Report

Current Biology

View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation

Highlights

- Biologically plausible learning rules yield view tolerance for faces
- The Oja rule predicts mirror-symmetric tuning in accordance with known physiology

Authors

Joel Z. Leibo, Qianli Liao, Fabio Anselmi, Winrich A. Freiwald, Tomaso Poggio

Correspondence

jzleibo@mit.edu

In Brief

Leibo et al. provide theoretical proof that a wide class of biologically plausible learning rules can wire up a feedforward network to compute a view-tolerant representation for bilaterally symmetric objects, like faces. One such rule, Oja's rule, yields a representation with mirror-symmetric head orientation tuning like that of the face patch AL.

Leibo et al., 2017, Current Biology *27*, 62–67
January 9, 2017
http://dx.doi.org/10.1016/j.cub.2016.10.015

CellPress

CellPress

Current Biology
Report

View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation

Joel Z. Leibo,[1,4,*] Qianli Liao,[1] Fabio Anselmi,[1,3] Winrich A. Freiwald,[1,2] and Tomaso Poggio[1]
[1]Center for Brains, Minds, and Machines and McGovern Institute for Brain Research at MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[2]Laboratory of Neural Systems, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA
[3]Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
[4]Lead Contact
*Correspondence: jzleibo@mit.edu
http://dx.doi.org/10.1016/j.cub.2016.10.015

SUMMARY

The primate brain contains a hierarchy of visual areas, dubbed the ventral stream, which rapidly computes object representations that are both specific for object identity and robust against identity-preserving transformations, like depth rotations [1, 2]. Current computational models of object recognition, including recent deep-learning networks, generate these properties through a hierarchy of alternating selectivity-increasing filtering and tolerance-increasing pooling operations, similar to simple-complex cells operations [3–6]. Here, we prove that a class of hierarchical architectures and a broad set of biologically plausible learning rules generate approximate invariance to identity-preserving transformations at the top level of the processing hierarchy. However, all past models tested failed to reproduce the most salient property of an intermediate representation of a three-level face-processing hierarchy in the brain: mirror-symmetric tuning to head orientation [7]. Here, we demonstrate that one specific biologically plausible Hebb-type learning rule generates mirror-symmetric tuning to bilaterally symmetric stimuli, like faces, at intermediate levels of the architecture and show why it does so. Thus, the tuning properties of individual cells inside the visual stream appear to result from group properties of the stimuli they encode and to reflect the learning rules that sculpted the information-processing system within which they reside.

RESULTS

The ventral stream rapidly computes image representations that are simultaneously tolerant of identity-preserving transformations and discriminative enough to support robust recognition. The ventral stream of the macaque brain contains discrete patches of cortex that support the processing of images of faces [8–11]. The face patches are arranged along an occipito-temporal axis (from the middle lateral [ML] and middle fundus [MF] patches, through the antero-lateral [AL] to the antero-medial [AM] patch; Figure 1A) [13]. Along this axis, response latencies increase systematically, suggesting sequential forward processing [7]. Selectivity to spatial position, size, and head orientation decrease from ML/MF to AM [7], replicating the general trend of the ventral stream [1, 14]. In accordance with these properties, many hierarchical models of object recognition [14–16] and face recognition [4, 17, 18] feature a progression from view-specific early processing stages to view-invariant later processing stages. They do so by successive pooling over view-tuned units.

Neurons in the intermediate face area AL, but not in preceding areas ML/MF, exhibit mirror-symmetric head orientation tuning [7]: an AL neuron tuned to one profile view of the head typically responds similarly to the opposite profile, but not to the front view (Figure 1B). This phenomenon is not predicted by classical and current hierarchical view-based models of the ventral stream, thus calling into question the idea that such models replicate the operations of the ventral stream.

Model Assumptions

Invariant information can be decoded from inferotemporal cortex, and the face areas within it, roughly 100 ms after stimulus presentation [19, 20]. This, it has been argued, is too fast of a timescale for feedback to play a large role [19, 21, 22]. Thus, whereas the actual face-processing system might operate in other modes as well, fundamental properties of shape selectivity and invariance need to be explained as a property of feedforward processing. We thus consider here a feedforward face-processing model.

