Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations

Ilker Yildirim (ilkery@mit.edu) [1]BCS, MIT [2]Lab of Neural Systems, RU

Tejas D. Kulkarni (tejask@mit.edu) BCS, MIT

Winrich A. Freiwald (wfreiwald@rockefeller.edu) Laboratory of Neural Systems, Rockefeller University

Joshua B. Tenenbaum (jbt@mit.edu) BCS, MIT

Abstract

A glance at an object is often sufficient to recognize it and recover fine details of its shape and appearance, even under highly variable viewpoint and lighting conditions. How can vision be so rich, but at the same time fast? The analysis-by-synthesis approach to vision offers an account of the richness of our percepts, but it is typically considered too slow to explain perception in the brain. Here we propose a version of analysis-by-synthesis in the spirit of the Helmholtz machine (Dayan, Hinton, Neal, & Zemel, 1995) that can be implemented efficiently, by combining a generative model based on a realistic 3D computer graphics engine with a recognition model based on a deep convolutional network. The recognition model initializes inference in the generative model, which is then refined by brief runs of MCMC. We test this approach in the domain of face recognition and show that it meets several challenging desiderata: it can reconstruct the approximate shape and texture of a novel face from a single view, at a level indistinguishable to humans; it accounts quantitatively for human behavior in "hard" recognition tasks that foil conventional machine systems; and it qualitatively matches neural responses in a network of face-selective brain areas. Comparison to other models provides insights to the success of our model.

Keywords: analysis-by-synthesis, 3d scene understanding, face processing, neural, behavioral.

Introduction

Everyday vision requires us to perceive and recognize objects under huge variability in viewing conditions. In a glance, you can often (if not always) identify a friend whether you catch a good frontal view of their face, or see just a sliver of them from behind and on the side; whether most of their face is visible, or occluded by a door or window blinds; or whether the room is dark, bright, or lit from an unusual angle. You can likewise recognize two images of an unfamiliar face as depicting the same individual, even under similarly severe variations in viewing conditions (Figure 1), picking out fine details of the face's shape, color, and texture that are invariant across views and diagnostic of the person's underlying physiological and emotional state. Explaining how human vision can be so *rich* and so *fast* at the same time is a central challenge for any perceptual theory.

The *analysis-by-synthesis* or "vision as *inverse graphics*" approach presents one way to think about how vision can be so rich in its content. The perceptual system models the generative processes by which natural scenes are constructed, as well as the process by which images are formed from scenes; this is a mechanism for the hypothetical "synthesis" of natural images, in the style of computer graphics. Perception (or "analysis") is then the search for or inference to the best explanation of an observed image in terms of this synthesis model: What would have been the most likely underlying scene that could have produced this image?

Figure 1: Same scene viewed at two different angles, illustrating level of viewing variability in everyday vision.

While analysis-by-synthesis is intuitively appealing, its representational richness is often seen as making inference highly impractical. There are two factors at work: First, in rich generative models a large space of latent scene variables leads to a hard search problem in finding a set of parameters that explains the image well. Second, the posterior landscape over the latent variables may have multiple modes or extended ridges of probability, making standard local search or stochastic inference methods such as Markov Chain Monte Carlo (MCMC) slow to burn in or mix, and potentially highly sensitive to the viewing conditions of scenes.

Here, we propose an efficient and neurally inspired implementation of the analysis-by-synthesis approach that can recover rich scene representations surprisingly quickly. We use a generic and powerful visual feature extraction pipeline to learn a recognition model with the goal of approximately "recognizing" certain latent variables of the generative model in a fast feed-forward manner, and then using those initial guesses to bootstrap a top-down search for the globally best scene interpretation. The recognition model is learned in an entirely self-supervised fashion, from scenes and corresponding images that are hallucinations from the generative model. We apply our approach to the specific problem of face perception, and find that (1) our recognition model can identify scene-generic latent variables such as object pose and lighting in a single feed-forward pass, and (2) brief runs of MCMC in the generative model are sufficient to make highly accurate inferences about object-specific latents, such as the 3d shape and texture of a face, when initialized by good guesses from the feed-forward recognition model.

