Using Explainable AI to Investigate Navigational Affordances in Real-world Scenes

UNIVERSITY OF AMSTERDAM

Clemens G. Bartnik and Iris I.A. Groen
Video & Image Sense Lab, Informatics Institute, University of Amsterdam, The Netherlands

Aims

We move with ease through our environment. We can choose different paths and use a variety of different navigational actions, such as walking, swimming or climbing.

Past research[1] focusing on indoor environments[2] suggests that CNNs[3] trained for scene recognition[4] contain useful representations of navigational affordances.

Here, we investigated the ability of CNNs to capture affordances in a broader set of environments and use different explainable AI feature visualizations (LRP[5], LIME[6], Grad-CAM[7]) and different task objectives/network models (depth perception[8], scene segmentation[9]).

Method

XAI feature visualizations:

Layer-wise Relevance Propagation (LRP)[5] identifies important pixels – those that contribute the most to the prediction get a higher relevance score.

Local Interpretable Model-Agnostic Explanations (LIME)[6] creates local explanations by perturbing the input and tracking how the predictions change.

Grad-CAM[7] visualizes image regions that were important for the classification decision.

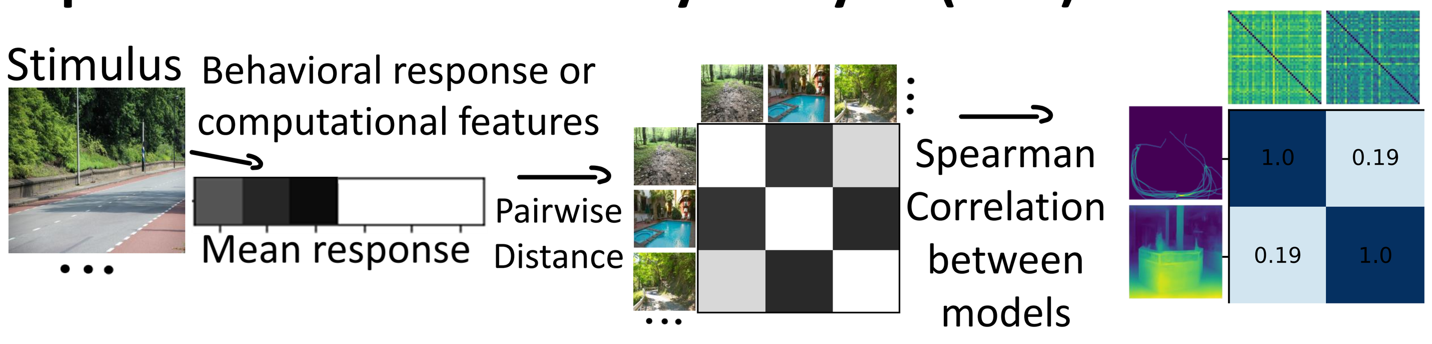
Models:

VGG16[3] trained for scene recognition on Places 365[4]

Scene Segmentation[9] trained on ADE20K – for stimuli[2] we use floor segmentation. For our stimuli we selected the segmentation class with the highest pixel overlap with the average paths.

Depth estimation[8] resulting depth map of monocular depth estimation model

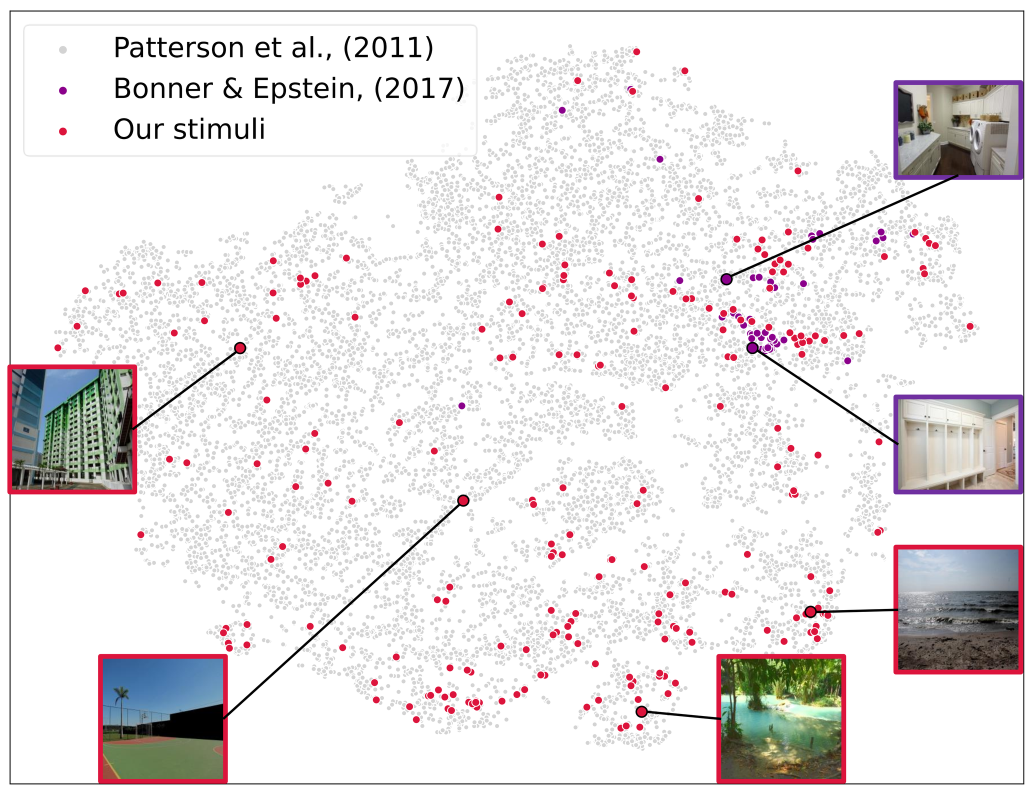
Representational similarity analysis (RSA)[11]



