Efficient inverse graphics with a differentiable generative model explains robustness of perception to unusual viewing angles

Hakan Yilmaz (hakan.yilmaz@yale.edu)
Department of Psychology, Yale University

Matthew Muellner (matt.muellner@yale.edu)
Department of Psychology, Yale University

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain & Cognitive Sciences, MIT

Katharina Dobs (katharina.dobs@gmail.com)
Department of Psychology, Justus-Liebig University Giessen

Ilker Yildirim (ilker.yildirim@yale.edu)
Department of Psychology, Yale University

Abstract

Upon foveating an object, in a few hundred milliseconds or less, we not only recognize the category or identity of that object, but also perceive its rich three-dimensional (3D) structure that causally underlies what we sense. Critically, this richness of perception is not brittle; our percepts may degrade under unusual or challenging viewing conditions, but they do so gracefully, typically remaining far above chance performance even when the best computer vision systems fail. What renders human perception distinct – with rich representations that are nevertheless robust to unusual viewing conditions – relative to the existing computer vision systems? Here we present a new computational architecture of perception that estimates 3D scene structure from real-world images in a fast bottom-up pass, and that can further refine this estimate via optimization under a differentiable generative model. In a case study of human face perception, we show that this model, and not the bottom-up only alternatives, matches human accuracy in both an upright and inverted face matching task. These results suggest that integrating discriminative and generative computations are needed to yield humanlike perception systems.

Keywords: inverse graphics; face perception; computational modeling; neural networks; robust vision

Introduction

When we encounter an object, we not only see its lower-level visual features (e.g., color, orientation), or attach high-level semantic labels (e.g., object category), but we also perceive rich three-dimensional (3D) scenes that causally underlie the inputs we sense (Olshausen, Mangun, & Gazzaniga, 2014). Such perceptual experiences come together with breathtaking speed, in a few hundred milliseconds or less (Grill-Spector & Kanwisher, 2005), and yet remain robust to unusual or challenging viewing conditions. In such atypical settings, perception is not brittle: It can certainly degrade and even slow down (both of which can be measured using objective, performance-based tasks), but performance typically remains far above chance, including in settings that render the accuracy of the best computer vision systems at chance level (e.g., Yildirim, Siegel, Soltani, Chaudhari, and Tenenbaum (2023)). What underlies the richness and robustness of biological vision – a significant goal post for machine vision systems?

Existing approaches individually do not address this question (DiCarlo et al., 2021). Discriminative models learn mappings from images to target variables (including in some cases over 3D scene structure (Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020)), offering fast, resource-rational mechanisms for visual inference, but they can be brittle when tested out of distribution (e.g., unusual viewing angles). Generative models encode a joint distribution of images and latent variables, and can in principle enable broader generalization with "fatter" tails, but inference can be costly.

Here, we present a new computational account of vision that integrates discriminative and generative computations in a single architecture. This architecture follows from a recent family of computer vision systems that combine test-time optimization with recognition models (e.g., Wu et al. (2017)). The model consists of a discriminative bottom-up pass to compute an initial percept – a 3D scene estimate – given an input image. This estimate is then refined via optimization over a differentiable generative graphics program (Fig. 1A). We call this model EDIG, short for efficient differentiable inverse graphics.

