Title
Testing Landmark Salience Prediction in Indoor Environments Based on Visual Information

Authors
Donabauer, Gregor
Ludwig, Bernd

Publication Date
2021-09-01

DOI
10.25436/E2SG62

Peer reviewed

Testing Landmark Salience Prediction in Indoor Environments Based on Visual Information

Gregor Donabauer ✉
Chair of Information Science, University of Regensburg, Germany

Bernd Ludwig ✉
Chair of Information Science, University of Regensburg, Germany

Abstract

We identify automated landmark salience assessment in indoor environments as a problem related to pedestrian navigation systems that has not yet received much attention but is nevertheless of practical relevance. We therefore evaluate an approach based on visual information using images to capture the landmarks' outward appearance. In this context we introduce the largest landmark image and salience value data set in the domain so far. We train various classifiers on domain agnostic visual features to predict the salience of landmarks. As a result, we are able to clarify the role of visual object features regarding perception of landmarks. Our results demonstrate that visual information has only limited expressiveness with respect to salience.

Supplementary Material https://github.com/doGregor/landmark-salience-prediction

1 Introduction

Pedestrians are often facing problems of self-orientation and wayfinding in environments they are not familiar with [7]. This challenge causes problems with route planning and decision making during navigation [17]. To support pedestrians in such situations they are increasingly provided digital assistance, e.g. Google Maps[1] on their smartphones [23].

Several studies highlight the need of landmarks for an adequate description of routes and to improve human orientation, e.g. [24, 27]. In general landmarks are conspicuous objects in space. Depending on varying semantic, structural and visual characteristics they can be perceived as differently salient reference points [25]. In the context of pedestrian navigation systems the question on how to identify suitable objects arises. Frequently, controlled field studies are conducted to let participants name relevant landmarks, e.g. [15]. This approach is not applicable in large-scale, unsupervised manner. Furthermore, it can be biased and does not allow to consistently identify appropriate objects [14].

From these observations we note a need for automated landmark identification and rating techniques. Research in this area so far has recommended to use crowd sourcing via OpenStreetMap[2] [22]. [2] suggest predicting landmark salience based on image data that moreover can be used to provide visual cues to pedestrians. Yet it is unclear which proportion of semantic, structural, and visual information is necessary to confidently deduce the salience score of an object. In prior research, the role of solely visual data so far has been rarely analyzed [21] leaving a research gap in operationalizing the approach in [2].

[1] https://www.google.com/maps/
[2] https://www.openstreetmap.org/

In this paper, we address this gap for image data of indoor environments. To the best of our knowledge, only [13, 9, 30] provide approaches to automated indoor landmark salience prediction. This contrasts to the urgent need for appropriate reference objects in indoor navigation instructions that is caused by a higher complexity of the environment [11]. Since navigation and related perception of objects for orientation are tasks non-trivial to model it will be interesting to see how much of salience visual information can encode.

To investigate this question, we introduce the largest dataset in this domain so far and try to draw conclusions using methods of machine learning. To foster reproducibility in the geographical information science [14] we provide our dataset and implementation via GitHub.

2 Related Work

Automated identification of landmarks for pedestrian navigation systems has attracted little attention in previous research, particularly in context of indoor environments. Yet some approaches have been proposed and we are briefly discussing them below.

We start with a look at techniques for outdoor areas: Some methods for example rely on external information sources like cartographical material or content in geographical databases [8, 6]. The data can be used to extract object related features that allow to deduce salience scores and thus suitable landmarks. Another study suggests data mining methods applied to online texts with spatial context [26]. This content, mainly originating from geographical information systems, can help identifying relevant objects. All three approaches rely on large-scale external information sources which usually are not publicly available for indoor environments. Additionally, they only consider data of structural and semantic nature.

[18] try to identify conspicuous buildings serving as landmarks based on the visual appearance of their facades. [19] also take into account the visual characteristics of building facades for salience determination in context of navigation through a virtual downtown environment. Both studies combine visual and semantic information within their approach making it difficult to assess the role of visual features. Either are reporting correlations between facades' colors and salience values. In [28] saliency maps are computed using DeepGaze to identify landmarks in images of virtual scenes. The authors report highly salient regions not to correlate with objects that attracted visual attention of test persons.

For indoor environments, [13] propose data mining methods specifically for the interior of buildings, but do not compare their findings with a ground truth of human salience ratings. Lastly, [9] consider 200 indoor-scene images that are used to let participants rank potential landmarks. Visual and semantic information are considered to train a genetic programming algorithm on the collected data to predict the objects' salience values. The authors were able to correctly identify the most salient landmark in 76% of scenes. In contrast to these studies that purely or partly rely on semantic and structural data, we focus on the influence of visual information on indoor-landmark salience. Most similar to our work is the study by [30]. The authors also consider visual salience, but apply features that we did not consider.