Our approach is inspired by and builds upon earlier proposals for efficient analysis-by-synthesis such as the Helmholtz machine and breeder learning (Dayan et al., 1995; Nair, Susskind, & Hinton, 2008), but it goes beyond prior work in several ways:

- We apply this approach to much richer generative models than previously considered, such as near-photorealistic graphics models of faces based on high-dimensional 3d shape and texture maps, lighting and shading models, and

varying (affine) camera pose. This lets us perceive and recognize objects under much greater variability in more natural scenes than previous attempts.

- We directly compare human perceptual abilities with our model, as well as other recently popular approaches to vision such as convolutional neural networks (Krizhevsky, Sutskever, & Hinton, 2012).
- We explore this approach as an account of actual neural representations arising from single-unit cell recordings.

Face perception is an appealing domain in which to test our approach, for several reasons. First, faces are behaviorally significant for humans, hence an account of face perception is valuable in its own right, although we also expect the approach to generalize to other vision problems. Second, virtually all approaches to computational vision have been tested on faces (e.g., Taigman, Yang, Ranzato, & Wolf, 2014), offering ample opportunities for comparing different models. Third, the shape and the texture of faces are complex and carry rich content. Therefore, it provides a good test bed for models with rich representations. Finally, recent neurophysiology research in macaques revealed a functionally specific hierarchy of patches of neurons selective for face processing (e.g., Freiwald & Tsao, 2010). As far as high-level vision is concerned, this level of a detailed picture from a neural perspective is so far unheard of. Therefore, faces provide an excellent opportunity to relate models of high-level vision to neural activity.

The rest of this paper is organized as follows. We first introduce our efficient analysis-by-synthesis approach in the context of face perception. Next, we test our model in a computationally difficult task of 3D face reconstruction from a single image. We then describe a behavioral experiment testing people's face recognition abilities under "hard" viewing conditions, and show that our model best accounts for people's behavior. Finally, we show that our model bears some qualitative similarity to neural responses in the Macaque face processing system. We conclude with a discussion of quantitative comparisons between our model and alternatives that use only variants of bottom-up, recognition networks.

Model

Our model takes an inverse graphics approach to face processing. Latent variables in the model represent facial shape, S, and texture, T, lighting direction, l, and head pose, r. Once these latent variables are assigned values, an approximate rendering engine, $g(\cdot)$ generates a projection in the image space, $I_S = g(\{S,T,l,r\})$. See Figure 2a for a schematic of the model.

Following (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015), we use the Morphable Face Model (MFM; Blanz & Vetter, 1999) as a prior distribution over facial shapes and textures, S and T, respectively. This model, obtained from a dataset of laser scanned heads of 200 people, provides a mean face (both its shape and texture) in a part-based manner (four parts: nose, eyes, mouth, and outline)

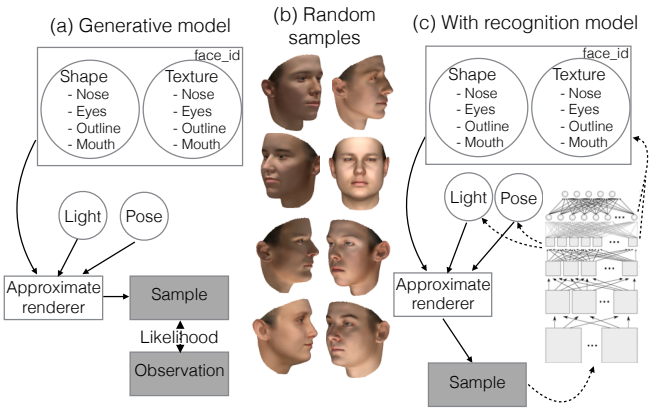

Figure 2: (a) Overview of the inverse graphics model. (b) Random draws from the model. (c) Training and the use of the recognition model.

and a covariance matrix to perturb the mean face to draw new faces by eigendecomposition. Accordingly, both the shape and texture take the form of multivariate Gaussian random variables: $S \sim N(\mu_{shape}, \Sigma_{shape})$ and $T \sim N(\mu_{texture}, \Sigma_{texture})$, where μ_{shape} and $\mu_{texture}$ are the mean shape and texture vectors respectively, and Σ_{shape} and $\Sigma_{texture}$ are the covariance matrices, each of which is set to be a unit diagonal matrix. The dimensionality of S and T are 200 each. The prior distributions over lighting direction and head pose are uniform over a discrete space (lighting direction could vary in elevation or azimuth in range -80° to 80°; the head pose could vary along the z-axis in range -90° to 90°, or on the x-axis in range -36° to 36°). Figure 2b shows several random draws from this model.

