View-invariance and mirror-symmetric tuning in a model of the macaque face-processing system

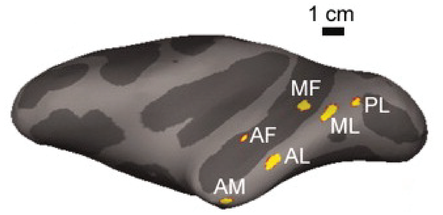
Joel Z Leibo, Fabio Anselmi, Jim Mutch, Akinori F Ebihara, Winrich A Freiwald, Tomaso Poggio

Center for Biological & Computational Learning

Introduction

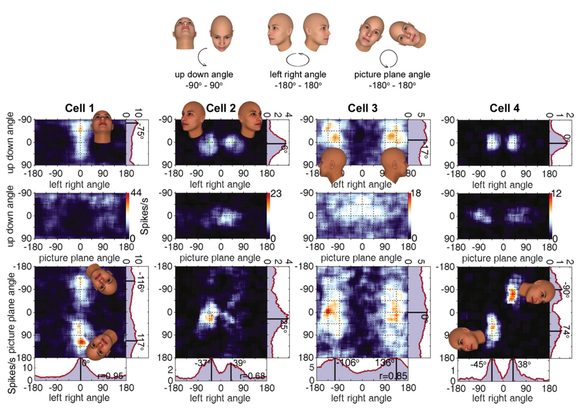
Recent experimental results characterizing the face processing network in macaque visual cortex pose a major puzzle. View-tuned units (found in patches ML/MF) are a natural step to a view-tolerant representation (found in patch AM), as predicted by several models. However, the observation that cells in patch AL are tuned to faces and their mirror reflections remains unexplained (cf. Freiwald and Tsao, 2010 & Leibo et al. 2011). We show that a model based on the hypothesis that the ventral stream implements a memory-based approach to transformation invariance predicts the main properties of ML/MF, AL and AM.

The macaque face-processing hierarchy



View-invariant ← View-specific

Patch AL

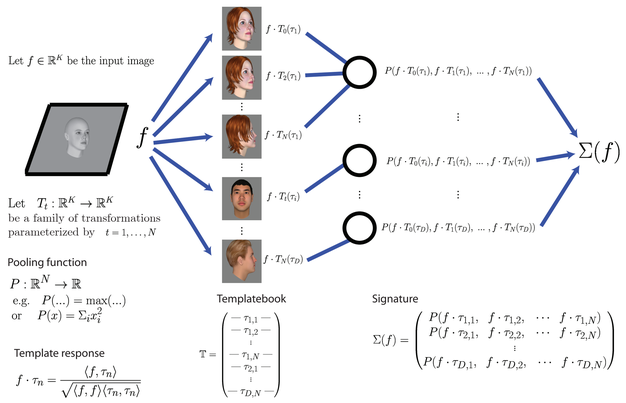


Freiwald et al. (2010) found that the macaque face patches differ qualitatively in how they represent identity across head orientations. Neurons in the middle lateral (ML) and middle fundus (MF) patches were view-specific; while neurons in the most anterior ventral stream face patch, the anterior medial patch (AM), were view-invariant. Puzzlingly, neurons in an intermediate area, the anterior lateral patch (AL) were tuned identically across mirror-symmetric views. That is, neurons in patch AL typically have bimodal tuning curves e.g., one might be optimally tuned to a face rotated 45 degrees to the left and 45 degrees to the right.

Average neuronal response latencies are 88 ms in ML/MF, 104 ms in AL, and 124 ms in AM; likewise, face stimuli evoke a change in the local field potential 126 ms after onset in ML/MF, 133 ms in AL and 145 ms in AM. Consistent with a hierarchical organization involving information passing from ML/MF to AM via AL, electrical stimulation of ML elicited a response in AL and stimulation in AL elicited a response in AM (Moeller et al. 2008). In addition, spatial position invariance increases from ML/MF to AL, and increases further to AM as expected for a feedforward processing hierarchy (Freiwald & Tsao 2010).

Many computational models of face (and object) recognition feature a progression from view-specific early processing stages to view-invariant later processing stages (similar to ML/MF and AM). However, it has thus far remained unclear why the brain would also utilize an additional intermediate step in which neurons confuse images and their mirror reflections (as in patch AL). We propose a new model of the macaque face-processing system in which an intermediate stage with properties similar to AL arises naturally as a consequence of the algorithm that we hypothesize is implemented by the ventral stream.