The population of neurons in ML/MF is highly face selective [9]. Thus, as in earlier computational models of face *perception* [18, 23, 24], we assume the existence of a functional gate that routes only images of face-like objects at the input of the face system.

We make the standard assumption that a neuron's basic operation is a pooled dot product between inputs x and synaptic weight vectors $\{w_i^k\}$, which, in our settings, correspond to rotation in depth, $g_i \in G$, of template w^k. This yields a complex-like cell computing

$$\mu^k(x) = \sum_{i=1}^{|G|} \eta\left(\langle x, g_i w^k \rangle\right), \quad k = 1, \ldots, K, \qquad \text{(Equation 1)}$$

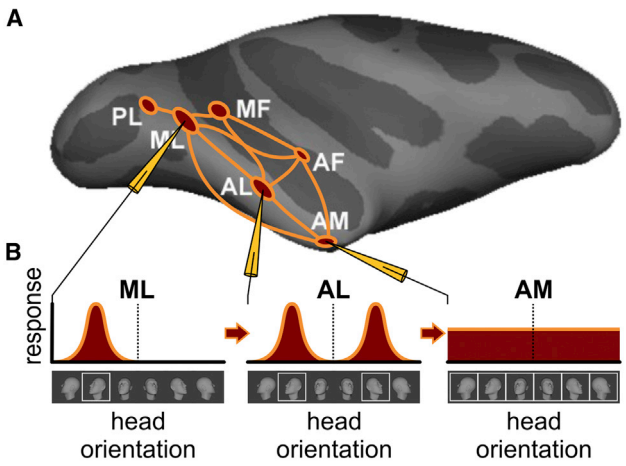

Figure 1. Schematic of the Macaque Face-Patch System

(A) Side view of computer-inflated macaque cortex with six areas of face-selective cortex (red) in the temporal lobe together with connectivity graph (orange) [13]. Face areas are named based on their anatomical location: AF, anterior fundus; AL, anterior lateral; AM, anterior medial; MF, middle fundus; ML; middle lateral; PL, posterior lateral, and have been found to be directly connected to each other to form a face-processing network [12]. Recordings from three face areas, ML, AL, and AM, during presentations of faces at different head orientations revealed qualitatively different tuning properties, schematized in (B).

(B) Prototypical ML neurons are tuned to head orientation, e.g., as shown, a left profile. A prototypical neuron in AL, when tuned to one profile view, is tuned to the mirror-symmetric profile view as well. And a typical neuron in AM is only weakly tuned to head orientation. Because of this increasing invariance to in-depth rotation, increasing invariance to size and position (not shown), and increased average response latencies from ML to AL to AM, it is thought that the main AL properties, including mirror-symmetry, have to be understood as transformations of ML representations and the main AM properties as transformations of AL representations [7].

