Interpretable prediction of human fixations from behavior-derived representational dimensions

Luca Kämmer (kaemmer@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences

Alexander Kroner
University of Osnabrück

Martin Hebart
Justus-Liebig University Gießen

Abstract

Eye movements are core to how humans perceive their environment, which is why understanding which properties guide fixations can help us better understand visual perception. Although there are many computational models trying to predict eye movements, they mostly aim to maximize performance without offering insights into why certain image regions draw our gaze. We addressed this question by leveraging 49 representational object dimensions that capture visual and semantic object information to predict human fixations on images. We weighted these dimensions with their relevance for an image to generate behaviorally-relevant feature maps, without specifically training on fixation data. Our approach outperformed a permutation-controlled baseline and matched the performance of a saliency model. Crucially, our predictions are interpretable, offering insights into which representational dimensions drive them. Lastly, we showed how predictive individual dimensions are of fixation in general which helps us better understand which features drive gaze allocation.

Keywords: Gaze prediction; Human fixations; Object representations; Computer vision; Interpretability

Introduction

Our visual world is made up of complex scenes composed of multiple objects. Overt attention is the mechanism by which humans fixate on important aspects of a visual scene to efficiently sample information from their environment (Posner, 1980). Researchers have long tried to predict where people will fixate in an image to infer how we visually sample information (Itti, Koch, & Niebur, 2002). While such models have progressed a lot in recent years, achieving remarkable accuracy (Linardos, Kümmerer, Press, & Bethge, 2021), they offer little insight into why certain regions draw our gaze more than others.

To bridge this gap, we used the core visual and semantic dimensions underlying perceived object similarity identified by Hebart, Zheng, Pereira, and Baker (2020). These dimensions were derived from behavioral responses to a similarity task on images from the THINGS database (Hebart et al., 2019, 2023). The authors trained a sparse non-negative embedding model to predict object similarities, yielding 49 interpretable object dimensions reflecting the information participants used to distinguish between objects. These dimensions have been demonstrated to be predictive of basic object behavior including categorization and typicality (Hebart et al., 2020) as well as memorability (Kramer, Hebart, Baker, & Bainbridge, 2023) and are predictive of responses in early and high-level visual cortex (Contier, Baker, & Hebart, 2024), underscoring their importance for our mental and cortical representation of objects. In this project, we use these dimensions to generate interpretable predictions of human fixations that help us shed light on the core objects dimensions influencing gaze behavior.

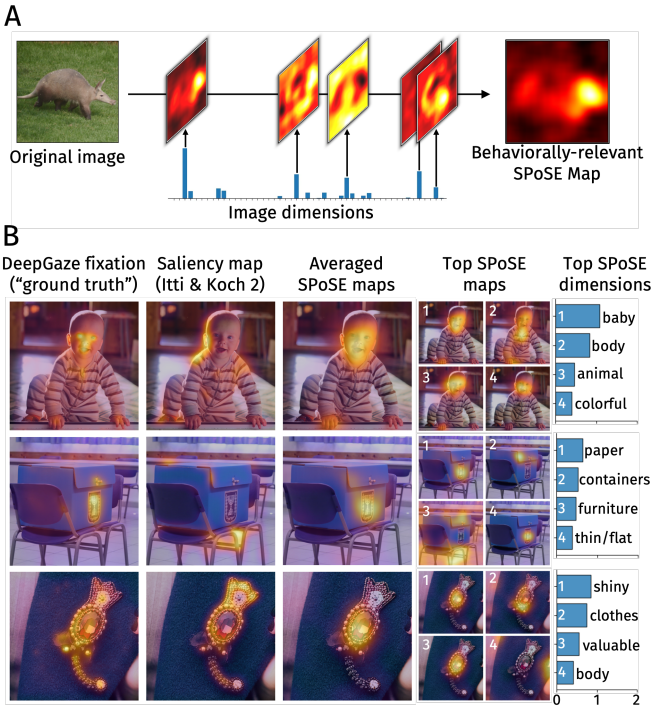

Figure 1: **A** Shows how SPoSE feature maps are extracted and combined into behaviorally-relevant SPoSE maps used for fixation prediction. **B** Examples of saliency models created by the different models, including a depiction of the most relevant SPoSE dimensions for each image.

Methods

While the method by which the SPoSE dimensions are learned is not inherently image-computable, Kaniuth, Mahner, Perkuhn, and Hebart (2024) created a model (Dimpred) that is able to predict SPoSE dimension scores for any natural image. The model extracts DNN activations from CLIP-ResNet in response to an image and then linearly maps them to the 49 interpretable SPoSE dimensions. The model is trained on a set of images for which the SPoSE embedding was computed based on similarity-judgements but it generalizes well to unseen images. The authors also utilized the interpretability method RISE (Petsiuk, Das, & Saenko, 2018), which can generate dimension-specific saliency maps for each image. This method repeatedly occludes parts of the input image using randomly generated masks to observe the effect on the predicted dimensions. Thus, it can compute a dimension-specific relevance score for each image region to generate feature maps that indicate where a certain dimension is represented on the input image.

These feature maps allow us to look at the distribution of an object dimension across an image, which is why they offer an ideal way to assess how these key visual and semantic dimensions are mirrored in where humans will fixate. Specifically, we hypothesized that the same features that are relevant for making similarity judgments of images might also be reflected in where on an image we fixate. To generate such an integrated

map, we weighted the feature map for each dimension with the associated SPoSE dimension score generated by Dimpred to create a ”behaviorally-relevant” feature map (Kaniuth et al., 2024), as depicted in Figure 1A.