Hierarchical memory-based models



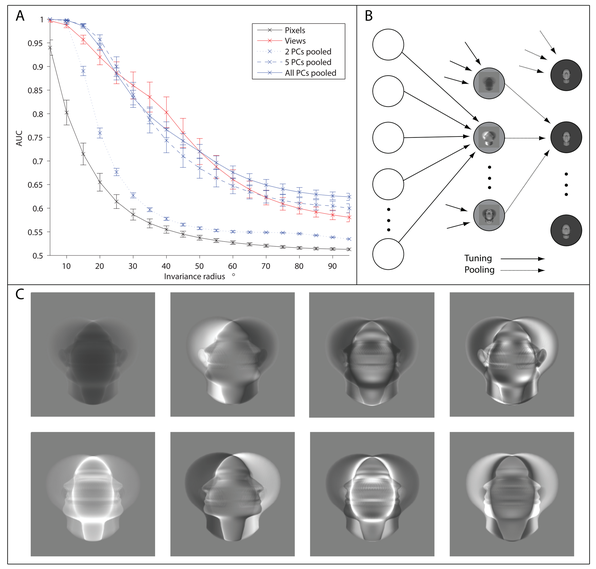
Assume a mechanism that stores "frames" as an initial pattern transforms from $t = 1$ to $t = N$ under the action of a specific transformation (such as rotation). This is the "developmental" phase of learning the templates. At run time, a set of normalized dot products with each of the stored templates and all their stored transformations is computed. The signature vector is produced by applying an aggregation function over the dot products with each template and its transformations. HMAX is a particular example of a hierarchical memory-based model that pools over translation and scaling (Riesenhuber & Poggio 1999, Serre et al. 2007). Leibo et al. (2011) (and others) investigated hierarchical memory-based models that pool over 3D-rotation-in-depth and showed that they can achieve good performance on viewpoint-invariant object recognition tasks.

Compressed models

As described above, the templates could be acquired directly as neural frames of the transformation video. However, there is no *a priori* reason to prefer these "directly sampled" templates. In fact, there are compelling arguments that a different method of obtaining templates is both algorithmically superior and more biophysically plausible.

We propose that the templates employed by the face patches are not directly sampled neural frames. Instead, they reflect a compressed version of the templatebook. More specifically, in our model, the templates represented in patch AL are the principal components (PC) of the templatebook.

Oja's rule (an approximation to the normalized version of Hebb's rule) converges to a solution where new inputs to the network are projected onto the first PC of the input's covariance (Oja 1992). There are also numerous extensions of Oja's rule which give additional PCs beyond the one with the highest eigenvalue. Thus, the assumption that synapses in the network are updated by a Hebb-like rule leads to the conclusion that the templates must be principal components of the videos of past visual experience.