3 Data and Methodology

3.1 Dataset

The landmark dataset is adopted from previous work [1]: 74 participants conducted a navigational experiment through an indoor environment on a route consisting of multiple segments. At each segment the subjects had to name four objects they would use to describe the current route section. In a follow-up questionnaire the salience values of the identified landmarks are measured according to [10]. We take three isolated images (masked scene information) from different angles of all identified objects in [1]. By that we can ensure capturing the landmarks' visual characteristics from multiple perspectives.

The final dataset consists of 1266 images $\mathbf{X}, \forall X \in \mathbf{X}\colon X \in \mathbb{R}^{298\times 224}$ related to 422 distinct landmark objects and their salience values $Y, \forall y \in Y : y \in \mathbb{R}$. The data are split in 0.8:0.2 ratio. We use 5-fold cross-validation and a shared evaluation/test set due to the size of our dataset. To test whether objects can confidently be grouped in high- and low-quality landmarks we also provide binary labels. The threshold for the split is calculated using k-means clustering and expectation maximization density estimation.

3.2 Landmark Salience Prediction

We subdivide our analysis into three steps. **(1)** we try to predict landmark salience directly on the image data using transfer learning based on a convolutional neural network and evaluate the results utilizing methods of explainable artificial intelligence **(2)** afterwards, we extract a set of different image features and evaluate them regarding significant differences compared to a random baseline **(3)** promising features are finally combined and treated as landmark representations that are used to train multilayer perceptrons.

Random baseline: since the distribution of salience values is similar to a Gaussian, we sample values according to mean and standard deviation of the train set in quantity of values in the respective test set. For binary classification we sample random values of 0 and 1 with 0.5 probability since both classes are approximately equally distributed.

(1) Transfer learning and XAI: We choose the well-known VGG19 CNN architecture as a frozen convolutional base, pretrained on the ImageNet dataset and stack further dense layers for salience value (linear, regression head) or binary label (sigmoid, classification head) prediction on top of it. For parametrization we refer to our GitHub resources. For evaluation, we use mean absolute error (MAE), mean squared error (MSE) and mean percentual error (MAPE). For simplicity we will only report the latter metric further on. Classification is evaluated using the accuracy metric.

We visualize pivotal pixels in the input data of the most and least precise predictions using *deep taylor decomposition* and *layer-wise relevance propagation.* The insights can foster interpretability of crucial image content, for example to select better features for step (2).

(2) Feature Evaluation: All features are evaluated regarding a confident prediction of salience values and landmark group labels, respectively. We adopt them from previous work in domain of semantic information mapping, like image classification and image retrieval. We fit random forest estimators on the data and compare the results against the random baseline using paired t-test. Only features that allow significant improvements in landmark salience prediction ($p < 0.01$ w.r.t. MAPE and Accuracy) are considered for the final MLP training. We are briefly introducing all utilized attributes below.

(a) *High level style:* We process the feature map output of the *conv5_v1* hidden layer of the pretrained VGG19 CNN to obtain high level landmark style representations. Correlations are calculated via gram matrices G^l through $G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}$ where G^l_{ij} is the dot product of the vectorized feature maps i and j in layer l. Subsequently, we reduce their dimensionality by applying a principal components analysis. [4, 3]

(b) *High level content:* To obtain high level content representations we adapt the same processing as with feature (a). Other than previously, we use VGG's *conv4_v2* layer to extract feature maps. [4, 3]

(c) *Complexity:* The application of XAI methods seems to reveal that complexity might be a useful landmark characteristic. We calculate spatial information for a grid of 18 sub-fields of each object-image by $SI_r = \sqrt{s_h^2 + s_v^2}$ per sub-field, where s_h and s_v are vertically and horizontally filtered Sobel-images. [29]

(d) *Colors, contrast and brightness:* Those features are popular, easy to compute image characteristics and yield basic information on the landmarks' visual appearance. Furthermore, they have been proposed as correlating with facade salience in [4, 13, 18, 19].

(e) *Scale-invariant feature transform (SIFT):* SIFT is popular in context of image retrieval, e.g. [16], and allows to obtain robust content representations independent of perspective. We extract the $|N| = 10$ most prominent features, where $\forall n \in N: n \in \mathbb{R}^{128}$ and reduce their dimension with PCA.

(f) *PCA, ICA, NMF, Dictionary Learning:* These methods are frequently used to extract meaningful features for face detection, e.g. [20, 5]. They represent details similar to information in primary visual cortex. We use $n = 50$ components.

