Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal 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 robust and fast? The analysis-by-synthesis approach to vision offers an account of the richness of our percepts, but it is typically considered too fragile to apply robustly to real scenes, and 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 and robustly, by combining a generative model based on a realistic 3D computer graphics engine with a recognition model based on a deep convolutional network fine-tuned by brief runs of MCMC inference. 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*, so *robust* 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 power is generally considered to come at the cost of making inference almost impossible. There are two factors at work: First, the generative model's rich representational power leads to a very large space of latent scene variables to be inferred, and hence a hard search problem in inverting the graphics pipeline to get the right set of parameters that can explain the scene reasonably well. Second, the posterior space of the latent variables is highly multimodal, which could lead to local minima and potential high sensitivity to the viewing conditions of scenes.

Here, we propose an *efficient implementation* of the analysis-by-synthesis approach that is faster and more robust to viewing conditions while also preserving rich representations. In particular, we use deep learning approaches from computer vision to learn a recognition network with the goal of "recognizing" certain latent variables of the generative model in a fast feed-forward manner, and then using those initial guesses to bootstrap a search for the globally best scene interpretation. The recognition network is trained with scenes that themselves are hallucinations from the generative model. Wrapping this recognition network in our generative model leads to speed and robustness: (1) our system can "recognize" scene-generic latent variables such as pose or lighting in a single feed-forward pass, and (2) our inference algorithm starts sampling from a good initial guess of object-specific latents such as 3d shape and texture arising from the feed-forward pass. Our approach is inspired by and builds upon earlier proposals such as the Helmholtz machine and breeder learning (Dayan et al., 1995; Nair, Susskind, & Hinton, 2008), but it goes beyond these earlier proposals in the following three ways.

- We apply this approach to much richer generative models than previously considered, including 3d shape and texture, graphics rendering, lighting, shading, and pose. This lets us apply to approach to harder invariance problems in much more natural scenes.
- We test this approach using psychophysics as an account

of human behavior, and we compare it to other recently popular approaches to vision such as convolutional networks (Krizhevsky, Sutskever, & Hinton, 2012).

- We explore this approach as an account of actual neural representations arising from single-unit cell recordings.

Here, in order to assess how our model performs in these three fronts, we picked *face processing* as our domain of application. Face processing is an appealing domain for several reasons. First, faces are behaviorally significant, perhaps more so than any other object category. Therefore, an account of vision in the case of faces is very valuable, and can be generalized to a certain extent. Second, almost all types of modeling approaches have been tested on faces (e.g., Taigman, Yang, Ranzato, & Wolf, 2014). Therefore, there is plenty of opportunity 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 start with introducing our model implementing efficient and robust analysis-by-synthesis. Next, we test our model in a computationally difficult 3D face reconstruction from a single image task. Next, we describe a behavioral experiment testing people's face recognition abilities under "hard" viewing conditions, and show that our model can account for people's behavior. Following this, we show that our model qualitatively accounts for neural face processing system as documented in macaque monkeys. And finally, we discuss our model by comparing it to several alternative models, before concluding our paper.

Model

Our model takes an inverse graphics approach to face processing. Latent variables in the model represent facial shape, S, and texture, T, lighting, 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.

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) and a covariance matrix to perturb the mean face to draw new faces by eigendecomposition. Therefore, we can consider both shape and texture as 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 mean shape and texture vectors respectively, and Σ_{shape} and $\Sigma_{texture}$ are the covariances defining variability in shape and texture respectively. MFM is flexible in terms of the length of mean vectors. In our simulations we chose S and T to be vectors of length 200 each. The prior distributions over light source location and head pose are selected to be uniform over a discrete space (the light source could be at locations that vary in the elevation axis frontal to the face from $-80°$ to $80°$; the head pose could vary along the z-axis, from one profile view to the other). Figure 2b shows several random draws from this model.