Distance metric: Euclidean distance

Stimuli

50 naturalistic real-world indoor stimuli[1]



t-SNE visualization of our stimuli covering the attribute space of large-scale dataset[10] (12k + Images) compared to previous research[2]

231 naturalistic real-world stimuli (equal No. of images: indoor, outdoor natural and outdoor manmade)

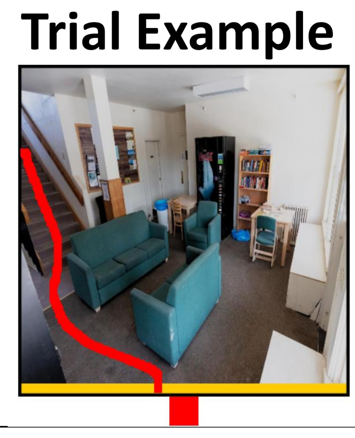
Behavior

In an online experiment we collected human annotations (N = 152) of possible paths through each scene (at least 20 per image) like[2]. Each participant was allowed to draw only one possible path but was not otherwise constrained.

Task Description

"Draw a route along which you would move/navigate in any way (walk, swim, ride, or climb) through each scene."

Trial Example

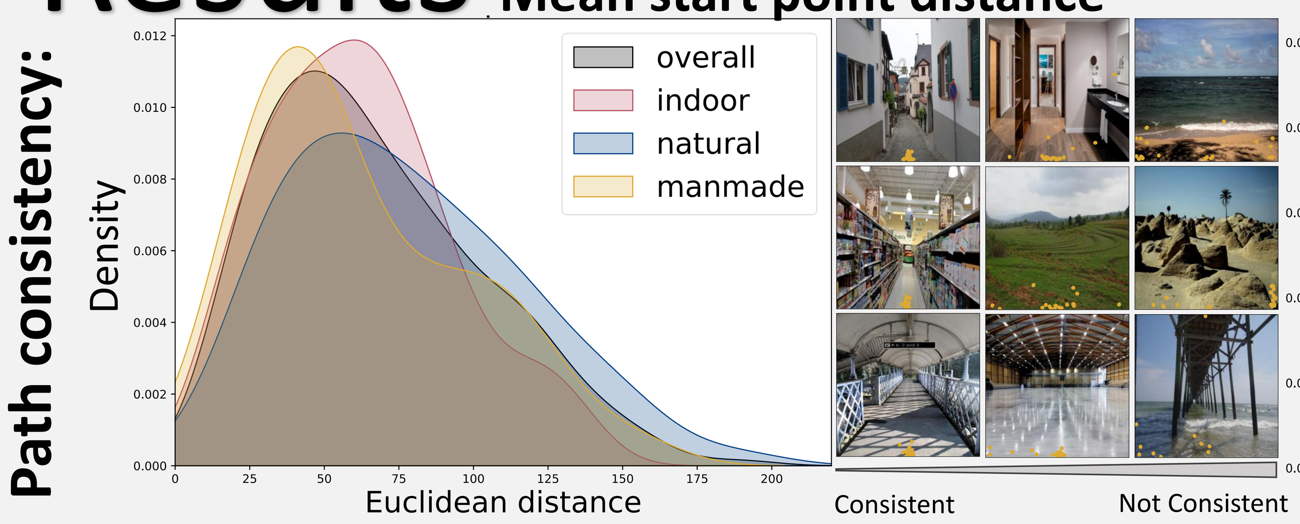


Average path across participants

Results

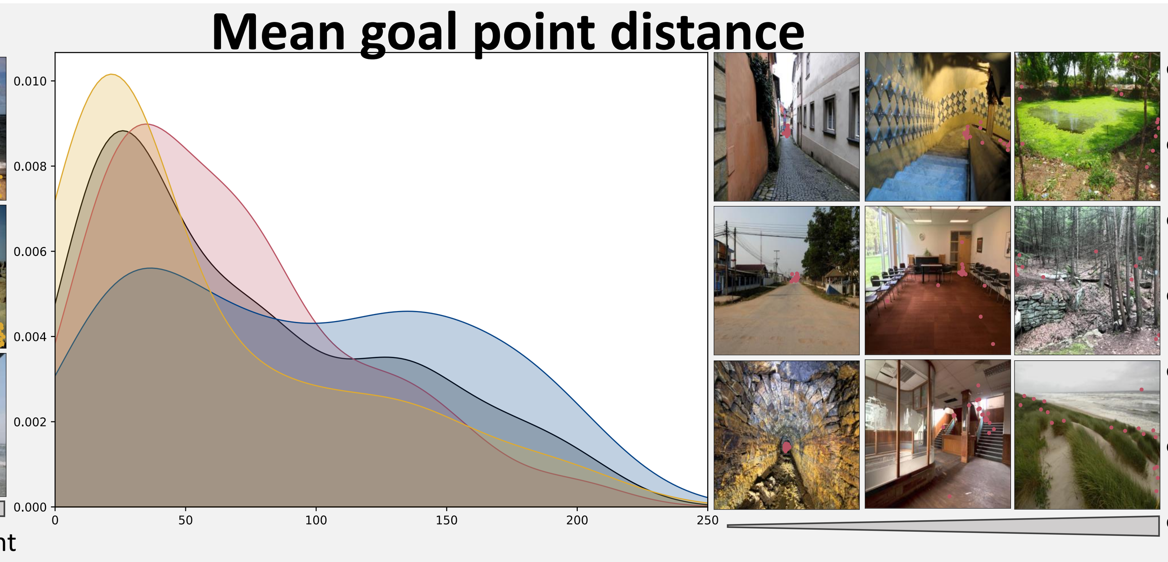
Path consistency:

Mean start point distance



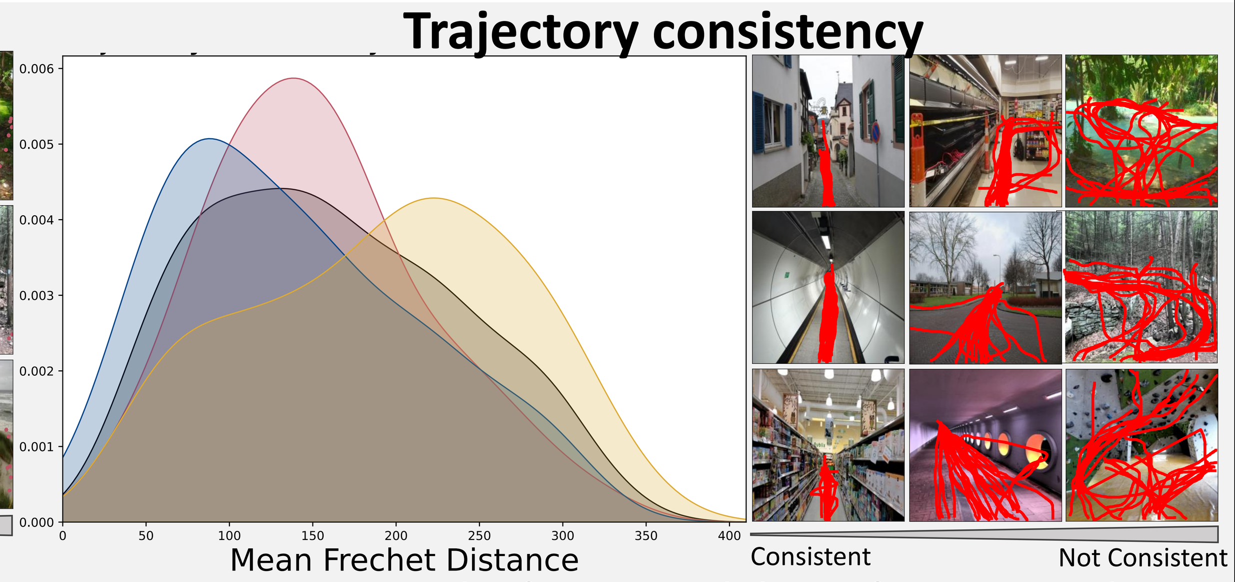
Overall consistent start points

Mean goal point distance



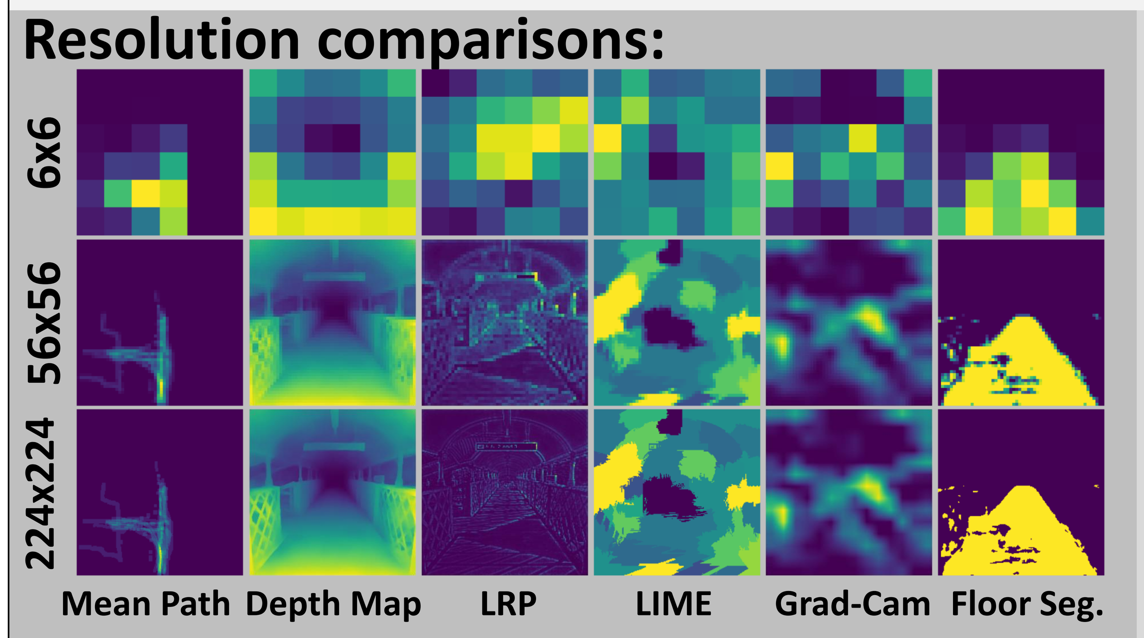
Overall high agreement in goal points – but for some natural scenes goal points less "clear"

Trajectory consistency