(3) MLP Classifier: It turned out that only features (a), (b) and (c) provide significant improvements over our random baseline. We therefore concatenate those features and create landmark representation vectors from it. Those are used to train multilayer perceptrons for prediction of salience values (linear, regression head) and binary labels (sigmoid, classification head). For architecture and parametrization we refer to our GitHub repository.

4 Results

We start introducing the random baseline scores which amount to 25.29 (MAPE) and 0.507 (Accuracy). Metrics are reported as average of all five cross validation sets.

Through transfer learning we are able to improve MAPE to 18.09 and accuracy to 0.629. In table 1 below, we are reporting MAPE and accuracy for the individual features based on random forest estimators. All characteristics significantly better than the random baseline at $p < 0.01$ regarding both metrics are combined for the final landmark representations.

(a)*	(b)*	(c)*	(d.1)	(d.2)	(d.3)	(e)	(f.1)	(f.2)	(f.3)	(f.4)
22.50	22.26	22.18	22.87	23.00	22.06	23.35	22.30	22.29	23.08	22.99
0.611	0.624	0.594	0.564	0.561	0.550	0.543	0.572	0.551	0.576	0.589

Table 1 Features, MAPE and Accuracy for landmark salience prediction; *significant at $p < 0.01$

We obtain the most meaningful visual characteristics of the landmarks using (a), (b) and (c) as features. Training MLPs on that data yields a MAPE of 17.80 and an accuracy score of 0.608. Overall, the results demonstrate that visual features of indoor-objects have limited expressiveness when speaking about perception of potential landmarks.

5 Discussion and Conclusion

Considering the remaining prediction error, we evaluated our data regarding correlations between true and estimated landmark perception as well as structural and semantic information. The results show negative correlations (Pearson's: -0.467, p-value: $2.97e^{-24}$) between MAPE and salience rating. A qualitative analysis reveals that especially low rated objects are highly influenced by non-visual characteristics. While we are able to make confident predictions for well rated objects, arguably also due to high visual expressiveness, we need further knowledge on the structural layout between landmarks to more confidently identify unsuitable reference points. As [12] state: "For instance, a red facade in an area where all facades are red will not stand out. But the same facade in a grey neighborhood stands out". Additionally, other factors influence the perception of landmarks, for example the position of an object in context of route directions. This validates findings of [28] who assume that image data represent too little of context to allow ideally identifying suitable objects.

Identifying salient objects to support human orientation in unfamiliar environments is not trivial to model. As we used domain agnostic features only, our results should generalize to other indoor environments. Unfortunately, for selecting appropriate landmarks it is not sufficient to extract these features from images: we could estimate the salience of landmarks in 62.9% of cases. For the remaining 37.1%, we conclude that the visual context of landmarks as well as additional semantic and structural knowledge is necessary to further improve prediction accuracy. This result is in line with the observation in [30]: visual information helps wayfinders if they are not familiar with the environment while structural and semantic information renders landmarks salient for wayfinders with good knowledge of the environment.

References

1 Christina Bauer. Unterstützung der orientierung im innenbereich: Analyse landmarkenbasierter karten-interfaces anhand des blickverhaltens der nutzer, September 2018. URL: `https://epub.uni-regensburg.de/37666/`.

2 David Caduff and Sabine Timpf. On the assessment of landmark salience for human navigation. *Cognitive processing*, 9:249–67, 12 2007.

3 Wei-Ta Chu and Yi-Ling Wu. Deep correlation features for image style classification. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 402–406, New York, NY, USA, 2016.

4 Vicente Dominguez, Pablo Messina, Denis Parra, Domingo Mery, Christoph Trattner, and Alvaro Soto. Comparing neural and attractiveness-based visual features for artwork recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, DLRS 2017, page 55–59, New York, NY, USA, 2017.

5 Bruce A. Draper, Kyungim Baek, Marian Stewart Bartlett, and J.Ross Beveridge. Recognizing faces with pca and ica. *Computer Vision and Image Understanding*, 91(1):115–137, 2003. Special Issue on Face Recognition.

6 Birgit Elias. Extracting landmarks with data mining methods. In Walter Kuhn, Michael F. Worboys, and Sabine Timpf, editors, *Spatial Information Theory. Foundations of Geographic Information Science*, pages 375–389, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

7 Christian Gaisbauer and Andrew U. Frank. Wayfinding model for pedestrian navigation. In *AGILE 2008*, pages 1–9. Springer, 2008.

8 Jana Götze and Johan Boye. Learning landmark salience models from users' route instructions. *Journal of Location Based Services*, 10(1):47–63, 2016.

9 Xuke Hu, Lei Ding, Jianga Shang, Hongchao Fan, Tessio Novack, Alexey Noskov, and Alexander Zipf. Data-driven approach to learning salience models of indoor landmarks by using genetic programming. *International Journal of Digital Earth*, 13(11):1230–1257, 2020.