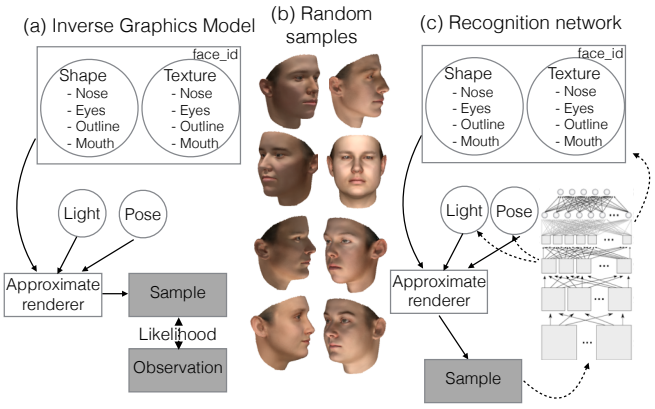

Figure 2: (a) Overview of the inverse graphics model. (b) Random draws from the model. (c) Recognition network to acquire good feature representations.

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 through the use of Bayes formula:

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

The likelihood of the model is chosen to be noisy Gaussian, $P(I_D|I_S) = N(I_D; I_S, \Sigma)$. Note that the posterior space is of high-dimensionality consisting of more than 400 highly coupled shape, texture, lighting, and head pose variables, rendering inference intractable.

Learning feature representations by training a recognition network

The idea of training a recognition network to invert generative models has been proposed in various influential forms (e.g., Dayan et al., 1995; Nair et al., 2008). Here, we build upon these ideas while taking advantage of the more successful feed-forward (i.e., bottom-up) architectures.

In particular, we took a Convolutional network (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), and fine-tuned its feature representations to turn the network to a bottom-up latent variable recognition pipeline for our generative model.[1] Similar in spirit to Nair et al.

[1] We used the Caffe system to train our recognition network (Jia et al., 2014).

(2008), we generated many hallucinations from our generative model for which we know the values of the latent variables, and then updated the weights of the ConvNet via back-propagation to predict these latents.

However, because the dimensionality of our latent space is very large, and because there is much coupling between the latent variables, this approach does not work as is. Instead, we fine-tuned the ImageNet trained network on a face identity dataset from our generative model to learn good feature representations. The dataset consisted of more than a 10000 labeled images of 200 facial identities projected under a variety of lighting conditions and poses. Pilot investigations suggested that the feature representations at layers "conv5" and "fc6" —the top convolutional layer and the first fully connected layer in the network— were most promising. As described below, we used linear mappings from these acquired feature representations to predict or to fix latent variables of the generative model in a bottom-up manner (Figure 2c).

Inference

We use the feature representations $\nu_{fc6}(\cdot)$ and $\nu_{conv5}(\cdot)$ arising from the fine-tuned ConvNet to learn a linear mapping from ConvNet features to each of the latent variables. In learning these mappings, we used Lasso linear regression, which is a linear model with L1 regularization on the weights:

$$\min_{w} \frac{1}{2N} \parallel \nu_{fc6}(I)w - v \parallel_2^2 + \alpha \parallel w \parallel_1 \qquad (2)$$

where $\nu_{fc6}(I)$ is the ConvNet representations of our training images, and v is a latent shape or texture variable, and α is the regularization term. We chose the value of α close to 0 on the basis of a held-out dataset. Similarly, we estimated linear mappings for the lighting, l, and pose, r, in the forms of $f_l : \nu_{conv5}(I) \rightarrow l$ and $f_r : \nu_{conv5}(I) \rightarrow r$. Our pilot experiments on a held out dataset indicated that these linear models worked very well for lighting and pose variables. Therefore, in our simulations, we decoupled l and r from each other and from the rest of the latent variables, and assigned them to the values "recognized" by our fine-tuned network in a bottom-up manner. However, linear mappings were not sufficient for the shape and texture latents. Instead, we resorted to Metropolis-Hastings sampling. We performed multi-site elliptical slice sampling (Murray, Adams, & MacKay, 2009) on the shape and texture latents. The basic idea is to define an ellipse using an auxiliary variable $x \sim N(0,\Sigma)$ and the current state of the random variables, and propose from an adaptive bracket on this ellipse based upon the log-likelihood function.