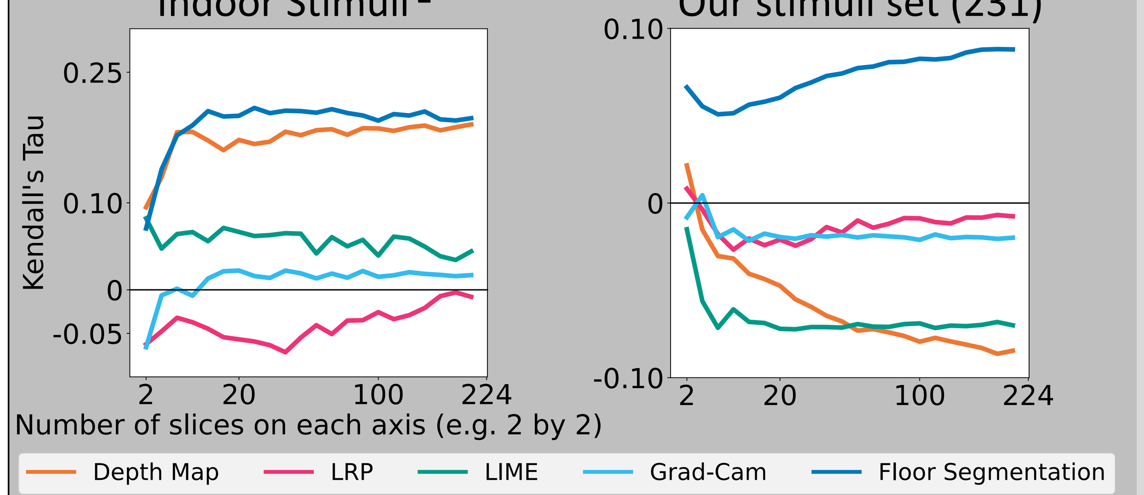


Most consistent paths for natural, least for manmade – but highly dependent on the image

Resolution comparisons:



Correlations of mean path with all others across resolutions



Correlations stabilize around 25x25 slices.

