Explaining monkey face patch system as deep inverse graphics

Ilker Yildirim, Tejas D. Kulkarni, Winrich A. Freiwald, Joshua B. Tenenbaum

1. Summary The spiking patterns of the neurons at different fMRI-identified face patches show a hierarchical organization of selectivity for faces: neurons at the most posterior patch (ML/MF) appear to be tuned to specific view points, AL (a more anterior patch) neurons exhibit specificity to mirror-symmetric view points, and the most anterior patch (AM) appear to be largely view-invariance, i.e., neurons there show specificity to individuals (Freiwald & Tsao, 2010). Here we propose and implement a computational characterization of the macaque face patch system. Our main hypothesis is that face processing is composed of a hierarchy of processing stages where the goal is to "inverse render" a given image of a face to its underlying 3d shape and texture. The model wraps and fine-tunes a convolutional neural network (CNN) within a generative vision model of face shape and texture. We find that different components of our model captures the qualitative properties of each of the three face patches. ML/MF and AL are captured by different layers of the fine-tuned CNN, whereas AM is captured by the latent variables of the generative model. This modeling exercise makes two important contributions: (1) mirror symmetry (as in the patch AL) requires dense connectivity from the layer below and requires the agent to observe mirror-symmetric images that belong to the same identity, (2) AM is best explained by a representation that consists not only of latent shape and texture variables but also of the latent variables for generic scene variables such as pose and light location, indicating that this most anterior patch should be equivariant.

2. More

Recent work on the neural basis of face processing in the monkey inferotemporal cortex (IT) draws the picture of a system that is richly structured (e.g., Ohayon et al., 2012; Freiwald et al., 2009; Freiwald & Tsao, 2010; Leopold et al., 2006). Surely, the current picture of a hierarchical organization of the face processing system is far from complete. But it is sufficient to generate many potentially fruitful questions about neural computation: for example, what is the contribution of the pose-specificity of ML/MF to face recognition? Why and how does mirror symmetry arise in AL? How is view-invariance achieved at AM? We believe that a computational characterization of the monkey face processing system can lead to progress on all these questions by generating neurally testable predictions, or by simply providing bidirectional computational and biological insights for better understanding face recognition.

Here, we propose and implement a computational characterization of the macaque face patch system. Our main hypothesis is that face processing is composed of a hierarchy of processing stages where the goal is to "inverse render" a given image of a face to its underlying 3d shape and texture (Figure 1). The model wraps a convolutional neural network within a generative vision model of face shape and texture.

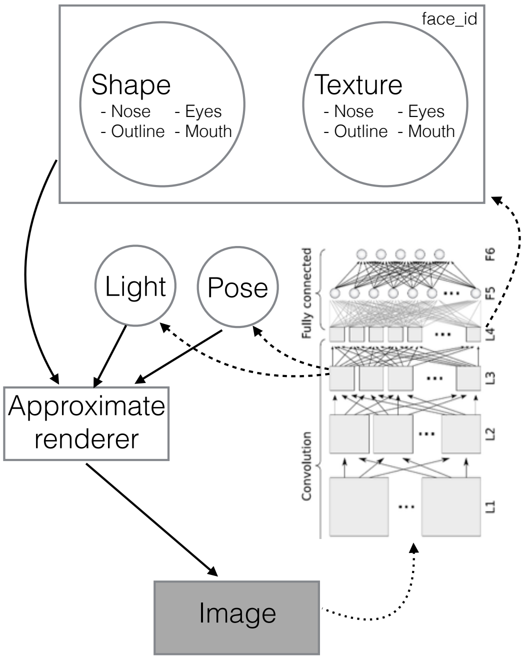

Figure 1: General overview of the inverse-rendering model.

The generative model treats the 3d shape and texture as well as the generic scene variables such as light source location and head pose as latent random variables. The prior over shape and texture is defined using the morphable face model (Blanz & Vetter, 1999), which provides a mean face (both its shape and texture) in a part-based manner (i.e., face is represented as composed of four parts: nose, eyes, mouth, and outline) and eigenvectors to perturb the mean face to draw new faces. The prior distributions over light source location and head pose are selected to be uniform over a discrete space.