3d reconstruction from single images

Clearly, human vision is computationally extremely powerful. Models that want to be on par should be able to perform computationally hard but relevant challenges. 3d reconstruction from a single image under challenging viewpoints is one such challenge. Here, we tested our model on this task using a held-out data set from (Blanz & Vetter, 1999). In Figure 3a, top row shows several inputs to our model, whereas

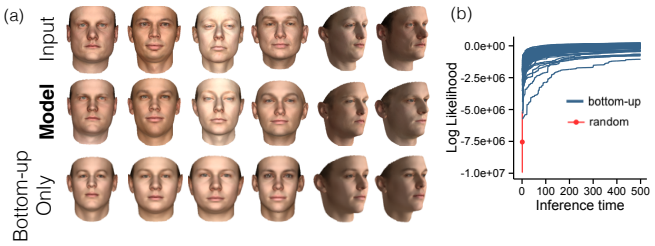

Figure 3: (a) Top: input images from a held-out laser scanned dataset (Blanz & Vetter, 1999). Middle: Our inverse graphics model's reconstructions. Bottom: Reconstructions on the basis of the initial bottom-up pass. (b) Log-likelihood scores arising from random vs. bottom-up initialization of Markov chains based upon a dataset of 2500 images. Error bars show standard deviation.

bottom row shows reconstructions based only on the bottom-up pass. Figure 3a, middle row shows the reconstructions by our model. 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.

The bottom-up initialization of the latent variables using a recognition network improves both the quality and the speed of inference. Figure 3b shows the log-likelihood traces of a number of chains for multiple input images that started with bottom-up initializations. As a comparison, we also show the initial log-likelihood score distribution of chains for 2500 different images that started from a random state. The bottom-up pass by itself gets about 70% of the whole improvement in log-likelihood score if one were to start from a random initialization, leading to more efficient search of the posterior space. Beyond that, in most cases, less than 100 iterations of inference is sufficient to achieve good results. And these results are robust to viewing conditions as shown Figure 3a, and as indicated by the successful modeling of behavior below.

Experiment

On common benchmark databases, face recognition appears to be solved by Machine Vision (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 deteriorated the performance of these systems. Building upon this observation, we asked the following research question: How well can people perform face recognition under "hard" invariance conditions? Here we operationally define "hard" invariance conditions as abundant variability in the viewing conditions. The task was as simple as the passport-photo verification task, where participants saw images of two faces sequentially, and their task was to judge whether the images belonged to the same person or to two different people. In addition to highlighting the extent of people's abilities, this experiment serves as a challenge for our model (and alternative models) to explain human behavior.