Given a single image of a face as observation, I_D, and an approximate rendering engine, $g(\cdot)$, face processing can be defined as inverse graphics in probability terms:

$$P(S,T,l,r|I_D) \propto P(I_D|I_S)P(I_S|S,T,l,r)P(S,T,l,r)\delta_{g(\cdot)} \quad (1)$$

The image likelihood is chosen to be noisy Gaussian, $P(I_D|I_S) = N(I_D; I_S, \Sigma)$. We set Σ to 0.05 in our simulations. Note that the posterior space is of high-dimensionality consisting of more than 400 highly coupled shape, texture, lighting direction, and head pose variables, rendering inference a significant challenge.

Recognition model

The idea of learning a recognition model to invert generative models has been proposed in various forms before (e.g., Dayan et al., 1995; Nair et al., 2008). We use a recognition model consisting of a generically trained deep Convolutional network (ConvNet) and linear mappings from that network to the latent variables in the generative model. To obtain this recognition model, we first used our generative model to hallucinate images from 300 different faces (each defined by a 3d shape and texture vector), and rendered each distinct face at 225 different viewing conditions (25 possible head poses $\times$ 9 possible lighting directions). Second, we used a ConvNet

trained on ImageNet (a labeled dataset of more than million images collected from the internet, Deng et al., 2009) that is very similar in architecture to that of (Krizhevsky et al., 2012) to obtain features for each of images in our dataset at all layers of the network.[1] In doing so, we first selected the "face-selective" units in each layer of the network by running a normal or a scrambled face test. The units that were activated twice as much to normal faces than to scrambled faces on the average (out of responses to 75 normal + 75 scrambled = 150 faces) were designated as "face-selective." Finally, we learn to construct bottom-up guesses for both scene-generic variables (pose and lighting direction) and object-specific latents (3d face shape and texture) via linear mappings from face-selective units in intermediate layers of the ConvNet. We have found that we can extract pose and lighting from the top convolutional layer (TCL) of the ConvNet, with close to perfect accuracy, using a linear support vector machine (SVM) for each combination of scene generic variables. We use a linear model with inputs from both TCL and the first fully connected layer (FFL) of the ConvNet to predict the shape and texture variables using Lasso regression (a schematic shown in Figure 2c).

Inference

Given an image, I_D, the recognition model described above makes fast bottom-up guesses about all latent variables in the generative model. Inference proceeds by fixing the head pose and the lighting direction variables to their "recognized" values, and then performing multi-site elliptical slice sampling (Murray, Adams, & MacKay, 2009), a form of MCMC, on the shape and texture vectors. At each MCMC sweep, we iterate a proposal-and-acceptance loop over eight groups of random variables: four shape vectors and four texture vectors, with one vector pair for each of four face parts (For more details see Kulkarni et al., 2015). In elliptical slice sampling, proposals are based on defining an ellipse using an auxiliary variable $x \sim N(0,\Sigma)$ and the current state of the latent variables, and sampling from an adaptive bracket on this ellipse based upon the log-likelihood function.

3d reconstruction from single images

Humans are capable of grasping much of the 3d shape and surface characteristics of faces or other objects from a single view, and can use that knowledge to recognize or imagine the object's appearance from very different viewpoints. We tested our model's capacity to perform this challenging task using a held-out set of faces (not among those used to build the generative or recognition models) from (Blanz & Vetter, 1999). Figure 3a shows several of these test faces as inputs, reconstructions based on only the bottom-up pass from the recognition model, and reconstructions from our full model after initializing with the recognition model and running MCMC to convergence. In addition to frontal faces, our model can reconstruct the shape and the texture of images of faces under non-frontal lighting and non-frontal pose, demonstrating robustness to non-standard viewing conditions and motivating the behavioral studies we describe below.

[1]We used the Caffe system to extract features, and also to train alternative networks that we describe later (Jia et al., 2014).