Results

Predicting fixations

We first tested this approach on images from the THINGS+ dataset (Stoinski, Perkuhn, & Hebart, 2022) which contains 1854 images depicting all the broadly sampled image concepts of the THINGS dataset (Hebart et al., 2019), making it a suitable candidate for addressing how object dimensions influence fixation patterns. As a proxy for empirical fixations, we generated fixation maps using the DeepGaze IIE model (Linardos et al., 2021), the state of the art model to simulate fixation patterns. To assess how well our SPoSE maps predict the DeepGaze fixation maps, we used a range of commonly used metrics for measuring the similarity between saliency heatmaps and fixation maps (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016; Judd, Durand, & Torralba, 2012). To control for biases (e.g. centerbias) of the dataset, that are independent of the image, we performed a permutation control where for each fixation map a fixation map from another random image is used as a predictor (Koehler, Guo, Zhang, & Eckstein, 2014). In addition, we assessed the revised Itti & Koch saliency model implemented by Harel, Koch, and Perona (2006). In general, our model performed much better than the permutation control and was comparable with the saliency model 1. To assess how well our model generalizes to other images, we replicated these results with the MIT1003 dataset, containing images with fixation maps of 1,003 natureal scenes (Judd, Ehinger, Durand, & Torralba, 2009) (Table 2).

Table 1: Performance on THINGS+ Dataset

	AUC-Judd ↑	SIM ↑	AUC-Borji ↑	AUC-shuf ↑	CC ↑	NSS ↑
SPoSE	0.81	0.81	0.76	0.66	0.56	1.45
Itti & Koch 2	0.81	0.74	0.79	0.68	0.59	1.44
Perm.-control	0.68	0.78	0.65	0.50	0.37	0.64

Table 2: Performance on MIT1003 Dataset

	AUC-Judd ↑	SIM ↑	AUC-Borji ↑	AUC-shuf ↑	CC ↑	NSS ↑
SPoSE	0.78	0.50	0.77	0.61	0.53	1.22
Itti & Koch 2	0.77	0.49	0.76	0.62	0.43	1.11
Perm.-control	0.68	0.34	0.68	0.50	0.20	0.49

Importantly, our model also shows which SPoSE dimensions were involved in generating each prediction and their individual contribution to the behaviorally-relevant SPoSE map (Figure 1B).

Contribution of individual dimensions

Lastly, to show how predictive each dimension is of fixations, we assessed their individual performance in both datasets Figure 2). We decided to use the shuffled AUC measure, as it removes the center bias contribution (Kummerer, Wallis, & Bethge, 2018). Both datasets showed a similar pattern of predictive dimensions, indicating that some dimensions (e.g. animal, fire, baby) draw our gaze more than others (e.g. flat pattern, furniture, long). Some dimensions even seem to dissuade people from fixating (below-chance accuracy).

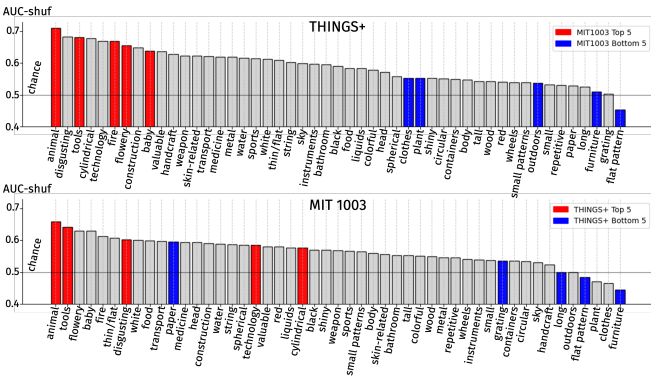

Figure 2: Relative predictive performance of individual dimensions for THINGS+ and MIT 1003

Conclusion

Overall, these results show that the SPoSE dimensions can provide critical information about the nature of the representations that may underlie eye movements on object images and natural scenes. These results offer exciting new opportunities because they do not only predict fixation patterns, but also allow for a more detailed examination of which object dimensions drive these predictions. In future projects, we aim to explore how our prediction may be improved by weighing different SPoSE dimensions. We also plan to validate our approach on additional datasets to uncover the central, generalizable dimensions directing our gaze.

References

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, *8*(11), 2179–2193.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, *19*.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., ... Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Itti, L., Koch, C., & Niebur, E. (2002). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. In *Mit technical report.*

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 ieee 12th international conference on computer vision* (pp. 2106–2113).

Kaniuth, P., Mahner, F. P., Perkuhn, J., & Hebart, M. N. (2024). A high-throughput approach for the efficient prediction of perceived similarity of natural objects. *bioRxiv*, 2024–06.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of vision*, *14*(3), 14–14.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science advances*, *9*(17), eadd2981.

Kummerer, M., Wallis, T. S., & Bethge, M. (2018). Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the european conference on computer vision (eccv)* (pp. 770–787).

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 12919–12928).

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2022). Things+: New norms and metadata for the things database of 1,854 object concepts and 26,107 natural object images. *Preprint at https://doi. org/10.31234/osf. io/exu9f*.