Disentangling belief and strategy in natural visual search

Hyunwoo Gu hwgu@stanford.edu
Justin L. Gardner jlg@stanford.edu
Department of Psychology, Wu Tsai Neurosciences Institute
Stanford University, Stanford, CA 94305, United States

Abstract

Our beliefs and strategies are not always aligned. For example, in visual search, an aligned strategy would be to choose to look where one believes the target to be (belief maximization). However, previous studies using simple stimuli have found that human search performance is comparable to an ideal observer which maximizes information (Najemnik & Geisler, 2009), a strategy which can sometimes select gaze to locations where little is known instead of where the target is believed to be. In naturalistic settings, however, visual search poses additional challenges; for example, targets can take on many possible appearances, and object affordances can suggest actions such as reaching, which may influence gaze strategies.

While widely used predictive models of visual saliency (Itti & Koch, 2001; Droste, Jiao, & Noble, 2020; Kümmerer, Bethge, & Wallis, 2022; Ding et al., 2022; Hosseini, Kazerouni, Akhavan, Brudno, & Taati, 2024; Yang et al., 2024) have achieved impressive accuracy in predicting human fixations from image features, they do not model belief or strategy. Moreover, these models are not *normative* as they do not specify the ideal criteria that can be compared to human performance, prohibiting a principled way to assess optimal belief propagation and visual search strategy. To address these challenges and characterize the gaze selection strategy, we generalized an ideal observer model (Najemnik & Geisler, 2005) to natural images with an explicit modular structure of belief and strategy. Across a publicly available dataset (COCO-search18) and a dataset we collected, we found that estimated strategies do not align with the beliefs, deviating from an intuitive, maximum-seeking strategy. Furthermore, we explicitly tested whether people's choice of eye movements matches their beliefs using a novel gaze-contingent paradigm, and we found that where people shift their gaze to and where they believe the target to be can differ substantially. Taken together, these results suggest that people tend to prioritize information-seeking over belief maximization in naturalistic visual search.

Keywords: Visual search; Gaze prediction; Ideal observer

Methods

Template matching. In signal detection theory, the optimal signal for detecting the unknown target location is given by template matching (Green, Swets, et al., 1966). Targets in natural scenes can have a multitude of different appearances making simple template matching impossible. Instead, we used features from CLIP-ViT (Radford et al., 2021) and

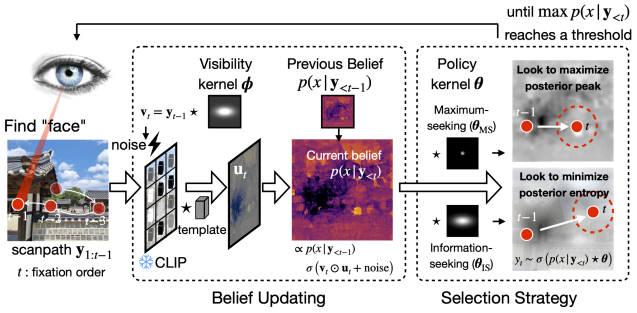

Figure 1: A normative framework that explicitly separates beliefs and gaze selection strategies in natural visual search.

modeled template matching as the cross-correlation $\mathbf{c}$ between the target text feature (in $\mathbb{R}^K$) and spatial features (in $\mathbb{R}^{H\times W\times K}$). The template responses were spatially normalized as $\mathbf{u}_t = \text{zscore}(\mathbf{c}) \in \mathbb{R}^{H\times W}$ and constant in a trial.

Optimal belief updating. A foveated observer should be constrained by their visibility, and we model such constraints with additive noise increasing with eccentricity. Optimal belief updating of the ideal observer, *knowing* their own visibility constraints, is given as Bayesian posterior updating (Najemnik & Geisler, 2005) about the target position x for each fixation order t in a given scan path $\mathbf{y}_{1:t}$,

$$p(x|\mathbf{y}_{<t}) \propto p(x|\mathbf{y}_{<t-1})\sigma\left(\mathbf{v}_t \odot \mathbf{u}_t - \mathcal{N}\left(\mathbf{v}_t/2, \text{diag}[\mathbf{v}_t]\right)\right) \quad (1)$$