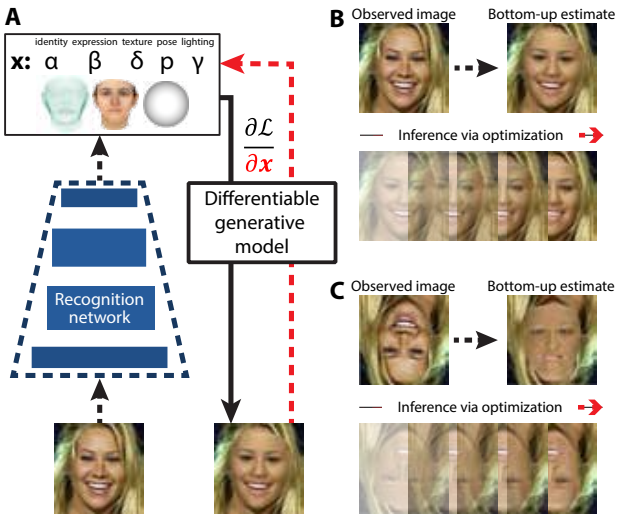

Figure 1: A: Our proposed architecture integrates a recognition network with a differentiable generative model. B & C: Visualization of bottom-up estimate and iterative optimization (upright & inverted).

As a case study, we implement EDIG in the context of the robustness of human face perception. It is well known that human face perception degrades when face stimuli are presented upside down; however, only a recent study quantified the extent of this performance gap using an identity matching task (Dobs, Yuan, Martinez, & Kanwisher, 2022). Surprisingly, their results indicate that the extent to which performance degrades in humans from upright to inverted faces is significantly less than the substantial drop of performance observed in a standard face identification network, illustrating the robustness of perception relative to standard vision models.

We show that EDIG matches human-level accuracy with just a single bottom-up pass for upright faces, whereas it requires several hundred iterations to match human-level accuracy in the inverted face recognition task. We report that this coincides with the increased response time in humans in the inverted task. We compare EDIG to its bottom-up-only ablation and a standard face-identification network finding that only EDIG reaches human level performance across tasks.

Computational model

Our model integrates a bottom-up discriminative recognition model with test-time optimization based on a differentiable generative graphics program. The generative model expresses a distribution over 3D scene structures based on the 3D Morphable Face Model (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009) – $\mathbf{x} = (\alpha, \beta, \delta, p, \gamma)$ denoting 3D shape,

expression, texture, pose, and lighting – and operates in both the top-down (for rendering) and the bottom-up (for gradient computation) directions. To realize this model for the perception of real-world human faces, we adapt the recognition network architecture introduced by Deng et al. (2019). This network is trained, in a semi-supervised manner, to map images of human faces to the generative model latents. To that end, Deng et al. (2019) use a pair of reconstruction-based objective terms for training: a photometric loss (comparing the input and predicted images) and a "perceptual loss" (comparing the encodings of the input and predicted images in a suitable embedding space – a later layer of a face classification network). In contrast to Deng et al. (2019), we employ the differentiable generative model also during test time, to refine the initial estimates generated by the recognition network.

During inference, we use the recognition network to initialize the latents **x**. For upright faces, this initialization is often remarkably accurate, with optimization leading to only marginal performance increase. On the other hand, when the observed image is inverted (i.e., flipped by 180°), we warm-start optimization by manually flipping the pose vector p and further refining it through an L2 loss based on five facial landmarks (eyes, mouth ends, nose) for 100 iterations. From this tuned pose vector and the rest of the bottom-up initialized latents, we backpropagate the photometric and feature encoding loss to optimize the latents for an additional 1000 steps. We use Adam with learning rate 0.001 as our test-time optimizer.

Training and alternative models We make comparisons to the bottom-up only component of the EDIG model, as well as a deep neural network trained for face identification reported in Dobs et al. (2022) that we refer to as ID. To equate these models with respect to their training experience, we train EDIG's bottom-up recognition model as well as the face identification network using the same 422k training images used in (Dobs et al., 2022). For each of these two networks, we tested two different architectures (ResNet50 and VGG16) and here we report the best performing architecture for each model.

Results

Identity matching task with upright and inverted faces We tested EDIG and alternative models on the identity matching task employed by Dobs et al. (2022). In this work, humans were asked to match the facial identity in a target image to one of two test images across 1560 unique match-to-sample triplets and without any time constraints. Across two experiments, images were shown upright or inverted (Fig. 2A). Average human accuracy for upright and inverted faces was 87.5% and 76.8%. Model accuracy was measured based on the Pearson dissimilarity using the shape + texture latents in EDIG and penultimate fully connected layer in ID between the target and each matching image.

