Iterative Bayesian inference explains the dynamics of perceptual organization of natural scenes

T. Biswas[1], **J. Vacher**[2], **P. Mamassian**[3], **S. Molholm**[1], **R. Coen-Cagli**[1]
1. Albert Einstein College of Medicine, 2. Université Paris Cité, 3. École Normale Supérieure Paris

Abstract

Perceptual decision making is classically conceptualized as evidence integration theory – the notion that sensory inputs are perceived by sequentially accumulating noisy samples from the environment and averaging out the noise. Modeling with evidence integration has captured perceptual and neural dynamics elicited by parametric stimuli in simple tasks, but studies of natural vision reveal richer dynamics that remain poorly understood.

In this study, we propose samples in time are not accumulated from a noisy external environment, but from internal representations formed through Bayesian inference where the statistics of sensory inputs are refined iteratively. Thus, we aim to test if iterative Bayesian inference determines perceptual dynamics when processing natural stimuli.

To test this, we focus on natural image segmentation. We measured human perceptual segmentation using a recently published experimental design: participants judged whether pairs of regions in an image were in the same segment ('yes') or not ('no'). Subjective segmentation maps were reconstructed for each participant with optimization on 'yes'/'no' responses per pair.

Examining responses where perceived segments were inconsistent with the segments established by the optimal subjective map, we observed that participants presented a bias toward responding 'yes' when the two regions were close and 'no' when far. Furthermore, decision times increased with distance for 'yes' responses, but decreased with distance for 'no' responses, and this effect was larger for participants with stronger bias.

For further inquiry, we developed image-computable segmentation models of the classical evidence integration and iterative Bayesian inference theories. Although both model types fit aggregate decision-time distributions similarly well, we found that the spatiotemporal dynamics observed in the data were captured only by iterative inference incorporating a Bayesian spatial proximity prior. This work highlights the importance of considering iterative Bayesian computations to understand human perceptual dynamics when exact inference is intractable, as in most real-life situations.

Keywords: normative modeling; visual perception; decision-making; perceptual dynamics

Experiments

We follow the design of the segmentation task in (Vacher et al., 2023) for quantifying aspects of natural image segmentation in human participants (Fig. 1). Our findings for "yes" response dynamics are consistent with past work (Jolicoeur et al., 1986; Korjoukov et al., 2012; Roelfsema, 2023), and current models (Adeli et al., 2023; Goetschalckx et al., 2023; Veerabadran et al., 2023), but the aforementioned studies do not present a framework for understanding the dynamics of "no" responses.

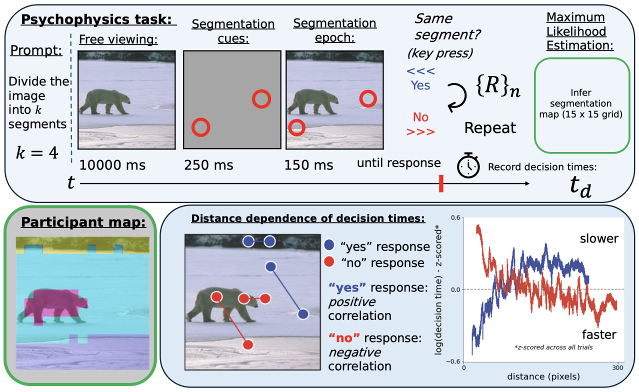
Psychophysics task:
Prompt:
Divide the image into k segments
$k = 4$
Free viewing:
Segmentation cues:
Segmentation epoch:
Same segment? (key press)
<<< Yes
No >>>
$\{R\}_n$
Repeat
Maximum Likelihood Estimation:
Infer segmentation map (15 x 15 grid)
10000 ms
250 ms
150 ms
until response
Record decision times:
t
t_d
Participant map:
Distance dependence of decision times:
"yes" response
"no" response
"yes" response: positive correlation
"no" response: negative correlation
log(decision time) - z-scored*
0.6
0.0
−0.6
0
300
distance (pixels)
slower
faster
*z-scored across all trials

Figure 1: **Psychophysics task:** The participant *mentally* partitions the image into k segments during the free-viewing epoch. Then, segmentation cues are briefly shown superimposed on a gray screen (red circles), followed by image presentation. After image presentation the participants respond "yes" or "no" to whether cues are in the same segment. We collected a set of responses $\{R\}_n$ to n distinct cues, and the corresponding decision times $\{t_d\}_n$. **Participant map:** A segmentation map is computed such that $\{R\}_n$ can be recapitulated with maximum likelihood. **Distance dependence of decision times:** Across 11 participants and 12 images ($n = 31,629$ pairs) we find that "yes" and "no" response decision times have positive and negative correlations with distance respectively (also shown in Figure 3c).