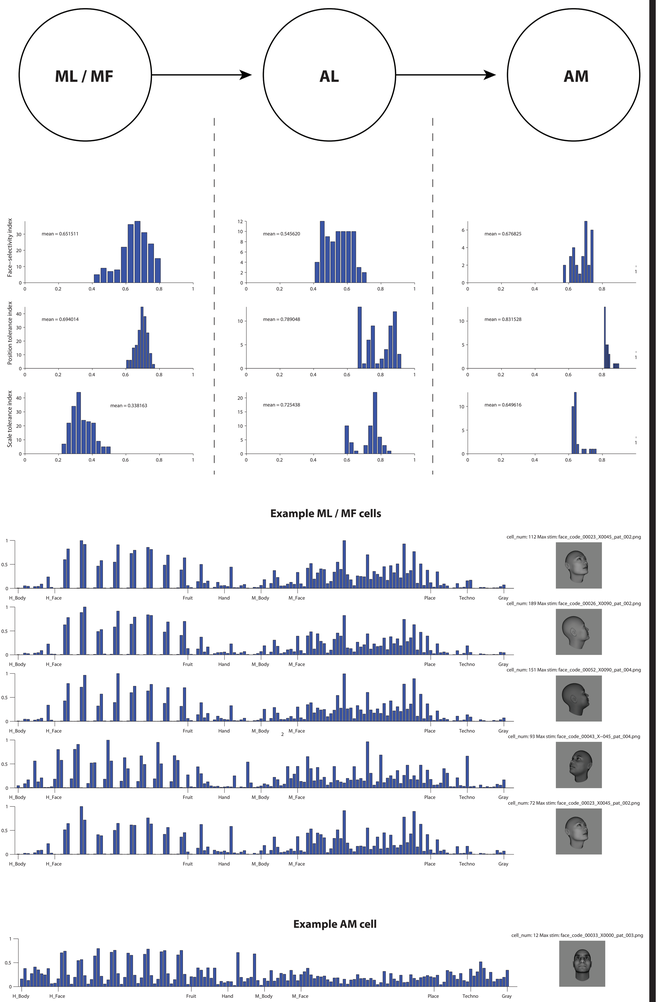


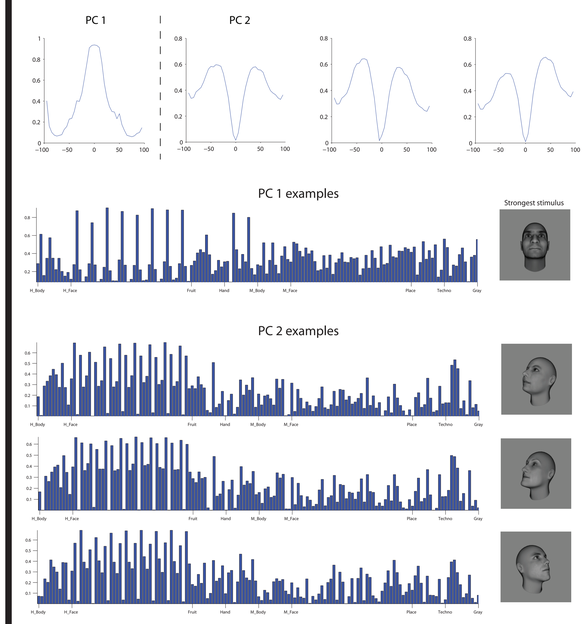
Principal components and mirror-symmetric tuning curves

A frontal view of a face is symmetric about its vertical midline. Thus equal rotations in depth e.g., 45 degrees to the left and 45 degrees to the right) produce images that are reflections of one another. Therefore, the templatebook obtained from a face's 3D rotation in depth must have a special structure. For simplicity, consider only symmetric transformation sequences, e.g., all the neural frames of the rotation from a left profile to a right profile. For each neural frame there must be a corresponding reflected frame in the templatebook. It turns out that as a consequence of its having this structure, the eigenfunctions of the templatebook will be even and odd. Therefore, the templates obtained from compressing the templatebook as though they were neural frames, are symmetric or anti-symmetric images.

Therefore, since we compute the template response using the absolute value of the normalized dot product of the input with a template, both even and odd templates yield tuning curves that show identical tuning to symmetric face views.

Quantitative predictions





Summary

We claimed that the goal of the ventral stream (including the face patches) is to compute a representation that is selective for objects (faces in this case) and invariant to identity-preserving transformations. A memory-based architecture can compute such a representation; by projecting the inputs onto neural frames (elements of the templatebook) and pooling. We propose that the ventral stream actually uses a compressed version of the templatebook (its principal components).

This proposal is supported by the fact that biologically-plausible learning rules (Hebb and Oja's rules) converge to solutions that project network inputs onto principal components of their covariance (Oja 1992). Furthermore, we showed that the PCs obtained from 3D rotations of faces are even and odd functions. Thus, we predict that neurons in the macaque face patch AL represent the (absolute value of the) projection onto these PCs. A simulated electrophysiology experiment shows that the tuning curves of the model's virtual AL cells resemble the tuning curves measured by Freiwald & Tsao (2010).

The same computational principle underlying the model of the macaque face-processing system also predicts Gabor tuning in V1. We conjecture that each ventral stream visual area sees its input through a different-sized aperture, and that this leads the cells in different areas to learn different optimal features that are useful for providing invariance to the transformations that are common for that aperture size.

Acknowledgments

Thanks to the Laboratory of Neural Systems at Rockefeller University and the McGovern Institute for Brain Research at MIT.

Support:
DARPA (IPTO and DSO), NSF and IIT.

Affiliations:
Center for Biological and Computational Learning, Cambridge MA 02139
McGovern Institute for Brain Research, Cambridge MA 02139
MIT Department of Brain and Cognitive Science, Cambridge MA 02139

References

1. Freiwald, W. A., Tsao D., Science. 330, 845 (2010)
2. Leibo J.Z., Mutch J., Poggio T. Advances in Neural Information Processing Systems (NIPS) (2011).
3. Oja E. Neural Networks 5:527-935 (1992).
4. Serre T., Wolf L., Bileschi S., Riesenhuber M., Poggio T. IEEE Trans. Pattern Anal. Mach. Intell. 29, 411-426 (2007).
5. Poggio T., Edelman S., Nature. 343, 6255 (1990).
6. Riesenhuber, M., Poggio, T., Nature Neuroscience. 1097-6256 (1999).
7. Moeller S., Freiwald W.A., Tsao, D. Science 320, 5881 (2008).
8. Poggio T., Mutch J., Anselmi F., Rosasco L., Leibo J.Z., Tacchetti A. MIT-TR-2012-035 *(2012)*.