EDIG explains human accuracy and response times Even though test-time optimization improves EDIG's performance in both the upright and inverted faces, this benefit is particularly pronounced for the inverted faces (Fig. 2B). We find that the particular choice of embedding space for optimization (alongside with the photometric loss) is important: Other choices such as the early or later layers of the bottom-up recognition network, or a later layer from an Imagenet-pretrained network performed worse. The alternatives we tested – the bottom-up recognition model and the ID network – did not reach human level accuracy, especially in the inverted setting (Fig. 2C).

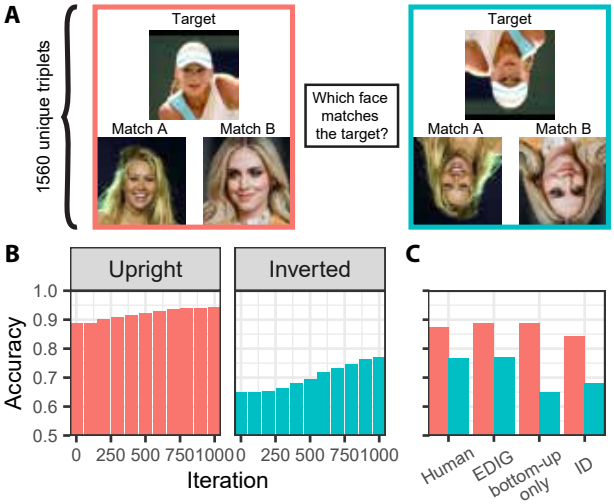

Figure 2: A: Identity matching task. B: EDIG task accuracy over optimization steps. C: Human vs. model comparisons.

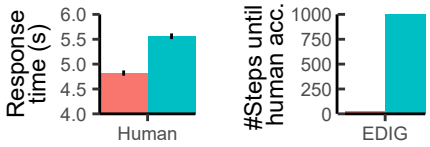

Figure 3: Human reaction time and EDIG optimization steps until human accuracy is reached.

Finally, we find that the amount of computation EDIG requires to match human-level performance parallels average human response times, who are significantly slower to respond in the inverted setting than the upright (Fig. 3, $p < .01$). In the upright setting, EDIG requires no iterations to match human performance, while 1000 iterations are needed to reach human accuracy in the inverted setting. Future work should make trial-level comparisons of response times.

Conclusion

We presented a new model of vision, EDIG, that combines a bottom-up inference network and test-time optimization in a unified architecture. In the domain of face perception, this architecture not only better explains robustness of human face perception to the inversion effect, but also shows signatures of higher computational demand when presented with unusual viewing angles. Future work will explore further integration of the recognition and generative models, afforded by the fact that both are differentiable and might be closely related to each other in their intermediate representations.

Acknowledgments

We thank the Yale Center for Research Computing for their support in managing the Milgram computing cluster.

References

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2019, June). Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 285–295). Long Beach, CA, USA: IEEE.

DiCarlo, J. J., Haefner, R., Isik, L., Konkle, T., Kriegeskorte, N., Peters, B., ... Yildirim, I. (2021, June). How does the brain combine generative models and direct discriminative computations in high-level vision? *CCN 2021 Workshop GAC*.

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2022, November). *Using deep convolutional neural networks to test why human face recognition works the way it does.* bioRxiv.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

Olshausen, B. A., Mangun, G., & Gazzaniga, M. (2014). 27 perception as an inference problem. *The cognitive neurosciences*, 295.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009, September). A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (pp. 296–301).

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W. T., & Tenenbaum, J. B. (2017). MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

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

Yildirim, I., Siegel, M. H., Soltani, A. A., Chaudhari, S. R., & Tenenbaum, J. B. (2023). 3d shape perception integrates intuitive physics and analysis-by-synthesis. *arXiv preprint arXiv:2301.03711*.