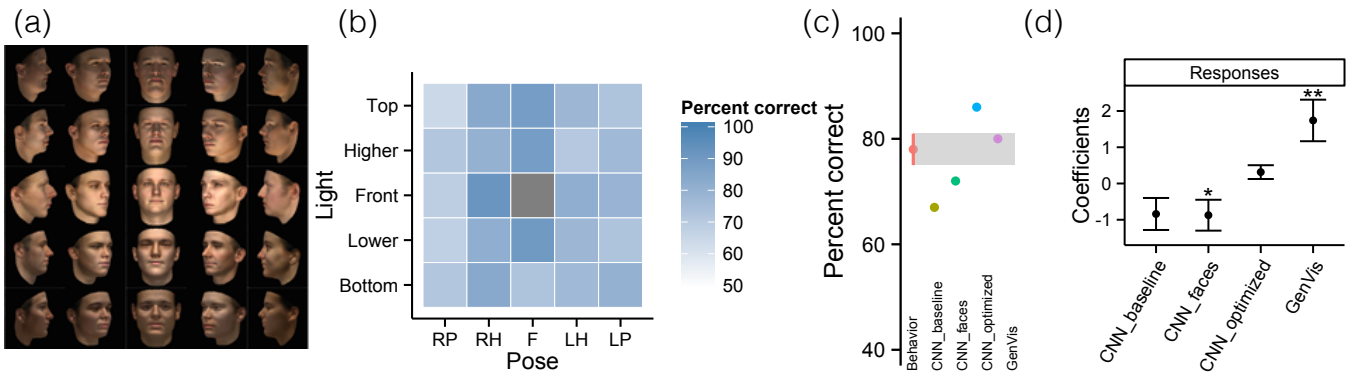

Figure 4: (a) Stimuli from the experiment illustrating the variability of lighting, pose, and identities. (b) Participants' average performance across all possible test viewing conditions. (c) Participants' and models' accuracy. Error bars shows 95% CIs. (d) Coefficients of mixed effects logistic regression analyses. Error bars show standard deviations.

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 conditions (at top to 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 briefly for 750ms. After a brief period of blank interval (750ms), they saw the test image, which remained visible until they responded. Subjects were asked to fixate to 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 conditions for the study image was always frontal lighting at frontal pose (e.g., center image in Figure 4a). The viewing conditions for the test image could be any of the remaining 24 possible combinations of lighting and pose. The participants' task was to judge whether the faces in the study and the test image belonged to the same person or to two different people. Participants entered their responses by pressing "s" for *same* or "k" for *different* on their keyboards.

There were a total number 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* *different* trials. The presentation order of the 96 pairs of images were randomized across subjects. None of the identities on the study images were repeated twice except the *same* trial identities, in which the identity presented twice (once in the study image and once in the corresponding test image). Care was given in choosing the faces of *different* trials so that they were not very different from each other.

Results

Participants performed remarkably despite the difficulty of the task: their performance was always above chance for all possible test face viewing conditions (Figure 4b; performance ranged between 65% for light at the bottom and right-profile pose to 92% for frontal light and right-half profile pose). In addition, there was not a strong pattern to suggest much systematic effect of viewing conditions over people's performance. Overall, participants performed at an average accuracy of 78% (red dot and the associated error bars in Figure 4c). This level of performance in this task challenges the most capable computational systems.

Simulation details

We ran our model on the identical 96 pairs of images as the participants in the experiment saw. For each pair, we ran our inference algorithm independently once for the study image and once for the test image for about 100 iterations. For a given image, the values of the latent shape and texture variables at the last iteration were taken as 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 manner. 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 measure its performance with respect to the ground truth. Similar to a 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. The

[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

search was such that the pairs of correlation values lower than the threshold were assigned *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 80% (the purple dot in Figure 4c, denoted as "GenVis"), closely matching to the participants' performance.[3] We note that matching people's overall performance 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 while allowing for a random slope for each participant. We performed this regression using the *lme4* package in R statistics toolbox (R Core Team, 2013). The coefficient and the standard deviation estimated for our model are shown in Figure 4d (denoted as "GenVis"). The internal representations of our model can strongly predict participants responses, providing further evidence for an inverse graphics approach to vision ($\hat{\beta} = 1.74, \sigma = 0.58, p < 0.01$).

Macaque face patch system as inverse graphics

Encouraged by the behavioral findings, we asked the following research question: Can our model explain the face processing hierarchy in the brain and generate testable predictions? To our advantage, face processing is the area of the visual neuroscience where we know most about higher-level vision. The spiking patterns of the neurons at different fMRI-identified face patches in macaques show a hierarchical organization of selectivity for faces: neurons at the more posterior patch (ML/MF) appear to be tuned to specific poses, AL (a more anterior patch) neurons exhibit specificity to mirror-symmetric poses, and the most anterior patch (AM) appear to be largely view-invariant, i.e., neurons there show specificity to individuals (Freiwald & Tsao, 2010).[4]

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 one lighting condition (e.g., Figure 5, top row). 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 model: $\nu_{conv5}(\cdot), \nu_{fc6}(\cdot)$ from the recognition network, and the latent shape, texture, pose, and lighting variables from the generative model.

[3] Our model with only bottom-up initializations of shape and texture performs at about 70%.