$$y_t \sim \sigma\left(p(x|\mathbf{y}_{<t}) \star \theta\right) \quad (2)$$

where the posterior probability given the fixations history $\mathbf{y}_{<t}$ up to order t, $p(x|\mathbf{y}_{<t})$ is proportional to the product of the posterior before the last fixation $p(x|\mathbf{y}_{<t-1})$ and the sensory evidence at t. $\mathbf{v}_t = \mathbf{y}_{t-1} \star \phi$ defines the visibility at the $(t-1)^{\text{th}}$ gaze point via the cross-correlation ($\star$) with a trainable visibility kernel ϕ. Gaze position is selected from the softmax (σ) of the cross-correlation of the current posterior $p(x|\mathbf{y}_{<t})$ with a policy kernel θ (Butko & Movellan, 2008).

Gaze selection strategy. Optimal gaze selection strategy depends on the *goal* that we assume for the observer. Previous strategies include selecting the maximum of the current posterior (Maximum-Seeking, MS, *e.g.*, Wolfe (1994)) or selecting the location predicted to minimize posterior entropy (Information-Seeking, IS). It can be shown that the MS strategy corresponds to a Dirac delta policy kernel θ (Butko & Movellan, 2008), and the IS corresponds to the policy kernel proportional to ϕ (Najemnik & Geisler, 2009).

Training. We trained the model parameters ϕ, θ by maximiz-

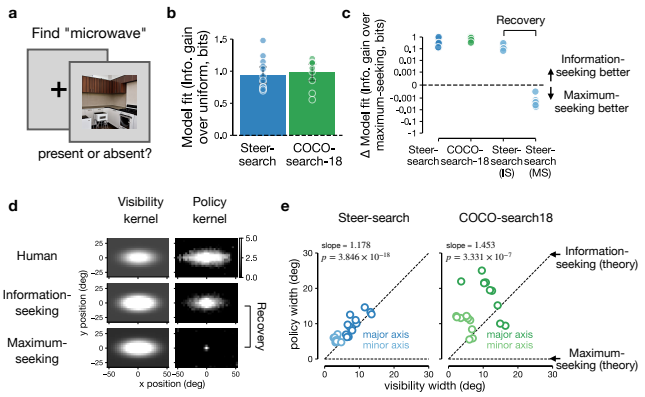

Figure 2: Normative modeling shows gaze selection deviates from maximum-seeking. **a**, Search tasks experiment structure. **b**, Cross-validated model performance (*information gain*, Kümmerer et al. (2015)) compared to the uniform baseline. **c**, Fits to IS and MS models, along with recovery fits from simulated gaze data of the fitted IS and MS models. **d**, Visibility and policy kernel estimates from fitting to COCO-search-18 (target-present) human data (top row) and estimates from fitting to simulated gaze data of IS and MS models (bottom rows). **e**, Comparison of the estimated kernel widths to the theoretical relationship of IS and MS searchers.

ing the per-fixation log-likelihood, next-fixation prediction goal.

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{T} \cdot \log p(\mathbf{y}_{1:T})\right] = \mathbb{E}\left[\frac{1}{T} \cdot \sum_{t=1}^{T} \log p\left(\mathbf{y}_t | \mathbf{y}_{<t}\right)\right] \quad (3)$$

with fixation length T. We approximate $\mathcal{L}$ via teacher forcing (Jordan, 1986), by sampling the posterior updates.

$$\mathcal{L} \approx \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \log p\left(\mathbf{y}_t | \hat{p}(x | \mathbf{y}_{<t})\right)\right] \quad (4)$$

Results

Normative model accurately predicts gaze positions

To test whether the normative models can accurately account for the human fixations, we collected human visual search data with EyeLink-1000 in saliency-controlled naturalistic images ("Steer-search" dataset, n_{subject}=13, n_{trial}=400, **Fig. 2a**). We cross-validated the model's prediction accuracy on both the steer-search and COCO-search18 datasets (Yang et al., 2020) and the models outperformed the uniform baseline in per-fixation log likelihoods (**Fig. 2b**).

Estimating gaze strategy via normative modeling

To infer the gaze selection strategy, we imposed the known relationship between visibility kernel (ϕ) and policy kernel (θ) in MS and IS searchers. Across datasets, MS models yielded higher validation log-likelihoods, supported by the recovery of MS and IS model fits within our framework (**Fig. 2c**). To further assess the strategy, we compared the widths of ϕ and θ.

