score below average on the personality trait extraversion." This is initial evidence from a real-world wayfinding experiment, reporting user-related and environmental aspects to be influential on the timing of the instructions.

2.2 Gaze Analysis in Instruction-based Wayfinding

Eye tracking provides insight into human cognitive processes [18, 11] complementing standard behavioral responses. Previous work has shown the potential of gaze for user modeling, such as including user's attention [1], cognitive load assessment [6], and spatial decision making [40, 22]. The knowledge obtained from user modeling can further be used for adapting the system behavior [16]. An example of gaze analysis in wayfinding is Brügger's study [8] where they examined the impact of navigation system behavior on human navigation behavior and performance in an outdoor experiment with 64 participants. They measured cognitive function and scene complexity using fixation frequency and duration, finding that average fixation duration varied across different tasks, i.e., incidental knowledge acquisition and knowledge retrieval.

De Cock et al. [10] used gaze behavior analysis to investigate the nature of navigation instructions in an indoor experiment with a VR-based adaptive route guidance system. They focused on photos or icons at start and end points, and photos or 3D simulations at turn junctions. They mainly used dwell time for their analysis and found that the detailed information of a static photo instruction is more difficult to transfer to the environment. Very recently in a paper by Ludwig et al. [29], the instruction needs in a real-world indoor multi-level wayfinding experiment were analyzed using the normalized mean square error between the observed dwell time distribution and its estimation from the distribution of the aggregated fixations between two routing instructions which were generated automatically by their system incorporating the most salient landmark close to the user. Their result suggested that instruction need tends to increase when there is a change in direction or level. These papers show practical evidence that gaze analysis can reveal aspects of the cognitive demands of processing navigation instructions.

2.3 Machine Learning for Wayfinding Activities' Prediction

The use of ML techniques for prediction and classification tasks in wayfinding has become increasingly important in recent years. Alinaghi et al. [4] predicted the direction in which the wayfinder would like to turn a few seconds before the turn action is performed, based on gaze behavior and environmental complexity. They tested several ML techniques, including SVM, DecisionTree, and XGBoost, and reported that XGBoost outperformed the other two with 91% accuracy. In another paper [3], the authors analyzed the effect of familiarity with the environment using the pre-trained XGBoost model from their earlier work and were able to show that the gaze behavior of familiar and unfamiliar wayfinders differed as they approached a turn decision point. There, the authors introduced a terminology as *matching-to-action* phase of wayfinding, which in their context refers to the part of the route from the point where an instruction is given to the turn to which the instruction refers. We have segmented our trial routes based on this notion (see Section 3.2). Liao et al. [28] trained a RandomForest classifier with gaze features from 38 participants in a real-world outdoor study to classify five common wayfinding tasks: Self-location and orientation, target search in the local environment, target search on the map, route memorization, and walking to the target, with an overall accuracy of 67%.

Another related experience with the RandomForest classifier is reported by Zhu et al. [44]. In a VR-based indoor study with 30 participants, the authors collected EEG recordings from participants performing a series of 10 wayfinding trials of varying difficulty, each exploring a portion of the virtual reality model (i.e., different sets of origin and destination locations were defined at different distances and on multiple floors for each trial). Using a combination of objective measurements (e.g., frequency of inputs to the VR controller) and behavioral recordings from two independent observers, a classifier was trained to predict uncertainty time segments during navigation trials. The overall predictive power of the model was reported to be 0.70 as measured by the area under the Receiver Operating Characteristics Curve (ROC-AUC). Although most work on activity recognition in the wayfinding domain uses the RandomForest model due to its simplicity in training and explainability, higher performance with other models is also reported in the broader domain of human activity recognition. For example, XGBoost, as an ensemble model of tree-based architectures that can better model more complex relationships and is as explainable as other tree-based models, has been successfully used in the literature for many different tasks (see, e.g., [5, 43]).

3 Data Collection, Pre-processing and Feature Extraction

This section summarizes the details of the data collection procedure, the pre-processing steps, and finally the feature extraction methods. The data we report on here was collected in 2020 and was first described in [17][1]. Parts of the data have already been used for analyses

[1] Parts of the data used in the current paper, will be made available at: `https://geoinfo.geo.tuwien.ac.at/resources/` (DOI: 10.5281/zenodo.4298703).

published in [4], [3] and [2]. The original dataset has 104 trials recorded from 52 participants (27 female and 25 males, $M(age) = 26$ years, $SD(age) = 8.3$). However, due to some sensor malfunctioning and data loss in eye-tracking data, here we have analyzed 71 trials from 45 participants.

3.1 Data Collection

The study was a within-subject experiment with two phases: an online phase for registration, demographic, Big Five personality traits [33], and the Spatial Strategies Questionnaire (FRS) [30] data collection; and an in-situ phase for recording participants' eye movements (using PupilLabs Invisible glasses with 200 Hz recording frequency) and trajectory data (using a PPM 10-xx38 GNSS receiver) as they walked familiar and unfamiliar routes[2]. The familiar trial was conducted in a region and to a destination that the participant reported being completely familiar with, as opposed to the unfamiliar trial. In both cases, however, participants were asked to walk a pre-calculated route by following the auditory turn-by-turn, German-language, landmark-based navigation instructions[3] provided to them upon request. Participants were given a clicking device that they held in their hands and could click on when and as often as they wanted. The obligation to follow the predefined route (which was unknown to participants) prompted all participants to request navigation instructions to find their way, whether or not they were in the familiar condition. This design provided us with the opportunity to examine primarily *when* they need instructions and then *why* they need the same instruction more than once.