We utilize a CNN to amortize inference in this model. The CNN is trained on a very large image corpus, which we fine-tune at different layers to predict different latent variables of the model. The fine-tuning of the network follows the idea of Helmholtz machine: model hallucinates faces from the prior where it knows the pose and light variables, and then fine-tunes the top convolutional layer of the network (layer 'conv5') to predict as best as possible those pose and light variables. Just above this top convolutional layer in the CNN is the first fully-connected layer (layer 'fc6'), which we can fine-tune to predict identity of the faces, with the idea that this will make the features at that layer more informative of the shape and texture.

During inference, light and pose variables are set in a bottom-up manner to the values predicted by the top convolutional layer of the network. This way of amortizing inference for light and pose on the basis makes sense in some analogy to variational inference: we introduce decoupling between the conditionally dependent latent variables. We expect this kind of amortizing to be generally useful for generative models of vision. Inference over shape and texture variables require MCMC sampling. We perform elliptical slice sampling to deal with the high dimensionality of the latents (Kulkarni et al., 2014). We penalize samples for both the CNN features and pixels.

We test the model on a dataset of faces generated from the morphable face model. Mimicking the FV image dataset from Freiwald & Tsao (2010), our dataset contains 7 views (left profile [LP], left half profile [LH], straight [S], right half [RH], right profile [RP], up [U], and down [D]) of 25 identities. We compare the representational similarity matrices of the neural responses from the data of Freiwald & Tsao (2010) (each of the ML/MF, AL, and AM while presented images from the FV image dataset) and the representational similarity matrices arising from the representations of the different components of our model (i.e., CNN layers conv5 and fc6, and the latent variables in the model). Our results suggest that ML/MF is characterized well by the top convolutional layer, whereas AL is best captured by the first fully connected layer in the network (Figure 2, the left two columns).

We capture AM patterns by the latent representations of the model. A representation that only considers shape and texture variables achieves almost full invariance (bottom third column, Figure 2). However, we can account for the neural data better only if we copy the generic but latent scene variables to the end of the shape and texture latents (last column, Figure 2). This results suggest that AM neurons consists of all the latent variables underlying a given scene, including the generic scene variables.

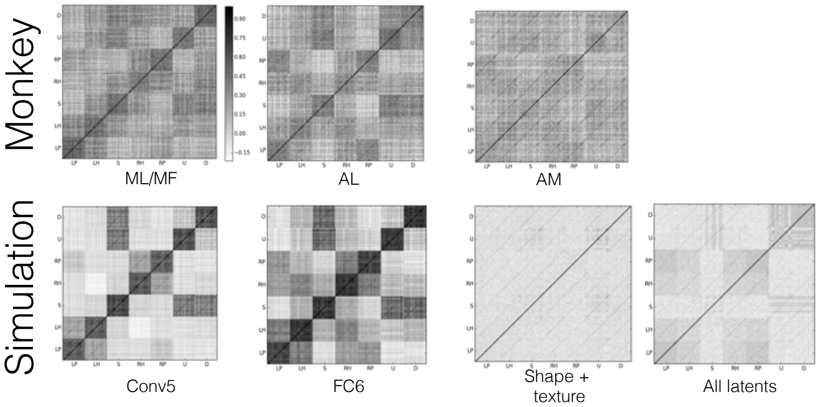

Figure 2: Side by side presentation of the neural data and the simulations.

Finally, what does give rise to mirror symmetry? We dissected our network in a 3 by 2 design experiment: possible computational operations (convolution or max pooling or dense connectivity) x kind of training data (mirror symmetric images of the same identity present or absent). We observed mirror symmetry only at fully connected layers (i.e., dense connectivity) and when the data contained mirror symmetric images of the same identity. Given that input is mirror symmetric for almost all types of objects as well as scenes in real world, and given the contribution of dense connectivity layers to increased classification accuracies in ANNs, we expect dense connectivity relatively frequently in more anterior stages of IT.