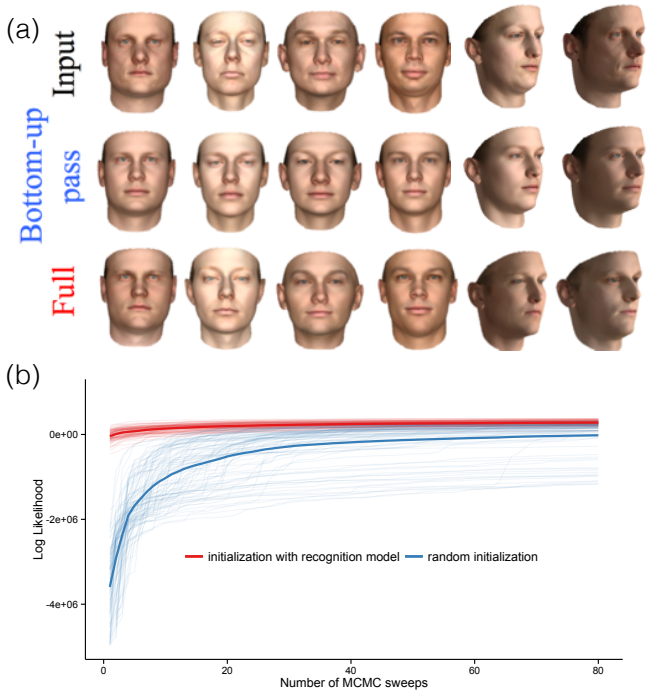

Figure 3: (a) Top: input images from a held-out laser scanned dataset (Blanz & Vetter, 1999). Middle: Reconstructions on the basis of the initial bottom-up pass. We used our generative process to visualize the shape and texture vectors obtained only from the recognition model. Bottom: Reconstructions after MCMC iterations. (b) The average and individual log-likelihood scores arising from randomly initialized 96 different chains vs. the recognition model initialized 96 chains. The recognition model initialized chains converge fast in less than 20 MCMC sweeps, and the variability across chains becomes much smaller.

Initializing inference for latent shape and texture variables using the recognition model dramatically improves both the quality and the speed of inference, as compared with the standard MCMC practice of initializing with random values (or samples from the prior). Figure 3b shows the log-likelihood traces of a number of chains for multiple input images that were initialized either randomly, or from the recognition model. Recognition-initialized chains converge much faster: In just a few MCMC sweeps, every chain reaches a log-likelihood that is almost as good as the best randomly initialized inference chains reach after tens or hundreds of sweeps. Furthermore, in comparison to the random initialization, recognition-model initialization leads to much lower variance: Inferences become uniformly good, very quickly.

Behavioral experiment

On common benchmark tasks for machine face recognition, the best systems now regularly report near-perfect perfor-