3.2 Data Pre-processing

To determine the motives behind instruction demand, we processed four of the data sources: GPS tracks to obtain the environmental features, online-phase data to obtain the demographics, personality traits, and spatial strategy scales, eye-tracking records for gaze behavior, and navigation instructions for their length and content. Of these, we only needed to preprocess the GPS tracks to match them with the Open Street Maps (OSM) data and extract urban complexity measures from them. Figure 1 depicts the six steps required for this preprocessing. First, we cleaned the GPS data and smoothed it by preserving the timestamps. Then, using the intersection framework [14], we extracted street intersections from the OSM data and matched them to the GPS tracks. The instruction-request events were also matched to the GPS tracks based on their timestamps. Then, we segmented the route based on the idea of *matching-to-action phase* [3] (See Subsection 2.3) with the small difference that we do not start this segment with the instruction request event, but with the previous turn intersection to also track how far the request event is from the last turn decision point, and we call the outcome chunks of the route *"unified segments"*. We also call the part of the route between any two intersections (whether it is a turn or not) a *"segment"*. Finally, a buffer of $30m$ (large enough to cover the buildings and Points of Interest (PoIs) from both sides of the road) was considered around the unified segments to extract building footprints and their relative attributes.

As for the prediction class, we labeled each event of the first request as *"1-click"*, *"2-clicks"*, or *"more-clicks"*. The 1-click class means that the instruction was requested only once, while the 2-clicks and more-clicks classes mean that the instruction was requested two or more

[2] Participants indicated their familiarity in three levels (region, route, and landmark) in the online phase.
[3] Route instruction pattern: TURN LEFT [IMPERATIVE] AT CAFE FABRIK [LANDMARK].

times. For example, in Figure 1, E1 and E2 events are marked as instances of class “1-click”, E3 as an instance of class “more-clicks”, and E4 as an instance of class “2-clicks”. In total, we had 234 samples of such first-request events in the unified segments, with an unbalanced distribution across the classes (1-click = 68.803%, 2-clicks = 20.940%, and more-clicks = 10.256%).

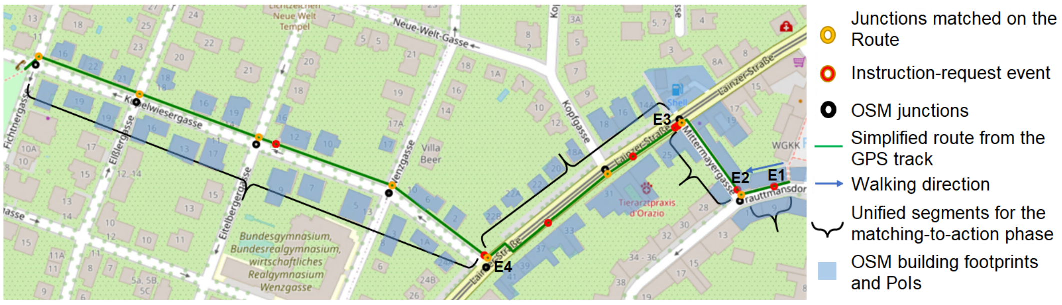

E1
E2
E3
E4
Neue-Welt-Gasse
Kopfgasse
Lainzer Straße
Wenzgasse
Fichtnergasse
Eitelbergergasse
Villa Beer
Bundesgymnasium, Bundesrealgymnasium, wirtschaftliches Realgymnasium Wenzgasse
Tierarztpraxis d'Orazio
WGKK
○ Junctions matched on the Route
○ Instruction-request event
● OSM junctions
— Simplified route from the GPS track
→ Walking direction
} Unified segments for the matching-to-action phase
■ OSM building footprints and PoIs

Figure 1 Preprocessing steps applied to the GPS tracks: smoothing (green line), OSM junction extraction and mapping to the route (black and yellow circles), instruction-event alignment (red circles), segmentation into unified segments with class labels (E1–E4), and extraction of land use and POI information from OSM in a 30m buffer around the route.

3.3 Feature Extraction

Selecting the right features for training a machine learning model is a very important step. Of course, any algorithm trained on any set of features will yield a model, but according to the principle of “garbage-in, garbage-out”, the model trained on inappropriate features, even if it performs well on the training data, fails to generalize and thus explain the results. Following the model of wayfinding decision situations [15], we selected features that reflect the complexity of the environment and instructions, as well as the characteristics of the user.

We extracted four groups of features from our data sources: 41 *Environmental*, 16 *Instruction-*, 12 *User-related*, and six *Gaze* features, yielding a total of 75 features. Table 1 summarizes these features into subcategories. For the environment category, we extracted seven features for the segments, including the length of the route and road segments, the elapsed time since the start of the trial, the distance to the previous and next non-turn intersections, and the distance to the previous and next turn intersections. The elapsed time, while not a direct environmental feature, may serve as a proxy for length/speed and may also be related to working memory (see Section 6). Based on the landcover codes of Urban Atlas data[4], we extracted 13 features describing land use in the buffer around the unified segments. These features are in fact the proportion of buffer area for each land use. For example, we calculated how much of the area is occupied by *“Green-urban-areas”* or *“Sports-and-leisure-facilities”*. Using the same approach, we extracted from the OSM data the semantic label of PoIs[5] and their density along the unified segments. For example, by counting the number of PoIs with the amenity tag *“shop”* per unit area along the unified segments, we calculated the density of the PoI *“shop”*. In total, there were 10 amenities in our experimental regions and we calculated the density for each. We also calculated the total PoI density to have a measure of the visual complexity of the environment. This yields a total of 21 PoI-related features for the environment category.

[4] `https://land.copernicus.eu/user-corner/technical-library/urban-atlas-mapping-guide/view`

[5] `https://wiki.openstreetmap.org/wiki/Key:amenity`

Table 1 Extracted features for predicting instruction demand: POIs (environmental and instruction-related) are extracted from OSM with standard amenity types (e.g. shop, touristic, etc.).

