It's a feature, not a bug: Multi-granular world models explain inattentional blindness

Mario Belledonne (mario.belledonne@yale.edu)
Dept. of Psychology, Yale University
100 College St., New Haven, CT 06510 USA

Ilker Yildirim (ilker.yildirim@yale.edu)
Dept. of Psychology, Yale University
100 College St., New Haven, CT 06510 USA

Abstract

Why do we sometimes miss what's right in front of us? Does our sometimes striking inability to notice a "gorilla" strolling in our midst repudiate the computational sophistication of human vision? Instead of regarding this and the related phenomena of inattentional blindess (IB) as a human quirk, here, we posit the opposite: that it is a signature of several advantageous computational adaptations. We realize this hypothesis by developing *multi-granular world models*, the first-ever model that reverse-engineers human IB by precisely capturing the elements relevant to our goals, while coarsely summarizing the rest scene.

Keywords: multi-granular world models; inattentional blindness; computational model; attention; perception

Introduction

Computational modeling has revealed multiple impressive facets of human perception, exhibiting invariance to geometric transformations (Han, Roig, Geiger, & Poggio, 2020) and efficiently inferring 3D objects (Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020). However, a large body of behavioral work (Mack, 2003) highlights a dramatic failure of human vision: we sometimes miss salient events in the world, especially when we are particularly focused on our goals. This phenomenon of "inattentional blindness" (IB) is often framed as a human quirk (Yuille & Liu, 2021).

Here, we reject this hypothesis and instead posit that IB is a result of sophisticated computation, not yet captured in computational vision. We hypothesize that IB arises from the *goal-conditioned approximation* of *multi-granular world models* in the mind, which flexibly and heterogeneously represent visual scenes, capturing precisely the elements more relevant to goals, while coarsely summarizing the rest. We realize this hypothesis with a computational architecture containing three core components: (1) *Multi-granular world models* - a novel class of generative model that supports representations at heterogeneous levels of resolution, (2) *Adaptive computation* - a dynamic algorithm to ration perceptual resources (Belledonne, Butkus, Scholl, & Yildirim, in press), and (3) *Granularity Optimizer* - a novel algorithm that adjusts object resolutions to efficiently satisfy task objectives.

We validate the resulting multi-granular world models in a case study of sustained IB (Most et al., 2001). In this paradigm, subjects are often unaware of the presence of an "invisible gorilla" - an additional object that moves through the center of the display - when focusing on a goal (counting the number of times the light objects bounce against the walls of the display; Fig.1). We recapitulate the primary pattern of human IB - that the "gorilla" is often missed when it appears similar to the task-irrelevant objects, and is often noticed when it appears similar to the task-relevant objects. Critically, we also show that this tradeoff is advantageous to performing the primary task: lesioned model variants significantly deviate from humans, detecting the gorilla more frequently and as a consequence, under performing in the primary task.

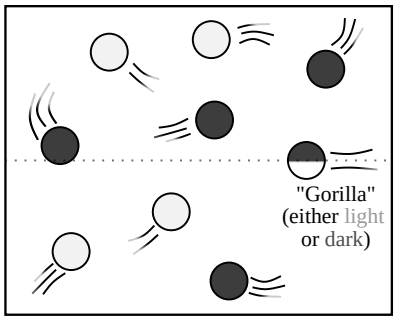

Figure 1: Sustained Inattentional Blindness task based on Most et al. (2001). Subjects count the number of times the light objects bounce against the walls of the display. At some random point, an additional object - the "gorilla", appearing either light or dark - moves through the center of the scene.

Model

Multi-granular world models define a generative model over how objects move and appear, critically, while also supporting these conditional probabilities at multiple levels of resolutions, simultaneously (Fig. 2, left). In the current case study, object representations take the form of either an individual or ensemble, collectively $S = \{s_1, ..., s_n\}$. An individual object consists of 2D position, instantaneous velocity, and shade (categorical: light or dark). An ensemble represents position with a 2D multivariate normal distribution with diagonal and constant covariance matrix, 2D velocity, and a mixture distribution over shade.

Objects move, $Pr(S^{t+1} \mid S^t)$, according to a transition kernel with Brownian dynamics over velocity[1]. In order to account of the appearance of the gorilla, a birth-death process (Karlin & McGregor, 1957) introduces a new object, once, anywhere in the display with a low probability (0.01).