Model comparisons:

Indoor Stimuli[2]

	Mean Path	Depth Map	LRP	LIME	Grad Cam	Floor Seg.
	1.0	0.19	-0.01	0.04	0.02	0.2
	0.19	1.0	-0.06	0.06	0.0	0.2
	-0.01	-0.06	1.0	-0.05	-0.05	0.01
	0.04	0.06	-0.05	1.0	0.07	0.03
	0.02	0.0	-0.05	0.07	1.0	0.07
	0.2	0.2	0.01	0.03	0.07	1.0

Our stimuli set (231)

	Mean Path	Depth Map	LRP	LIME	Grad Cam	Floor Seg.
	1.0	-0.08	-0.01	-0.07	-0.02	0.09
	-0.08	1.0	0.01	-0.03	0.05	0.07
	-0.01	0.01	1.0	-0.0	0.02	0.02
	-0.07	-0.03	-0.0	1.0	-0.03	-0.03
	-0.02	0.05	0.02	-0.03	1.0	0.03
	0.09	0.07	0.02	-0.03	0.03	1.0

In the stimuli[2] we see substantial correlations between certain feature visualizations. These are much lower in our new stimulus set with diverse environments.

Indoor

	Mean Path	Depth Map	LRP	LIME	Grad Cam	Floor Seg.
Mean Path	1.0	-0.0	-0.02	-0.01	0.0	0.18
Depth Map	-0.0	1.0	-0.01	0.01	0.05	0.12
LRP	-0.02	-0.01	1.0	-0.04	-0.04	-0.03
LIME	-0.01	0.01	-0.04	1.0	0.02	0.03
Grad Cam	0.0	0.05	-0.04	0.02	1.0	0.05
Floor Seg.	0.18	0.12	-0.03	0.03	0.05	1.0

Outdoor manmade

	Mean Path	Depth Map	LRP	LIME	Grad Cam	Floor Seg.
Mean Path	1.0	-0.15	-0.09	-0.1	-0.07	0.07
Depth Map	-0.15	1.0	-0.01	-0.03	0.05	0.09
LRP	-0.09	-0.01	1.0	0.09	0.06	0.05
LIME	-0.1	-0.03	0.09	1.0	-0.03	-0.05
Grad Cam	-0.07	0.05	0.06	-0.03	1.0	-0.01
Floor Seg.	0.07	0.09	0.05	-0.05	-0.01	1.0

The correlation pattern changes for different types of environments. Our indoor subset shows similar pattern to stimuli[2] but the other types of environments differ.

Outdoor natural

	Mean Path	Depth Map	LRP	LIME	Grad Cam	Floor Seg.
Mean Path	1.0	-0.04	0.09	-0.09	-0.0	0.04
Depth Map	-0.04	1.0	0.09	-0.04	0.17	0.13
LRP	0.09	0.09	1.0	-0.06	0.05	0.03
LIME	-0.09	-0.04	-0.06	1.0	-0.1	-0.03
Grad Cam	-0.0	0.17	0.05	-0.1	1.0	0.03
Floor Seg.	0.04	0.13	0.03	-0.03	0.03	1.0

Conclusions

- Human participants are consistent in annotating possible paths in diverse natural scenes. In most scenes they independently pick consistent goal points and use consistent paths to navigate the presented scene. This shows that diagnostic features for potential paths are present.
- We can confirm that scene layout[1] captured by depth estimation and floor segmentations captures affordances of indoor environments. But this does not generalize to other types of environments (outdoor manmade, outdoor natural)..
- None of the CNN feature representation we tested, alone adequately captures navigational affordances across all types of environments. It is still unclear how humans incorporate multiple relevant features to perceive a possible path through a natural scenes.

Code

References

[1] Bonner & Epstein (2018), *PLoS Comput. Biol.* | [4] Zhou et al. (2017), *IEEE PAMI* | [7] Selvaraju et al. (2020) *Int. J. Comput. Vis.* | [10] Patterson & Hays (2012), *CVPR* |
[2] Bonner & Epstein (2017), *PNAS* | [5] Bach et al. (2015), *PLoS One* | [8] Miangoleh et al. (2021), *CVPR* | [11] Kriegeskorte et al. (2008), *Front. Syst. Neurosci* |
[3] Kalliatakis (2017), *GitHub* | [6] Ribeiro et al. (2016), *Proc. ACM SIGKDD* | [9] Zhou et al. (2018), *Int. J. Comput. Vis.* |

Contact

c.g.bartnik@uva.nl
github.com/cgbartnik