Environmental Features	**41 (7+13+21)**	**Instruction Features**	**16 (2+14)**
unified-segment	*7*	*length-related*	*2*
distance from/to previous/next turn junctions	2	number of words	1
distance from/to previous/next non-turn junctions	2	number of characters	1
segment-length	1	*content-related*	*14*
route-length	1	OSM PoI	11
time passed since start	1	landmark OSM type	1
landuse	*13*	contains-street-names (boolean)	1
PoI	*21*	last instruction (boolean)	1
User Features	**12 (3+5+4)**	**Gaze Features**	**6**
demographics	*3*	fixation count	1
gender (binary)	1	min/max/sd fixation	3
age (in years)	1	mean fixation duration	1
familiarity (binary)	1	fixation duration skewness	1
Big Five Personality traits	*5*		
Spatial Strategies Questionnaire FRS	*4*		

For instructions, since we assumed that the length of the instruction may affect the demand, we calculated the number of words and characters in the instructions. We also extracted 11 OSM-PoI features as one-hot encoding (the 11 PoIs used as landmarks in the instructions), and three features describing whether the instruction contains a street name, whether it is the last instruction, or what kind of spatial object it refers to (e.g., point or an area). In terms of user-related characteristics, we had age, gender, and a binary measure of familiarity as demographic data from the online study; five values for personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) obtained from the Big Five personality test; and four values for spatial strategy scales (preference for egocentric, allocentric, cardinal directions, and the sense of direction score).

Finally, for the gaze-based features, as one goal of our analysis was also to determine how quickly we could predict instruction demand after the first request, we segmented the gaze data into 1- to 10-second windows immediately after the first event, and extracted features within these windows. In this way, we were able to find the minimal set of gaze data for the prediction task. Eye-tracking datasets were collected in a mobile eye-tracking scenario with free head/body movements. It is well known that head movements strongly influence the calculation of saccade length and velocity [2], so no saccadic features were computed. We, therefore, computed only basic fixation-based features. Fixations were extracted using the Dispersion-Threshold Identification (I-DT) algorithm [36] (gaze-dispersion threshold: 1 *deg*; duration threshold: 100 *ms*). Fixation count and five statistical measures from fixation duration (mean, minimum, maximum, standard deviation, and skewness of the frequency distribution), were extracted from the data.

By extracting all these features, we obtained a dataset of size $234 * 75$. These 75 features were extracted based on our assumptions about influential factors on instruction demand. After training the models and analyzing the feature importance (see Section 5), we were able to prune this list and extract the most relevant features and interpret their effect.

4 Machine Learning Experiments

To predict instruction demand, i.e., how often a wayfinder requests an instruction, we trained three classifiers: SVM, RandomForests, and XGBoost. The choice of models was based on previous experience [4] and other successful reports of the performance of these models in the literature of both human activity recognition and wayfinding (see Section 2.3). For instance, our experience with the SVM-RBF, CART, Random Forest, and XGBoost algorithms for a similar task of predicting pedestrians' turning activity based on their gaze behavior (whether they turn left/right at an intersection or continue straight ahead) yielded test accuracies of .58 ±.05, .61 ±.06, .77 ±.06, and .91 ±.08, respectively. Before applying any of these models, we first normalized the data and converted all categorical values to numerical values. To deal with the imbalance of the data, we also compared two methods, sample weighting, and oversampling. It is known that if the number of samples in the minority class is too small and the samples are much farther away from the other classes, the highly weighted samples, although drawing the edge of the decision function to themselves, may not be effective enough to actually lie within the decision function [19]. However, we tried both methods and obtained better results ($\approx 5.6\%$) with Synthetic Minority Over-sampling Technique (SMOTE) [9], which may also be an observation of the same phenomenon in sample weighting. To find out how fast we can predict the demand, in a pilot testing, we trained an XGBoost classifier for each of the 1- to 10- second windows gaze data and plotted the performance metrics to select the optimal window size (see Section 5 for details).

Once we had the data ready, we set up the experimental pipeline in a way to avoid both data leaks and unwanted effects of participants' individuality in training and testing. We split the data into 70% and 30% (train, test) using the leave-one-group-out (LOGO) method to ensure that test participant samples were not part of the training. Then, oversampling was applied as a pipeline separately on the test data and within the 10-fold cross-validation to ensure that there was no data leakage not only between the training and test datasets but also within the validation folds. The training results were checked for overfitting by *mlogloss* and *merror* with cross-entropy loss function. Finally in order to interpret the results and prune the features, and see how much the model's predictions are influenced by every feature, we applied both the Tree SHapley Additive exPlanations (SHAP) method and feature importance by permutation (i.e., leaving the features one by one out ordered by their importance and monitor the drop in accuracy). The interpretation of the results using features' importance is presented and discussed in Section 6.

5 Results

This section summarizes the results of our analysis. First, we report on the SVM and RandomForest classifiers, which achieved test accuracy of 62.88% and 69.78%, respectively, when trained on all the data, being the highest accuracy in both cases. However, the XGBoost classifier outperformed both by an average of 11.2% across all feature combinations. Therefore, only the XGBoost results are presented here in more detail.

We begin with the gaze window sizes. Figure 2 plots the train and test accuracies for 1- to 10-second windows, with the 8- and 2-second windows representing the best and second-best prediction results for instruction demand. The fact that we can predict demand 6 seconds faster while losing less than 1% in accuracy makes window size two more attractive for the application domain. However, since the 8-second window still has the highest accuracy, we continue to report results for this window size.

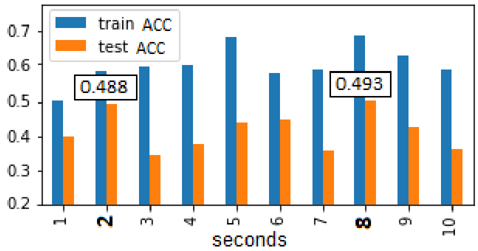


Figure 2 Compares the XGBoost results trained on the gaze data of different window sizes.