Model

We compare and contrast two image-computable, normative models of decision making described in Figure 2. One uses classical evidence integration while the other uses iterative Bayesian inference.

Let $\mathbf{t}_d$ represent participant decision times and let $\hat{\mathbf{t}}_d$ represent the time variable in the model. Model parameters (defined in Fig. 2b,d) are fit across all pairs for each participant and image by minimizing the negative log-likelihood of observing the human distribution $\mathbf{t}_d$ under the model distribution $\hat{\mathbf{t}}_d$ across all pairs.

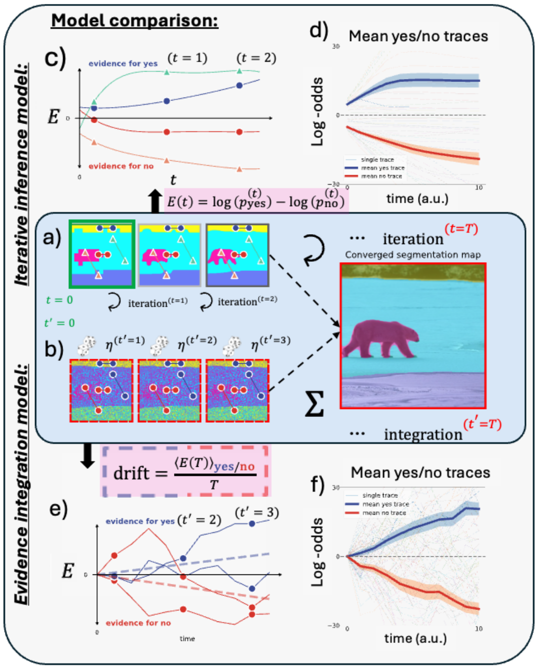

Figure 2: **Center panel (light blue)**: The model assigns a per-segment probability to each image pixel, here we overlay colors indicating the *most likely* segment. Across time, both models converge to the same percept (dashed arrows). **(a)** In iterative inference, the low-resolution participant segmentation map is the initial guess at $t = 0$ (green square). Segment probabilities are refined by iterative Bayesian inference until time T when inference cannot improve (Vacher et al., 2022). **(b)** In evidence integration, the converged map (large red square) is perturbed (dashed red squares) by independent noise at draw-time t', $\eta^{(t')}$. Evidence is integrated (represent by Σ) over T samples. **Outer panels (gray):** Pink boxes indicate equations used in transforming pixel probabilities to an evidence space $E \in (-\infty, \infty)$. **(c)** In iterative inference, $E(t = 0)$ comes from the initial guess and spatial proximity prior which allows for distinct E for close pairs (circles) and far pairs (triangles). E may stop when segment probabilities become static (light blue triangle or red circle). In the cases with highest certainty, evidence grows monotonically (blue circle, light red triangle). **(d)** In iterative inference, mean traces elucidate how a decision time can be calculated by evaluating when the log-odds become static $(dE/dt \to 0)$, or when reach a fittable y-axis boundary, i.e. $E(t) = b$. **(e)** In evidence integration, the model computes a drift rate by averaging over converged evidence for yes/no responses (dashed blue/red lines). The model drifts from $E(t' = 0) = 0$ as independent noise is added at each t'. **(f)** Mean traces elucidate how a decision time can be calculated by evaluating when the log-odds reach a fittable y-axis boundary. Unlike in iterative inference, evidence is accumulating at a constant rate.

Results and Discussion

Our results show that the iterative inference model fits human decision time data at least as well as the classical evidence integration mode (Fig. 3a). However, only iterative inference captures the distance dependence of reaction times (Fig. 3b,c).