At given timestep t, objects appear as an unordered set of detections, $X^t = \{x_1, ..., x_m\}$, where x_i denotes a noisy "detection" mask containing 2D location and shade. It's possible that an object generates more than one (or none) detections, with individual objects generally generating a smaller number of detections in tighter groupings (in terms of location and shade) than ensembles. Since no inherent mapping of $S^t \rightarrow X^t$ exists, the likelihood $Pr(X^t \mid S^t)$ is defined using random finite sets (RFS), which marginalizes across all valid mappings (Vo, Singh, & Doucet, 2005).

In theory, perception could invert this novel generative model to infer object states given a sequence of detections $Pr(\vec{S} \mid \vec{X})$, and decision-making could then use those states to count bounces $Pr(\pi \mid \vec{S})$. However, the flexibility of representation raises a critical question: **What level of granularity?**.

Adaptive computation We answer this question in two parts. First, in order to determine what objects are relevant, we integrate adaptive computation with a perception and decision making model for this case study. Adaptive computation (Belledonne et al., in press) is a general and dynamic attention algorithm for that rations perceptual computations C_k

[1] and also covariance in the case of ensembles

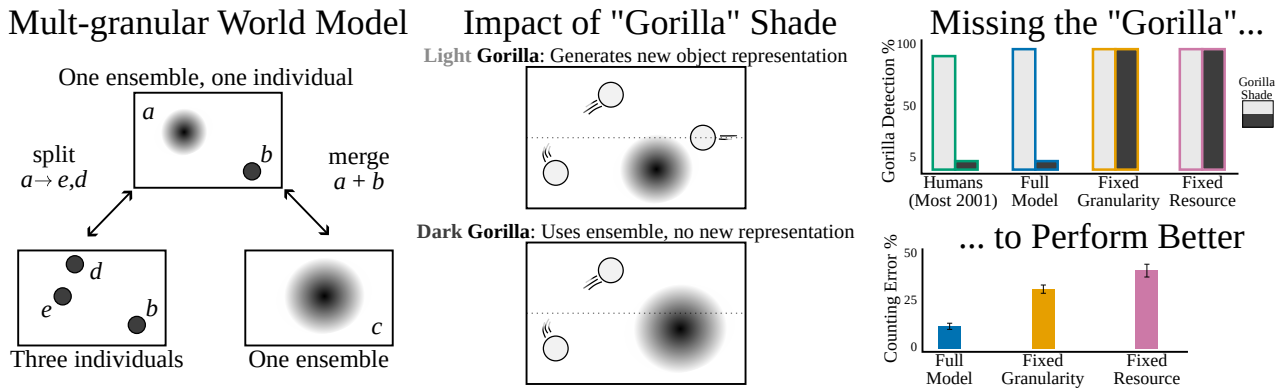

Figure 2: **Left**: Three distinct granularity schemas for a scene corresponding to three objects. **Middle**: Visualization of perceived object states after some period of granularity optimization. When the "gorilla" appears similar to the relevant objects (a light shade), it's sensory signal induces the model to generate a new, individual object representation - inducing a noticeable change in the model's mental state. When the gorilla instead appears similar to irrelevant objects (a dark shade), the model's coarse ensembles "explains away" its sensory signal without generating any new representations. **Right**: Both multi-granularity and adaptive computation are necessary to capture human patterns of inattentional blindness. Moreover, lesion models performed worse at the primary task, while also noticing the gorilla more often.

across objects and moments. It does so according to a task-relevance measure that bridges how computations readily improve object states $\delta_k S$ and further inform decision-making $\delta_k \pi$, $\Delta_k = \delta_k \pi \cdot \delta_k S$. Here, perceptual computations are defined in the context of perception implemented with a particle filter (Doucet, Freitas, & Gordon, 2001), where a small number of independent guesses over object states (e.g., 5 particles) approximate the posterior distribution. C_k corresponds to "rejuvenation moves" that iteratively improve the object states in a particle, e.g., $C_k(S) \rightarrow S'$ (object $a_k \in S \rightarrow a'_k \in S'$). Intuitively, only the light, individual objects are task-relevant $\Delta_k > 0$, since altering their state alone impacts counting $\delta_k \pi > 0$, especially so as they approach the walls [2].