As explained in Subsection 3.3, we extracted different categories of features to determine which categories are most important for our prediction task. We conducted 15 experiments in which the pipeline was set to different categories of features or combinations thereof. Table 2 summarizes these results in terms of accuracy, f1 score, precision, recall, and Cohen's kappa. Exp. 1, 2, 3, and 4 are based on single categories with 6, 16, 12, and 41, gaze, instruction, user, and environmental features, respectively. We also run some experiments with the subcategories of these features, which are summarized in Table 1. These experiments are labeled **.1* through **.3*. For example, in Exp. 3.2, the model is trained with the five features of the Big Five personality test to see how personality alone can define instruction demand. In Exp. 6, the model was trained with 16 features that were among the top 4 features from Exp. 1, 2, 3, and 4. Exp. 7 is the result of applying the feature selection by permutation approach. In each iteration of the permutation step, a subset of the features was selected by removing each individual feature based on its importance rank and the model was trained on that subset. The best feature set to report was defined as the smallest subset with the least deviation from the best resulting model, in this case the model trained on all 75 features. As can be seen in Table 2, the three best performances of the model belong to experiments using all features (Exp. 5), permutation-selected features (Exp. 7), and a combination of the four most important features of individual categories (Exp. 6), with a test accuracy of 79.1%, 78.9%, and 77.4%, respectively. However, the kappa values show a better agreement result for the permutation-selected features. Using only gaze features, the model achieves the lowest performance of 49.3%. After that, the instruction-related features provide an accuracy of 58%. User-related features, despite having a smaller number of features than instruction-related features, provide the model with an accuracy of 65.6%. The environment-related features with the highest number of features among the individual categories are close to the best-performing model with less than 4% difference in test accuracy (i.e., 75.8%).

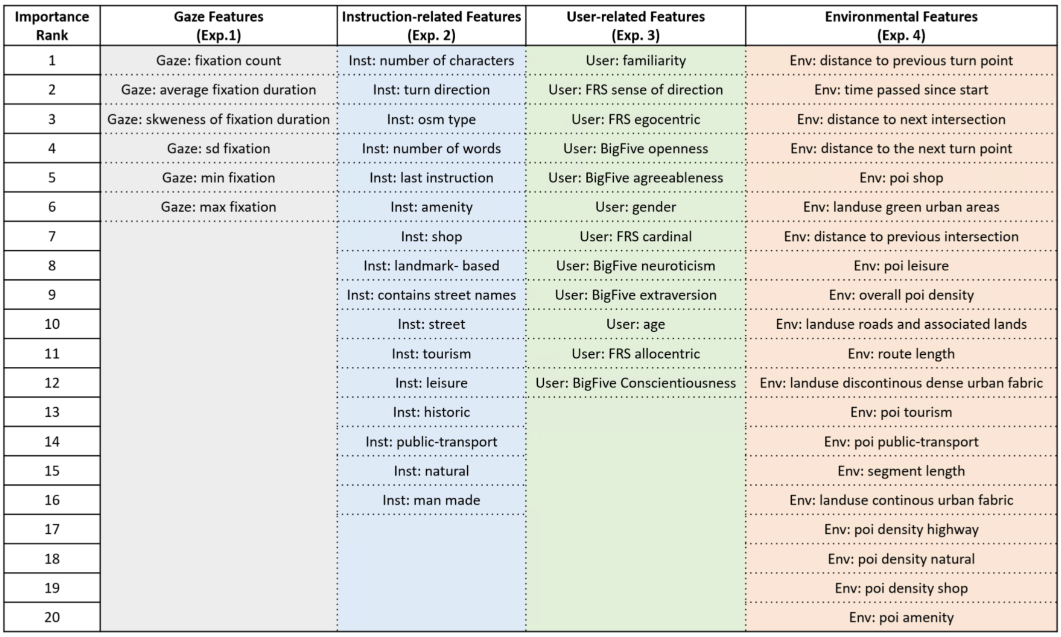
Figures 3 and 4 show the SHAP ranks of each feature category (i.e., Exp. 1, 2, 3, and 4) and the top three best-performing experiments (i.e., Exp. 5, 6, and 7), respectively. In Figure 3, the top four features from each category are used to train the model in Exp. 6. These features are distance to and from turn points and the nearest non-turn intersection, time elapsed since the start of the experiment, familiarity, sense of direction, openness, egocentric preference, length of the instruction, content of the instruction in terms of type of landmarks used and turn direction (left or right), number of fixations, standard deviation of fixations, average, and skewness of fixation duration. In Figure 4, environmental characteristics are among the top four important features in all cases (Exp. 5, 6, and 7). Among user-related features, which are the second most important category, gender, the personality trait of openness, and sense of direction are the most important ones. Instruction length, measured by the number of characters, and instruction content, measured by the type of landmark and turn direction, are as well present in the list of the most important features. Fixation features are at the bottom of all three lists, with average duration and maximum fixation being the most important ones. These results are further discussed in Section 6.

Table 2 Summarizes the results of the trained XGBOOST classifier for different combinations of features in terms of accuracy, f1 score, precision, recall, and Cohen's kappa. Experiment 7 with 21 features yields the best performance after experiment 5 with 75 features.