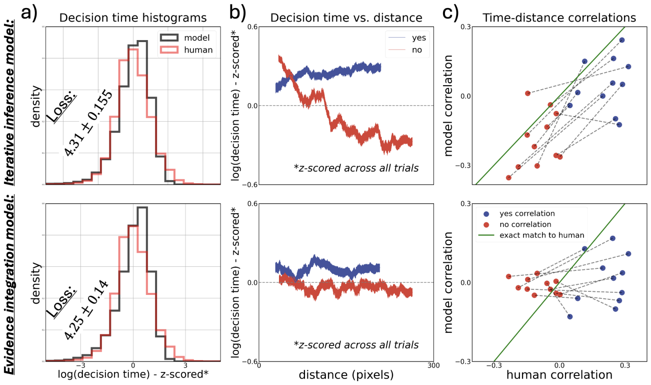

Figure 3: Results from model fits per participant and image. **(a)** Model decision time distributions (black) plotted against human distributions (red) for each model type, annotated with the loss (negative log likelihood) per participant per trial. Mean loss values are not significantly different between models $(p > 0.4)$ **(b)** Iterative inference model decision times conditioned on response type and distance (top middle) are qualitatively more similar to human dynamics than evidence integration model decision times (bottom middle). **(c)** Each marker indicates the Spearman correlation between time and distance for a single participant across all images. In iterative inference, similarity to human dynamics is evident per-participant (as well as per-image, not shown). Correlation *differences* between yes and no responses in iterative inference are particularly aligned with human data as shown by the slopes of dashed lines matching the slope of the green line, regardless of marker location.

In other words, although neither model was trained with knowledge of the distance between pairs, the iterative Bayesian inference model captures a human-like distance dependence because of the spatial proximity prior. Ablation experiments (not shown here due to lack of space) indicate that it is indeed the spatial proximity prior in iterative inference that is responsible for this effect, and not the human-aided initial guess or non-Bayesian pre-processing of image features in the model. In summary, this work provides a normative explanation for the recurrent computations that have been shown to be involved in segmentation.

Acknowledgments

This work was supported by NIH grants EY030578, EY031166 and by a grant from the Simons Foundation Research in Autism Initiative (SFARI).

References

Adeli, H., Ahn, S., Kriegeskorte, N., & Zelinsky, G. (2023). *Affinity-based attention in self-supervised transformers predicts dynamics of object grouping in humans.* Retrieved from https://arxiv.org/abs/2306.00294

Goetschalckx, L., Govindarajan, L. N., Ashok, A. K., Ahuja, A., Sheinberg, D. L., & Serre, T. (2023). *Computing a human-like reaction time metric from stable recurrent vision models.* Retrieved from https://arxiv.org/abs/2306.11582

Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, *14*(2), 129–140. (Place: US Publisher: Psychonomic Society) doi: 10.3758/BF03198373

Korjoukov, I., Jeurissen, D., Kloosterman, N. A., Verhoeven, J. E., Scholte, H. S., & Roelfsema, P. R. (2012, December). The time course of perceptual grouping in natural scenes. *Psychological Science*, *23*(12), 1482–1489. doi: 10.1177/0956797612443832

Roelfsema, P. R. (2023, April). Solving the binding problem: Assemblies form when neurons enhance their firing rate—they don't need to oscillate or synchronize. *Neuron*, *111*(7), 1003–1019. doi: 10.1016/j.neuron.2023.03.016

Vacher, J., Launay, C., & Coen-Cagli, R. (2022, May). Flexibly regularized mixture models and application to image segmentation. *Neural Networks*, *149*, 107–123. doi: 10.1016/j.neunet.2022.02.010

Vacher, J., Launay, C., Mamassian, P., & Coen-Cagli, R. (2023, January). Measuring uncertainty in human visual segmentation.
(arXiv: 2301.07807)

Veerabadran, V., Ravishankar, S., Tang, Y., Raina, R., & Sa, V. R. d. (2023, November). *Adaptive recurrent vision performs zero-shot computation scaling to unseen difficulty levels.* arXiv. Retrieved 2025-04-10, from http://arxiv.org/abs/2311.06964 (arXiv:2311.06964 [cs]) doi: 10.48550/arXiv.2311.06964