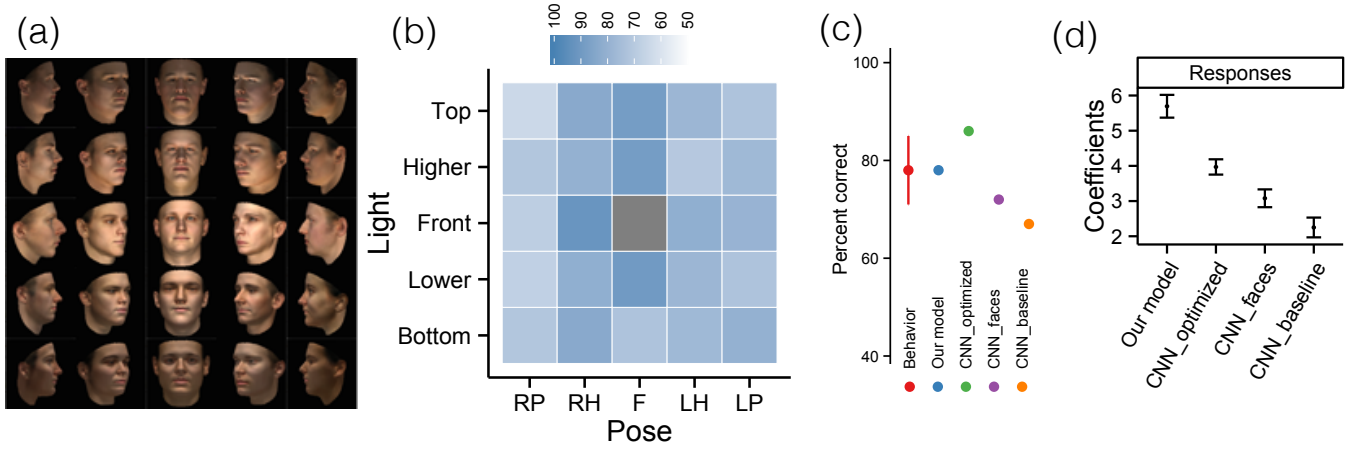

Figure 4: (a) Stimuli from the experiment illustrating the variability of lighting, pose, and identities. (b) Participants' average performance across all possible test viewing conditions. (c) Participants' and models' accuracy. (d) Coefficients of mixed effects logistic regression analyses. Error bars show standard deviations.

mance (e.g., Taigman et al., 2014). However, Leibo, Liao, and Poggio (2014) observed that most face databases are "easy", in the sense that the faces in the images are often frontal and fully visible. They found increasing viewing variability severely hurt the performance of these systems. Building upon this observation, we asked how well people can perform face recognition under widely varying pose and lighting conditions. The task was a simple passport-photo verification task: participants saw images of two faces sequentially, and their task was to judge whether the images showed the same person or two different people. Explaining human behavior in this task provides a challenging test for our model, as well as alternatives from the literature.

Participants

24 participants were recruited from Amazon's crowdsourcing web-service Mechanical Turk. The experiment took about 10 minutes to complete. Participants were paid \$1.50 (\$9.00/hour).

Stimuli and Procedure

The stimuli were generated using our generative model described above (Figure 2a). A stimulus face could be viewed at one of the five different poses (right profile to left profile) and under five different lighting directions (from top to from bottom), making a total of 25 possible viewing conditions. A subset of the facial identities and the 25 possible viewing conditions are shown in Figure 4a.

On a given trial, participants saw a study image for 750ms. After a brief period of blank interval (750ms), they saw the test image, which remained visible until they responded. They were asked to fixate a cross in the center of the screen at the beginning of each trial and between the study and the test stimuli presentations. The viewing condition for the study image was always frontal lighting at frontal pose (e.g., center image in Figure 4a). The viewing condition for the test image could be any of the remaining 24 possible combinations of lighting and pose. Participants judged whether two face images (study and test) belonged to the same person or to two different people, by pressing keys "s" for *same* or "k" for *different* on their keyboards.

There were a total of 96 trials, with 48 of the trials being *same* trials. Each test image viewing condition was repeated four times ($4 \times 24 = 96$), split half between *same* and *different* trials. The presentation order of the 96 pairs of images was randomized across participants. None of the identities was repeated except in the *same* trials, where the same identity was presented between the study image and corresponding test image. On *different* trials, faces were chosen to be as similar as possible while still remaining discriminable on close scrutiny.

Results

Participants performed well despite the difficulty of the task: Performance was above chance for all possible test face viewing conditions (Figure 4b), and ranged between 65% for light at the bottom and right-profile pose to 92% for frontal light and right-half profile pose. Overall, participants performed at an average accuracy of 78% (red dot and the associated error bars in Figure 4c), a level of performance that challenges even the most capable machine-vision systems.

Simulation details

We ran our model on the same 96 pairs of images that experimental participants saw. We ran at least 18 and at most 24 chains for each of the study and test images. Once initialized with the recognition model, each chain was run for 80 MCMC sweeps. Each chain simulated a participant in our study. For a given image, the values of the latent shape and texture variables from the last sample were taken as the model's representation of identity. We denote the representation of the study image i as $study_i$, and of the test image i as $test_i$ for $i \in 1, \ldots, 96$.

We calculated the performance of our model (and the alternative models that we introduce later) in the following man-

ner. We first scaled the study and the test image representations independently to be centered at 0 and have a standard deviation of 1.[2] Then, for each pair i, we calculated the Pearson correlation coefficient between the representations of the study and the test images, denoted as $corr_i$. Below, we used these pair-specific correlation values to model people's binary responses (*same* vs. *different*) in regression analyses.

Finally, we need to obtain *same* vs. *different* judgments from the model, to compare its performance with ground truth. Similar to an ROC analysis, we searched for a threshold correlation $\in [-1,1]$ such that the model's performance will be highest with respect to ground truth. Pairs of correlation values lower than the threshold were called *different*, and the pairs of equal or higher correlation values than the threshold were assigned *same*. We report results based upon the threshold that gave the highest performance.