Exp.	**Features**	#	**Split (LOGO)**	**Evaluation Metrics**				
				Accuracy	**F1 Score**	**Precision**	**Recall**	**Kappa**
1	**gaze**	6	train	0.683	0.683	0.682	0.682	–
			test	0.493	0.493	0.495	0.443	0.432
2	**instruction**	16	train	0.689	0.690	0.692	0.689	–
			test	0.580	0.523	0.560	0.567	0.526
2.1	instruction length	2	train	0.513	0.506	0.511	0.512	–
			test	0.44	0.394	0.393	0.469	0.413
2.2	instruction content	14	train	0.586	0.565	0.627	0.608	–
			test	0.565	0.568	0.581	0.565	0.505
3	**user**	12	train	0.731	0.729	0.729	0.730	–
			test	0.656	0.632	0.678	0.641	0.632
3.1	user demographics	3	train	0.537	0.527	0.529	0.536	–
			test	0.527	0.485	0.492	0.584	0.415
3.2	user BigFive	5	train	0.682	0.683	0.689	0.683	–
			test	0.530	0.505	0.528	0.529	0.531
3.3	user FRS	4	train	0.634	0.634	0.634	0.634	–
			test	0.569	0.521	0.614	0.548	0.489
4	**environment**	41	train	0.855	0.854	0.855	0.854	–
			test	0.758	0.711	0.744	0.757	0.752
4.1	environment segment	7	train	0.841	0.836	0.849	0.840	–
			test	0.696	0.642	0.655	0.670	0.644
4.2	environment PoI	21	train	0.675	0.677	0.681	0.675	–
			test	0.656	0.638	0.702	0.682	0.638
4.3	environment landuse	13	train	0.668	0.671	0.680	0.668	–
			test	0.521	0.495	0.528	0.532	0.567
5	**all**	75	train	0.896	0.897	0.902	0.896	–
			test	**0.791**	0.746	0.758	0.757	**0.742**
6	**manual selection of features based on their importance rank in experiments 1, 2, 3, and 4**	16	train	0.879	0.879	0.882	0.879	–
			test	**0.774**	0.764	0.784	0.792	**0.763**
7	**selection by permutation**	21	train	0.896	0.896	0.897	0.896	–
			test	**0.789**	0.778	0.794	0.802	**0.783**

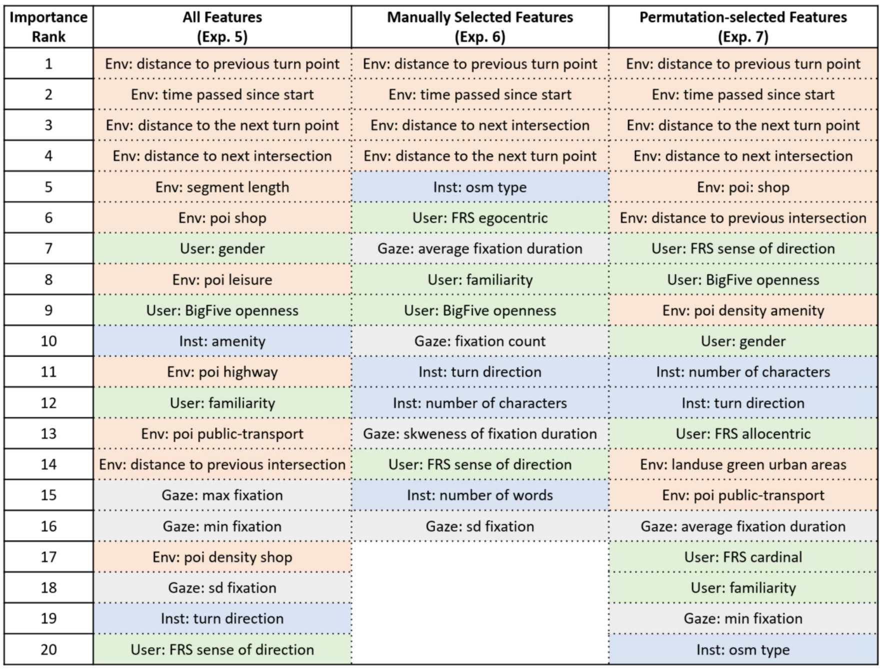
6 Discussion

Here we predicted the instruction demand in a pedestrian wayfinding scenario, after the first instruction was given, based on a set of feature categories. In a real-world application scenario, we assume that three of these categories, namely environmental features, user-related features, and instruction-related features, are fixed once the route to be navigated is selected. That is, once the navigation system computes the optimal route and generates the instructions for it, the environment- and instruction-related features can be easily computed. Similarly, user-related features can be collected in the sign-up information. In contrast, gaze behavior is not static and is heavily influenced by the task, stimulus processing, movements, and so on [11]. Gaze features should therefore be computed immediately after the first instruction. To determine how much gaze data should be recorded for this prediction, we tested different window sizes and found that as little as two seconds of fixation behavior recording can support the prediction with slightly lower accuracy than 8 seconds (**Exp. 1**). It is well known from the gaze analysis literature that fixation behavior can be interpreted in terms of frequency and duration as a sign of cognitive load on information processing, attention, and scene perception (see, e.g., [38, 21]). Our results are consistent with these findings, but also show that in a real-world situation, these features do not by themselves encode enough



Importance Rank	Gaze Features (Exp.1)	Instruction-related Features (Exp. 2)	User-related Features (Exp. 3)	Environmental Features (Exp. 4)
1	Gaze: fixation count	Inst: number of characters	User: familiarity	Env: distance to previous turn point
2	Gaze: average fixation duration	Inst: turn direction	User: FRS sense of direction	Env: time passed since start
3	Gaze: skweness of fixation duration	Inst: osm type	User: FRS egocentric	Env: distance to next intersection
4	Gaze: sd fixation	Inst: number of words	User: BigFive openness	Env: distance to the next turn point
5	Gaze: min fixation	Inst: last instruction	User: BigFive agreeableness	Env: poi shop
6	Gaze: max fixation	Inst: amenity	User: gender	Env: landuse green urban areas
7		Inst: shop	User: FRS cardinal	Env: distance to previous intersection
8		Inst: landmark- based	User: BigFive neuroticism	Env: poi leisure
9		Inst: contains street names	User: BigFive extraversion	Env: overall poi density
10		Inst: street	User: age	Env: landuse roads and associated lands
11		Inst: tourism	User: FRS allocentric	Env: route length
12		Inst: leisure	User: BigFive Conscientiousness	Env: landuse discontinous dense urban fabric
13		Inst: historic		Env: poi tourism
14		Inst: public-transport		Env: poi public-transport
15		Inst: natural		Env: segment length
16		Inst: man made		Env: landuse continous urban fabric
17				Env: poi density highway
18				Env: poi density natural
19				Env: poi density shop
20				Env: poi amenity

Figure 3 The SHAP feature importance rank for Experiments 1 to 4. Top 4 features of each category is manually selected for training the model in experiment 6.