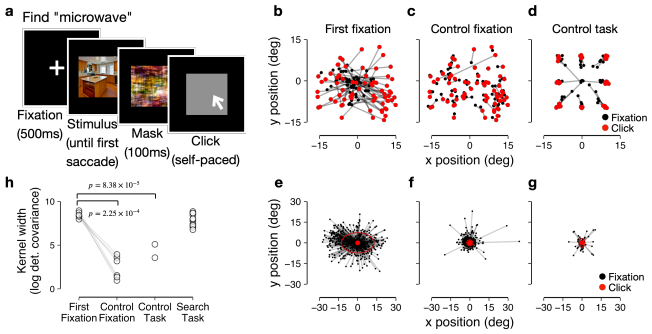

Figure 3: A gaze-contingent paradigm reveals a dissociation between belief and fixation selection strategies. **a**, Task structure. **b-d**, Example participant's locations of first fixations and click (**b**), control fixations and click (**c**), and first fixations in a control (directly-look) task (**d**). **e-g**, Population merged fixation points (first and control fixations and control task), aligned to the click locations. Red contour denotes the half-maximum of the Gaussian component in the Gaussian+uniform fit. **h** Click-aligned kernel widths (Gaussian determinant) compared to normative model policy kernels in **Fig. 2**.

Theoretically, MS observers use an infinitely narrow θ that allows the peak-selecting of posterior belief, while IS observer use θ shaped like ϕ, which allows selecting fixations further from the maximum points. Simulations confirmed that recovered policy kernels matched these expectations (**Fig. 2d**). Following this rationale, we fitted human data with Gaussian-parameterized ϕ and θ and found their width parameters were approximately matched, consistent with IS predictions and deviating from the constant-width MS baseline (**Fig. 2e**).

Estimating gaze strategy via behavior paradigm

To explicitly characterize the gaze selection strategy, we analyzed the relationship between initial saccadic intentions and target beliefs by designing a gaze-contingent paradigm using the same stimuli as Steer-search data (n_{subject}=6, n_{trial}=200; **Fig. 3a**). Participants viewed an image, which disappeared upon saccade detection, and clicked where they believed the target was. Misalignment between the initial saccade choice and the click indicates exploratory behavior, using their saccade to scan an area even if they do not necessarily believe the target to be, while alignment suggests the MS strategy. We compared first saccades to the "control saccade" (final saccades before clicking) and a "control task" where participants directly fixated and clicked on simple red-dot target images (**Figs. 3b-d**). Saccade endpoints were aligned by centering them on click positions to evaluate the dispersion of the strategy kernel (**Figs. 3e-g**). The estimated strategy matched the normative model's policy kernel width (**Fig. 3h**), deviating from the controls, together supporting that the human gaze selection deviates from maximum-seeking.

Conclusions

We tested whether human fixations in natural visual search reflect a strategy of belief maximization or information seeking, and found converging evidence that people's behaviors matched information-seeking policies. Taken together, these results suggest that people tend to prioritize information-seeking over belief maximization during natural visual search.

References

Butko, N. J., & Movellan, J. R. (2008). I-pomdp: An infomax model of eye movement. In *2008 7th ieee international conference on development and learning* (pp. 139–144).

Ding, Z., Ren, X., David, E., Vo, M., Kreiman, G., & Zhang, M. (2022). Efficient zero-shot visual search via target and context-aware transformer. *arXiv preprint arXiv:2211.13470*.

Droste, R., Jiao, J., & Noble, J. A. (2020). Unified image and video saliency modeling. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part v 16* (pp. 419–435).

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hosseini, A., Kazerouni, A., Akhavan, S., Brudno, M., & Taati, B. (2024). Sum: Saliency unification through mamba for visual attention modeling. *arXiv preprint arXiv:2406.17815*.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(3), 194–203.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 8).

Kümmerer, M., Bethge, M., & Wallis, T. S. (2022). Deepgaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), 7–7.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision research*, *49*(10), 1286–1294.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*, 202–238.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., ... Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 193–202).

Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., & Samaras, D. (2024). Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1683–1693).