Simulation results

Our inverse graphics model performs at 78% (see Figure 4c), matching the participants' average performance. Matching average level participant is an important criteria in evaluating a model, but only a crude one. We also tested whether the internal representations of our model ($corr_i$ for $i \in 1,\ldots,96$) could predict participants' *same/different* responses on unique stimuli pairs. We performed mixed effects logistic regression from our model's internal representations ($corr_i$ for $i \in 1,\ldots,96$) to participants' judgments, where we allowed a random slope for each participant using the *lme4* package and R statistics toolbox (R Core Team, 2013). The coefficient and the standard deviation estimated for our model are shown in Figure 4d. The internal representations of our model can strongly predict participants responses, providing evidence for an inverse graphics approach to vision ($\hat{\beta} = 5.69, \sigma = 0.26, p < 0.01$).

Macaque face patch system as inverse graphics

Encouraged by these behavioral findings, we next asked whether our model could explain neural responses in the face-processing hierarchy in the brain. Face processing is perhaps the best understood aspect of higher-level vision at the neural level. The spiking patterns of neurons at different fMRI-identified face patches in macaque monkeys show a hierarchical organization of selectivity: neurons in the most posterior patch (ML/MF) appear to be tuned to specific poses, neurons in AL (a more anterior patch) exhibit specificity to mirror-symmetric poses, and those in the most anterior patch (AM) show specificity to individuals but appear largely viewpoint-invariant (Freiwald & Tsao, 2010).[3]

We ran our model on a dataset of faces generated using our generative model, which mimicked the FV image dataset from Freiwald and Tsao (2010). Our dataset contained 7 head poses of 25 different identities under a fixed frontal lighting direction. We compared the representational similarity matrices of the population responses from Freiwald and Tsao (2010) in patches ML/MF, AL, and AM, and the representational similarity matrices arising from the representations of the different components of our recognition model: face-selective TCL units, face-selective FFL units, and the fast bottom-up guesses for shape and texture vectors.

ML/MF representations were captured best by the TCL activations (pearson correlation 0.67), suggesting that pose-specificity arises from a computational need to make inverse graphics tractable. Our results also suggest that this layer might carry information about the lighting of the scene, which is experimentally not systematically tested yet. AL representations were best accounted by the FFL activations (pearson correlation 0.67). Our model also provides a reason *why* mirror symmetry should be found in the brain: Computational experiments showed that mirror symmetry arises only at fully connected layers (i.e., dense connectivity) and only when the training data contains images of the same face from viewpoints distributed across both left and right sides. Our model captures AM patterns best via inferred latent shape and texture representations. The shape and texture vectors obtained just using the recognition model (that is, without running any MCMC iterations) captured AM responses best in comparison to all other layers in the recognition model (pearson correlation 0.42), suggesting the possibility of a generative 3D representation of shape and texture in the patch AM.

Discussion: Comparison to other models

In comparing our model against other approaches, we concentrated on alternatives that are based upon ConvNets, due to their success in many visual tasks including face recognition (DeepFace, Taigman et al., 2014), and the fact that they are architecturally similar (or even identical, in some cases) to our recognition model.[4] We considered three alternative models: (1) a baseline model, which simply is a ConvNet trained on ImageNet (CNN_baseline), (2) a ConvNet that is trained on a challenging real faces dataset called SUFR-W introduced in Leibo et al. (2014) (CNN_faces), and (3) a ConvNet that is selected from a number of networks that were all fine-tuned using samples from our generative model (CNN_optimized).

We focused on these alternative models' ability to explain our behavioral data. But we should note that ConvNets, on their own, cannot do 3D reconstruction. Also, even though each ConvNet can partially account for the neural data such as the pose specificity at patch ML/MF, they are worse at explaining the other two patches. The performance of all alternative models on our behavioral task was assessed just as for our model, with the only difference being that internal repre-

[2] This scaling step was not crucial for our model, but it was required to obtain the best out of other models that we will introduce below.

[3] Recent studies suggest homologue architecture between human and macaque face processing systems.