Importance Rank	All Features (Exp. 5)	Manually Selected Features (Exp. 6)	Permutation-selected Features (Exp. 7)
1	Env: distance to previous turn point	Env: distance to previous turn point	Env: distance to previous turn point
2	Env: time passed since start	Env: time passed since start	Env: time passed since start
3	Env: distance to the next turn point	Env: distance to next intersection	Env: distance to the next turn point
4	Env: distance to next intersection	Env: distance to the next turn point	Env: distance to next intersection
5	Env: segment length	Inst: osm type	Env: poi: shop
6	Env: poi shop	User: FRS egocentric	Env: distance to previous intersection
7	User: gender	Gaze: average fixation duration	User: FRS sense of direction
8	Env: poi leisure	User: familiarity	User: BigFive openness
9	User: BigFive openness	User: BigFive openness	Env: poi density amenity
10	Inst: amenity	Gaze: fixation count	User: gender
11	Env: poi highway	Inst: turn direction	Inst: number of characters
12	User: familiarity	Inst: number of characters	Inst: turn direction
13	Env: poi public-transport	Gaze: skweness of fixation duration	User: FRS allocentric
14	Env: distance to previous intersection	User: FRS sense of direction	Env: landuse green urban areas
15	Gaze: max fixation	Inst: number of words	Env: poi public-transport
16	Gaze: min fixation	Gaze: sd fixation	Gaze: average fixation duration
17	Env: poi density shop		User: FRS cardinal
18	Gaze: sd fixation		User: familiarity
19	Inst: turn direction		Gaze: min fixation
20	User: FRS sense of direction		Inst: osm type

Figure 4 The SHAP feature importance rank of the three best-performing feature sets: All features, manually selected features, and permutation selected features. In all groups, environmental features are among the most important features.

information to predict instructional demand.

According to Table 2, instruction-related features (**Exp. 2**) are more informative for the model than gaze alone. Among them, the length of the instruction, encoded with the number of words as a language-independent proxy and the number of characters as a German-language proxy, is more important. This is consistent with the well-known word length effect, which states that longer words (which are common in German) are less well remembered [7]. We assume that both length measures correlate with the memory and recall performance of the instruction. This assumption is inline with previous findings suggesting that longer and more complex navigation instructions may lead to increased cognitive load and decreased wayfinding performance, making the instructions less memorable [23, 26]. The observation that landmark type and turn direction (two important content-related features) are also among the top five features, also supports this assumption. It means that *what information* and *how much information* to be processed are important for predicting instruction needs. This pattern is also found in the three best-performing experiments (Figure 4).

Exp. 3, was based on user-related characteristics only. Familiarity played the most important role here, followed by the sense of direction and egocentric preference scores. These observations are consistent with previous research on the importance of familiarity in activity recognition in wayfinding [39, 3] and the fact that instructions were given in an egocentric viewpoint. The same pattern for spatial strategy scores was also observed in **Exp. 3.3**. Among the personality traits, openness, which correlates with eagerness to learn and experience new things, is also the most important factor in **Exp. 3.2**. The relationship between openness and need for instruction is not well studied, but some studies have reported that individuals with higher openness tend to prefer more creative and less structured tasks [37] and therefore may need less detailed instructions. Our results suggest the same: Wayfinders with higher openness scores tend to show a lower need for instruction. However, further research is needed to decipher the relationship between this personality trait and prior experience (i.e., familiarity), task description and complexity, etc.

The result of **Exp. 4** and its sub-experiments with environmental features shows that segment-related features, including distance measurements to and from the immediate non-turn intersections and the previous and next turn points (which are the target of the instructions), as well as the time passed since onset, are the most important. This means that the longer the distance to these decision points the more probable it is to need instructions again. Analysis of this feature between the two familiarity conditions shows that distance to and from turn points is equally relevant for both conditions, but the distance to and from non-turn points is less dominant for familiar cases. These observations are consistent with those of [17], in which the authors showed that unfamiliar participants ask for instructions earlier on longer segments, and because the upcoming decision point is likely to be seen later on long segments, unfamiliar wayfinders might experience higher levels of uncertainty due to their less developed mental representation of the area. This explanation is also valid for our observation. All these measures that somehow capture the distance to/from different points on the route (time since start or length of segments) may also be related to the capacity of the wayfinder's working memory, which means that the longer the distance, the more likely it is that someone will forget the instruction and need it again. However, in our study, we did not control for this characteristic, and according to the psychological literature, working memory as a cognitive process of temporarily storing and processing information in the mind to perform tasks, while playing an important role in following instructions, is not the only factor that provides an advantage in a complex task environment [42]. Other factors that play a role in a complex task environment, such as problem-solving ability, creativity, or other cognitive skills, may provide an advantage beyond good working memory [13]. However, we cannot relate our results to these explanations because our data do not provide corresponding information on these merits. Secondly important are PoI-based features, including the overall density of PoIs and *shop* in particular. Overall PoI density can be considered as a good indicator of the density of the urban landscape, which affects human-environment interactions including information seeking for wayfinding [31]. Our results suggest that the higher the PoI density, the higher the cognitive effort required to match the received instruction to the environment, and the greater the probability of requesting the same instruction again.

Finally, for the land use characteristics (e.g. *green urban area*, *roads*, etc.), which are equally important as PoIs in **Exp. 4** but not in **Exp. 5**, **6**, and **7**, no precise explanation can be offered because land use is likely biased by the study design. The familiar trials were conducted closer to the center of the city, while the unfamiliar trials were largely in the outskirts with different land use. This may justify the absence of land use characteristics in the experiments with combined features, as familiarity already encodes this difference. The effect of this environmental aspect needs to be further explored.