10 Markus Kattenbeck. Empirically measuring salience of objects for use in pedestrian navigation, Juli 2016. URL: `https://epub.uni-regensburg.de/34145/`.

11 Markus Kattenbeck, Manuel Ullmann, Christina Bauer, and Bernd Ludwig. Der weg ist das ziel - fußgängernavigation ist forschung zu information behavior. *Information - Wissenschaft & Praxis*, 66:45–55, 02 2015.

12 Alexander Klippel and Stephan Winter. Structural salience of landmarks for route directions. COSIT'05, page 347–362, Berlin, Heidelberg, 2005. Springer-Verlag.

13 Hao Lyu, Zhonghai Yu, and Liqiu Meng. *A Computational Method for Indoor Landmark Extraction*, pages 45–59. Springer International Publishing, Cham, 2015.

14 Bartosz Mazurkiewicz, Markus Kattenbeck, Peter Kiefer, and Ioannis Giannopoulos. Not arbitrary, systematic! average-based route selection for navigation experiments. In Krzysztof Janowicz and Judith A. Verstegen, editors, *GIScience 2021 - Part I*, volume 177 of *LIPIcs*, pages 8:1–8:16, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

15 Alexandra Millonig and Katja Schechtner. Developing landmark-based pedestrian-navigation systems. *IEEE Transactions on Intelligent Transportation Systems*, 8(1):43–49, 2007.

16 Gholam Ali Montazer and Davar Giveki. Content based image retrieval system using clustered scale invariant feature transforms. *Optik*, 126(18):1695 – 1699, 2015.

17 Daniel R. Montello. *Navigation*, page 257–294. Cambridge Handbooks in Psychology. Cambridge University Press, 2005.

18 Clemens Nothegger, Stephan Winter, and Martin Raubal. Selection of salient features for route directions. *Spatial Cognition & Computation*, 4(2):113–136, 2004.

19 Denise Peters, Yunhui Wu, and Stephan Winter. Testing landmark identification theories in virtual environments. In Christoph Hölscher, Thomas F. Shipley, Marta Olivetti Belardinelli, John A. Bateman, and Nora S. Newcombe, editors, *Spatial Cognition VII*, pages 54–69, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

20 Menaka Rajapakse and L. Wyse. Nmf vs ica for face recognition. In *ISPA 2003*, volume 2, pages 605–610 Vol.2, 2003.

21 Martin Raubal and Stephan Winter. Enriching wayfinding instructions with local landmarks. In Max J. Egenhofer and David M. Mark, editors, *Geographic Information Science*, pages 243–259. Springer Berlin Heidelberg, 2002.

22 Kai-Florian Richter and Stephan Winter. Harvesting user-generated content for semantic spatial information: The case of landmarks in openstreetmap. In *Surveying & Spatial Sciences 2011*, pages 75–86, 2011.

23 Gian-Luca Savino, Miriam Sturdee, Simon Rundé, Christine Lohmeier, Brent Hecht, Catia Prandi, Nuno Jardim Nunes, and Johannes Schöning. Maprecorder: analysing real-world usage of mobile map applications. *Behaviour & Information Technology*, 0(0):1–17, 2020.

24 Alexander W. Siegel and Sheldon H. White. The development of spatial representations of large-scale environments. volume 10 of *Advances in Child Development and Behavior*, pages 9–55. JAI, 1975.

25 Molly E. Sorrows and Stephen C. Hirtle. The nature of landmarks for real and electronic spaces. In Christian Freksa and David M. Mark, editors, *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science*, pages 37–50, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

26 Taro Tezuka and Katsumi Tanaka. Landmark extraction: A web mining approach. In Anthony G. Cohn and David M. Mark, editors, *Spatial Information Theory*, pages 379–396, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

27 Steffen Werner, Bernd Krieg-Brückner, Hanspeter A. Mallot, Karin Schweizer, and Christian Freksa. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. In Matthias Jarke, Klaus Pasedach, and Klaus Pohl, editors, *Informatik '97*, pages 41–50, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

28 Demet Yesiltepe, Ayse Ozbil Torun, Antoine Coutrot, Michael Hornberger, Hugo Spiers, and Ruth Conroy Dalton. Computer models of saliency alone fail to predict subjective visual attention to landmarks during observed navigation. *Spatial Cognition & Computation*, 21(1):1–28, 2020.

29 Honghai Yu and Stefan Winkler. Image complexity and spatial information. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 12–17, 2013.

30 Zhiyong Zhou, Robert Weibel, and Haosheng Huang. Familiarity-dependent computational modelling of indoor landmark selection for route communication: a ranking approach. *International Journal of Geographical Information Science*, 0(0):1–33, 2021. doi:10.1080/13658816.2021.1946542.