[4] We attempted to evaluate the DeepFace on our behavioral dataset. However, email exchanges with its authors suggested that a component of the model (3d spatial alignment) would not work with images of profile faces. Accordingly, we estimate the performance of that model on our behavioral dataset to be around 65%.

sentations of images were obtained as the FFL layer activations when that image is input to the model. For the mixed effects logistic regressions, a given pair of study and test images is represented by the correlation of the FFL activations for each of the two images.

The baseline model (CNN_baseline) performed at 67% (Figure 4c). This is impressive given that the model was not trained to recognize faces explicitly, and arguably justifies our use of ConvNets as good feature representations. The ConvNet trained on SUFR-W dataset (CNN_faces) performed at 72% (Figure 4c), closer to but significantly worse than human-level performance. We should note that CNN_faces is remarkable for its identification performance on a held-out portion of the SUFR-W dataset (67%; chance level $= 0.25\%$). The last ConvNet, CNN_optimized, performed *better* than people did with 86% (Figure 4c).

We are not the first to show that a computer system can top human performance in unfamiliar face recognition. However, we argue that the discrepancy between people and CNN_optimized points to the computational superiority of human face processing system: our face processing machinery is not optimized for a single bit information (i.e., identity), but instead can capture much richer content from an image of a face. This comes with a cost of accuracy in our same vs. different task. Our model accounts for the rich content vs. accuracy trade-off by acquiring much richer representations from faces while performing only slightly worse on identity matching than an optimized ConvNet. [5]

But, do people actually undertake the difficult challenge of 3d reconstruction when they look at unfamiliar faces? Our data suggests so: internal representations of the CNN_optimized, $corr_i$ for $i \in 1,\ldots,96$, is a worse fit to people's responses (also using a mixed effects logistic regression model; $\hat{\beta} = 3.97, \sigma = 0.22, p < 0.01$; Figure 4d). Indeed, none of the alternative models could account for participants' precise patterns of same/different responses as well as our model did(Figure 4d).

Do our computational and behavioral approaches extend to other object categories? A representational aspect of our model that lets us account for behavioral and neural data at the same time is that it represents 3D content in the form of a vector. Therefore, our approach should easily extend to other classes of 3D objects that can be represented similarly by vectors. Immediate possibilities include bodies, classes of animals such as birds, generic 3D objects such as vases, bottles, and so on. These object classes, in particular bodies, are exciting future directions, where revealing neural results have also been accumulating, our psychophysics methods can be straightforwardly extended to, and a generalization of our model already efficiently handles 3D reconstruction tasks (Kulkarni et al., 2015).

[5] We should also note that if we average the results across all the chains that we ran for each image, our model's performance significanly increases to 89% too.

Conclusion

This paper shows that an efficient implementation of the analysis-by-synthesis approach can account for people's behavior on a "hard" visual recognition task. The same model also solves a computationally challening task of reconstructing 3d shape and texture from a single image. Finally, it accounts qualitatively for the main response characteristics of neurons in the face processing system in macaque monkeys. None of the alternative ConvNet models, lacking a generative model and the capacity for top-down model-based inference, can account for all three of this phenomena. These results point to an account of vision with inverse graphics at its center, supported by bottom-up recognition models that can be learned from generative model fantasies in a self-supervised fashion, that allow top-down processing to refine their initial guesses but still do most of the work of inference in a bottom-up fashion, and that thereby enable even very rich model-based inferences to proceed almost as quickly as the fast feedforward processing of neural networks.

Acknowledgments

We thank the reviewers for their helpful comments. This research was supported by the Center for Brains Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and by ONR grant N00014-13-1-0333.

References

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, *7*(5), 889–904.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, ieee conference on* (pp. 248–255).

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: An imperative probabilistic programming language for scene perception. In *Computer vision and pattern recognition, IEEE conference on.*

Leibo, J. Z., Liao, Q., & Poggio, T. (2014). Subtasks of unconstrained face recognition. In *International joint conference on computer vision, imaging and computer graphics.*

Murray, I., Adams, R. P., & MacKay, D. J. (2009). Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*.

Nair, V., Susskind, J., & Hinton, G. E. (2008). Analysis-by-synthesis by learning to invert generative black boxes. In *Icann* (pp. 971–981). Springer.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Cvpr* (pp. 1701–1708).