None of the experiments with a single category achieved practically high accuracy, with the exception of the environment category. Our combinatorial experiments aimed to find a pruned list of features that are most informative for predicting instruction demand. **Exp. 5**, **6**, and **7** are the best-performing experiments, and we consider the last experiment to be the optimal one with one-third of the features and only less than 1% loss in accuracy. In the last three experiments, we can see some similarities in the features. However, even in terms of features, we consider the permutatively selected features (listed in Figure 4 Exp. 7) to be the most informative features for predicting instruction needs. This list of features encodes well the wayfinding situation in terms of the instruction needs related to the wayfinder's relative position to the next and previous decision points (both turn and non-turn points); the effect of cognitive load on fixation behavior caused by two factors: the density of the urban landscape and PoIs in the environment and the processing of the instructional information (length and content); and finally, the characteristics of the user, from personality to preference for spatial strategies, gender, and familiarity.

7 Conclusion and Future Work

This paper presents the results of 15 ML experiments that predict the need for navigation instructions with an accuracy of 78.4%. The predictions are based on a combination of factors such as the wayfinder's position, the next decision points, cognitive load, the amount of visual information, the length and content of the instructions, and the user's personality traits and spatial strategies. The findings have theoretical and practical implications for better understanding the cognitive aspects of wayfinding and for adapting navigation instructions in real time.

The features used in the experiments encode several aspects of the wayfinder's situation: The wayfinder's position with respect to the starting point, the previous and upcoming decision points (both turn and non-turn), the cognitive load due to information processing and environmental perception reflected in fixation behavior and caused by the density of the landscape and the amount of visual information, the length and content of the given instruction, and finally the user's characteristics in terms of personality traits and spatial strategies. In our experimental design, instruction demand was defined as the frequency with which the same instruction is requested after it has been heard once. In other navigation systems with different HCI components, this demand can be defined, for example, as the frequency of transition between the map screen and the environment, or as the number of fixations on the augmented information in AR-based systems after a first viewing. In any case, this prediction offers advantages from both theoretical and application perspectives: Behavior before and after an instruction is retrieved contains valuable information about the cognitive aspects of human-environment interaction and spatial perception. Knowing what external and internal features affect this demand can help us better understand the spatial-cognitive aspects of wayfinding and even mental states of uncertainty, being lost, or needing reassurance that are common in wayfinding but not yet well explored. A further research question would be when and in which stage of wayfinding we feel such needs more strongly and for which purpose (e.g. self-localization or route planning), the repeated instructions may be useful.

Since our prediction results are based on unseen data, it is very likely that a pre-trained model can perform on-the-fly predictions as a module of the navigation system, which can be beneficial for real-time instruction adaptation. Off-line predictions can also be used as a measurable metric for evaluating navigation instructions. However, further research is needed to examine the generalizability of our observations for different modalities.

References

1 Y. Abdelrahman, A. A. Khan, J. Newn, E. Velloso, Sh. Ashraf Safwat, J. Bailey, A. Bulling, F. Vetere, and A. Schmidt. Classifying attention types with thermal imaging and eye tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), 2019. `doi:10.1145/3351227`.

2 N. Alinaghi and I. Giannopoulos. Consider the head movements! saccade computation in mobile eye-tracking. In *2022 Symposium on Eye Tracking Research and Applications*, 2022.

3 N. Alinaghi, M. Kattenbeck, and I. Giannopoulos. I can tell by your eyes! continuous gaze-based turn-activity prediction reveals spatial familiarity. In *15th Intl. Conf. on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

4 N. Alinaghi, M. Kattenbeck, A. Golab, and I. Giannopoulos. Will you take this turn? gaze-based turning activity recognition during navigation. In *11th Intl. Conf. on Geographic Information Science (GIScience 2021)-Part II*. Leibniz-Zentrum für Informatik, 2021.

5 L. S. Ambati and O. El-Gayar. Human activity recognition: a comparison of machine learning approaches. *J. of the Midwest Association for Information Systems (JMWAIS)*, 2021(1):4, 2021.

6 T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, En. Kasneci, and P. Gerjets. Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 Intl. Conf. on Multimodal Interaction*, ICMI '19, pages 154–163, New York, NY, USA, 2019. Association for Computing Machinery.

7 A. D Baddeley, N. Thomson, and M. Buchanan. Word length and the structure of short-term memory. *J. of verbal learning and verbal behavior*, 14(6):575–589, 1975.

8 A. Brügger, K. Richter, and S. Fabrikant. How does navigation system behavior influence human behavior? *Cognitive research: principles and implications*, 4:1–22, 2019.

9 N. V Chawla, K. W Bowyer, L. O Hall, and W Ph. Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. of artificial intelligence research*, 16:321–357, 2002.

10 L. De Cock, N. Van de Weghe, K. Ooms, I. Saenen, N. Van Kets, G. Van Wallendael, P. Lambert, and P. De Maeyer. Linking the cognitive load induced by route instruction types and building configuration during indoor route guidance, a usability study in vr. *Intl. J. of Geographical Information Science*, 36(10):1978–2008, 2022.

11 W. Dong, H. Liao, B. Liu, Z. Zhan, H. Liu, L. Meng, and Y. Liu. Comparing pedestrians' gaze behavior in desktop and in real environments. *Cartography and Geographic Information Science*, 47(5):432–451, 2020. `doi:10.1080/15230406.2020.1762513`.

12 M. Duckham, S. Winter, and M. Robinson. Including landmarks in routing instructions. *J. of location based services*, 4(1):28–52, 2010.

13 S. Dunham, E. Lee, and A. M Persky. The psychology of following instructions and its implications. *American J. of Pharmaceutical Education*, 84(8), 2020.