Granularity Optimization Using the task-relevance across the representations $\vec{\Delta}$ for a given granularity schema G (e.g., Fig. 2, left), we can determine the efficiency of the schema with $\mho_G = \frac{\|\vec{\Delta}\|_2}{|\vec{\Delta}|}$. To build an intuition in the current domain, the schema where all dark objects are one ensemble and light objects are individuals maximizes $\mho_G$, as the dark objects have $\Delta_k \approx 0$. Interestingly, this schema (Fig. 2, middle) would lead to drastically different perceptual inferences over the presence of the "gorilla", as a dark gorilla's detections would be "explained" away by the dark ensemble, but for the light gorilla, generating a new object (the birth kernel) is more probable.

In order for the model to refine granularity on-the-fly, we implement a hierarchical particle filter (Yang, Duraiswami, & Davis, 2005), where each hyper-particle is itself a particle filter entertaining a granularity schema. Periodically[3], the algorithm estimates $\mho_G$ for each schema, and culls the weaker schemas. The surviving hyper-particles then "mutate", either refining (splitting) relevant representations or coarsening (merging) irrelevant representations to yield an altered schema for the next period. This process naturally yields more perceptually tractable coarse representations while ensuring that task-relevant details are preserved.

Results and Discussion

Across 20 trials following the procedure from Most et al. (2001), we evaluated the full model as well as two lesion models matched in terms of overall C_k: (1) *fixed-granularity* with all individual objects and adaptive computation, and (2) *fixed-resource* same as before but without adaptive computation[4]. Each model was evaluated with 10 independent runs per trial, and initialized to represent all object's initial positions with individual representations. For each run, we considered the model aware of the gorilla if it had tracked the gorillas detections using an individual object representation with above 50% confidence[5] for at least 6 frames.

Only the model with goal-conditioned multi-granular world models captured human IB patterns, almost never generating an individual object to track detections arising from the gorilla when it matched task *irrelevant* objects, while almost always doing so when the gorilla appeared similar to task relevant objects (Fig. 2, right). In contrast, the lesion variants almost always detected the gorilla in both conditions, and critically, performed substantially worse at the primary task. Thus, this case study shows that multi-granular world models provides the first-ever model of human IB while also illustrating the utility of such sophisticated computations. Future work will explore domains, such as Atari-like games, that more emphatically connect performance with capturing precisely those dimensions of the scene that matter most.

[2] generally $\delta_k S > 0$, unless an object has been repeatedly attended to

[3] after several observations

[4] e.g, uniform C_k

[5] under the RFS marginal

Acknowledgments

We'd like to thank the members of the Cognitive & Neural Computation and the Perception & Cognition labs at Yale. This work was supported by AFOSR grant # FA9550-22-1-0041.

References

Belledonne, M., Butkus, E., Scholl, B. J., & Yildirim, I. (in press). Adaptive computation as a new mechanism for human attention. *Psychological Review*.

Doucet, A., Freitas, N., & Gordon, N. (2001). *Sequential monte carlo methods in practice* (Vol. 1) (No. 2). Springer.

Han, Y., Roig, G., Geiger, G., & Poggio, T. (2020). Scale and translation-invariance for novel objects in human vision. *Scientific reports*, *10*(1), 1411.

Karlin, S., & McGregor, J. (1957). The classification of birth and death processes. *Transactions of the American Mathematical Society*, *86*(2), 366–400.

Mack, A. (2003). Inattentional blindness: Looking without seeing. *Current directions in psychological science*, *12*(5), 180–184.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological science*, *12*(1), 9–17.

Vo, B.-N., Singh, S., & Doucet, A. (2005). Sequential monte carlo methods for multitarget filtering with random finite sets. *IEEE Transactions on Aerospace and electronic systems*, *41*(4), 1224–1245.

Yang, C., Duraiswami, R., & Davis, L. (2005). Fast multiple object tracking via a hierarchical particle filter. In *Tenth ieee international conference on computer vision (iccv'05) volume 1* (Vol. 1, pp. 212–219).

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science advances*, *6*(10), eaax5979.

Yuille, A. L., & Liu, C. (2021). Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, *129*(3), 781–802.