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

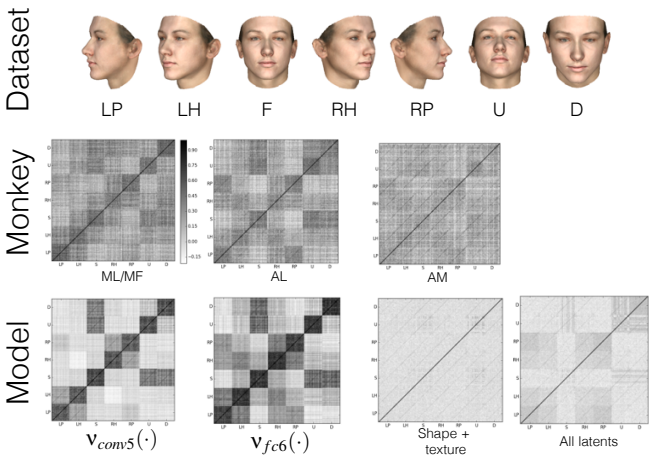

Figure 5: Side by side presentation of the neural data and the simulations. Middle row is included from (Freiwald & Tsao, 2010) with permission.

Figure 5 shows the results. ML/MF representations are captured best by the top convolutional layer of our recognition network $\nu_{conv5}(\cdot)$, 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 first fully connected layer of the network, $\nu_{fc6}(\cdot)$. Our models 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 when the training data contains mirror symmetric images of the same identity.

Our model captures AM patterns by the latent representations of the model. A representation that only consists of shape and texture variables achieves almost full invariance (bottom third column, Figure 5). However, we account for the neural data better only if we copy the scene-generic latent variables (lighting and pose) to the end of the shape and texture latents (last column, Figure 5), suggesting a *separable* and *equivariant* code at AM.

Discussion: Comparison to other models

We wish to compare our model against other approaches. This can be daunting if we were to consider all families of face processing models. Instead, we concentrated on models that are based upon ConvNets due to their success in many visual tasks including face recognition (DeepFace, Taigman et al., 2014).[5] 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

[5] 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%.

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 mainly compared these alternative models' performances in explaining 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, and mirror symmetry at AL, they cannot explain the observed view-invariance at AM.

All alternative models' performance on our behavioral dataset was obtained the same way we evaluated our model. The only difference in that for a given model, the internal representation of an image was obtained as the "fc6" layer activations given that image as 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 internal representations of the two images.

The baseline model (CNN_baseline) performed at 67% (Figure 4c). This is remarkable given that the model was not trained to recognize faces explicitly. This 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 tap 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 the 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 slightly worse than an optimized ConvNet.

This argument is supported by the behavioral data: internal representations of the CNN_optimized, $corr_i$ for $i \in 1,\ldots,96$, unlike our model, could not account for people's responses (another mixed effects logistic regression model; $\hat{\beta} = 0.31, \sigma = 0.19, p = 0.098$; Figure 4d). For that matter, none of the alternative models could account for participants' responses (Figure 4d).[6]

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 be easily extended to the 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, etc. These object classes, in particular bodies, are exciting future directions, where revealing neural results have been accumulating, our psychophysics methods can be straightforwardly extended to, and a generalization of our model already efficiently handles 3D reconstruction for these classes of objects (Kulkarni, Kohli, Tenenbaum, & Mansinghka, submitted).

[6]The significant negative coefficient for CNN_faces indicates that this model's error patterns should be different from that of people, which we confirmed quantitatively in another set of logistic regression models.

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. This same model also achieves the computationally very difficult task of reconstructing 3d shape and texture from a single image. Furthermore, it also accounts well for the currently known aspects of face processing system in macaque monkeys. None of the alternative ConvNet models can account for all three. These results point to an account of vision with inverse graphics at its center, where it is supported by recognition networks that provide speed and robustness via good feature representations.

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. (submitted). Picture: An imperative probabilistic programming language for scene perception..

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/` (ISBN 3-900051-07-0)

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Computer vision and pattern recognition, IEEE conference on* (pp. 1701–1708).