14 P. Fogliaroni, D. Bucher, N. Jankovic, and I. Giannopoulos. Intersections of our world. In *10th Intl. Conf. on geographic information science*, volume 114, page 3. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

15 I. Giannopoulos, P. Kiefer, M. Raubal, K. Richter, and T. Thrash. Wayfinding Decision Situations: A Conceptual Model and Evaluation. In *Proc of GIScience 2014*, 2014.

16 F. Goebel, K. Kurzhals, V. R. Schinazi, P. Kiefer, and M. Raubal. Gaze-adaptive lenses for feature-rich information spaces. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3379155.3391323.

17 A. Golab, M. Kattenbeck, G. Sarlas, and I. Giannopoulos. It's also about timing! when do pedestrians want to receive navigation instructions. *Spatial Cognition & Computation*, 22(1-2):74–106, 2022.

18 G. Gunzelmann, J. R Anderson, and S. Douglass. Orientation tasks with multiple views of space: Strategies and performance. *Spatial Cognition and Computation*, 4(3):207–253, 2004. doi:10.1207/s15427633scc0403_2.

19 H. He and E. A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

20 H. Huang, M. Schmidt, and G. Gartner. Spatial knowledge acquisition with mobile maps, augmented reality and voice in the context of gps-based pedestrian navigation: Results from a field test. *Cartography and Geographic Information Science*, 39(2):107–116, 2012.

21 M. Adam Just and Patricia A. C. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

22 M. Keskin and P. Kettunen. Potential of eye-tracking for interactive geovisual exploration aided by machine learning. *Intl. J. of Cartography*, pages 1–23, 2023. doi:10.1080/23729333.2022.2150379.

23 A. Klippel, H. Tappe, and Ch. Habel. Pictorial representations of routes: Chunking route segments during comprehension. In *Spatial Cognition III: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning 8*. Springer, 2003.

24 A. Klippel, H. Tappe, L. Kulik, and P. U Lee. Wayfinding choremes—a language for modeling conceptual route knowledge. *J. of Visual Languages & Computing*, 16(4):311–329, 2005.

25 A. Klippel and S. Winter. Structural salience of landmarks for route directions. In *Spatial Information Theory: Intl. Conf., COSIT 2005, Ellicottville, NY, USA, September 14-18, 2005. Proceedings 7*, pages 347–362. Springer, 2005.

26 J. Krukar, V. Joy Anacta, and A. Schwering. The effect of orientation instructions on the recall and reuse of route and survey elements in wayfinding descriptions. *J. of Environmental Psychology*, 68:101407, 2020.

27 A. Lakehal, S. Lepreux, L. Letalle, and Ch. Kolski. From wayfinding model to future context-based adaptation of hci in urban mobility for pedestrians with active navigation needs. *Intl. J. of Human–Computer Interaction*, 37(4):378–389, 2021.

28 H. Liao, W. Dong, H. Huang, G. Gartner, and H. Liu. Inferring user tasks in pedestrian navigation from eye movement data in real-world environments. *Intl. J. of Geographical Information Science*, 33(4):739–763, 2019.

29 B. Ludwig, G. Donabauer, D. Ramsauer, and K. al Subari. Urwalking: Indoor navigation for research and daily use. *KI - Künstliche Intelligenz*, 2023.

30 S. Münzer and Ch. Hölscher. Entwicklung und validierung eines fragebogens zu räumlichen strategien. *Diagnostica*, 2011.

31 C. Nothegger, S. Winter, and M. Raubal. Selection of salient features for route directions. *Spatial cognition and computation*, 4(2):113–136, 2004.

32 L. Pillette, G. Moreau, J. Normand, M. Perrier, A. Lecuyer, and M. Cogne. A systematic review of navigation assistance systems for people with dementia. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

33 B. Rammstedt, Ch. Kemper, M. Céline Klein, C. Beierlein, and A. Kovaleva. Eine kurze skala zur messung der fünf dimensionen der persönlichkeit: big-five-inventory-10 (bfi-10). *Methoden, Daten, Analysen (mda)*, 7(2):233–249, 2013.

34 M. Raubal and S. Winter. Enriching wayfinding instructions with local landmarks. In *Intl. Conf. on geographic information science*, pages 243–259. Springer, 2002.

35 K. Richter, M. Tomko, and S. Winter. A dialog-driven process of generating route directions. *Computers, Environment and Urban Systems*, 32(3):233–245, 2008.

36 D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78, New York, NY, USA, 2000. ACM. doi:10.1145/355017.355028.

37 D. Twomey, E. Burns, and Sh. Morris. Personality, creativity, and aesthetic preference: Comparing psychoticism, sensation seeking, schizotypy, and openness to experience. *Empirical Studies of the Arts*, 16(2):153–178, 1998.

38 P Unema. Differences in eye movements and mental work-load between experienced and inexperienced motor vehicle drivers. *Visual search*, pages 193–202, 1990.

39 J. M Wiener, S. J Büchner, and C. Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9(2):152–165, 2009.

40 J. M Wiener, Ch. Hölscher, S. Büchner, and L. Konieczny. Gaze behaviour during space perception and spatial decision making. *Psychological Research*, 76(6):713–729, 2012. doi:10.1007/s00426-011-0397-5.

41 S. Winter, M. Tomko, B. Elias, and M. Sester. Landmark hierarchies in context. *Environment and Planning B: Planning and Design*, 35(3):381–398, 2008.

42 T. Yang. *The role of working memory in following instructions*. PhD thesis, University of York, 2011.

43 W. Zhang, X. Zhao, and Z. Li. A comprehensive study of smartphone-based indoor activity recognition via xgboost. *IEEE Access*, 7:80027–80042, 2019.

44 B. Zhu, J. G Cruz-Garza, Q. Yang, M. Shoaran, and S. Kalantari. Identifying uncertainty states during wayfinding in indoor environments: An eeg classification study. *Advanced Engineering Informatics*, 54:101718, 2022.