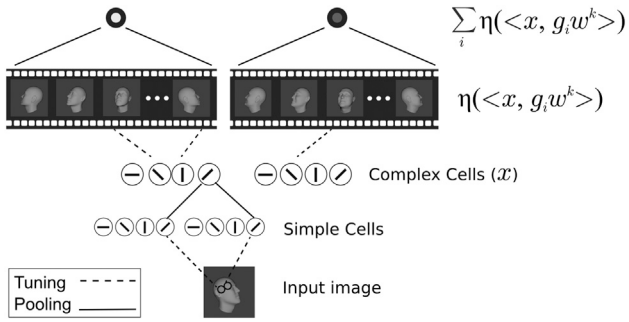

Figure 2. Illustration of the Model

Inputs are first encoded in a V1-like model. Its first layer (simple cells) corresponds to the S1 layer of the HMAX model. Its second layer (complex cells) corresponds to the C1 layer of HMAX [14]. In the view-based model, the V1-like encoding is then projected onto stored frames $g_i w^k$ at orientation i, from videos of transforming faces $k = 1, \ldots, K$. Finally, the last layer is computed as $\mu^k = \sum_i \eta(\langle x, g_i w^k \rangle)$. That is, the k^{th} element of the output is computed by summing over all responses to cells tuned to views of the k^{th} template face. In the PCA model, the V1-like encoding is instead projected onto templates w_i^k describing the i^{th} PC of the k^{th} template face's transformation video. The pooling in the final layer is then over all of the PCs derived from the same identity. That is, it is computed as $\mu^k = \sum_i \eta(\langle x, w_i^k \rangle)$. In both the view-based and PCA models, units in the output layer pool over all of the units in the previous layer, corresponding to projections onto the same template individual's views (view-based model) or PCs (PCA model).

where $\eta : \mathbb{R} \to \mathbb{R}$ is a nonlinear function, e.g., squaring [25] (Supplemental Mathematical Appendix, section 1.1). We call $\vec{\mu}(x) \in \mathbb{R}^K$ the signature of image x.

Approximate View Invariance

The model described so far encodes a given image of a face x by its similarity to a set of stored images of rotated in depth familiar faces (called templates) acquired during the algorithm's (unsupervised) training. The key observation to understand why the computation in Equation 1 gives a view-tolerant representation is that the similarity function of two faces has a sharp peak when they are at the same orientation. That is, $\langle x, g_i w^k \rangle$ is maximal when x and $g_i w^k$ depict faces at the same orientation, even when they depict different identities. Thus, because Equation 1 sums over all template orientations, it is always dominated by the contribution from the templates at orientations matching x. In other words, it is approximatively unchanged by rotation (we prove this in the Supplemental Mathematical Appendix, section 1).

Because this model is based on stored associations of frames (see Figure 2), it can be interpreted as taking advantage of temporal continuity to learn the simple-to-complex wiring from their view-specific to view-tolerant layers. They associate temporally adjacent frames from the video of visual experience as in, e.g., [26]. This yields a view-tolerant signature because, in natural video, adjacent frames almost always depict the same object [26–30]. Short videos containing a face almost always contain multiple views of the same face. There is considerable evidence from physiology and psychophysics that the brain employs a learning rule, taking advantage of this temporal continuity [31–34].

Thus, our assumption here is that, in order to get invariance to non-affine transformations (like rotation in depth), it is necessary to have a learning rule that takes advantage of the temporal coherence of object identity. More formally, this procedure achieves tolerance to rotation in depth because the set of rotations in depth approximates the group structure of affine transformations in the plane (see the Supplemental Mathematical Appendix, section 1). For the latter case, there are theorems guaranteeing invariance without loss of selectivity [2, 35].

Biologically Plausible Learning

The algorithm described so far can provide an invariant representation but is potentially biologically implausible: it requires the storing of discrete views observed during development. Instead, we propose a more biologically plausible Hebb-like mechanism. Instead of storing separate frames, cortical neurons update their synaptic weights according to a Hebb-like rule. Over time, they become tuned to basis functions encoding different combinations of the set of views. Different Hebb-like rules lead to different sets of basis functions, such as independent components (ICs) or principal components (PCs) [36]. Because each of the neurons becomes tuned to one of these basis functions instead of one of the views, a set of basis functions replaces the $g_i w^k$ in the pooling equation. The question then is whether invariance is still present.

The surprising answer is that supervised backpropagation learning and most unsupervised learning rules will learn approximate invariance to viewpoint for an appropriate training set (see section 2 of the Supplemental Mathematical Appendix for proof). This is the case for unsupervised Hebb-like plasticity rules, such as Oja's, Foldiak's trace rule, and independent component analysis (ICA), all of which provide bases over which the pooling equation provides invariance. One such Hebbian learning scheme is Oja's rule [37, 38]. It can be derived as the first-order expansion of a normalized Hebb rule. The assumption of this normalization is plausible, because homeostatic plasticity mechanisms are widespread in cortex [39].

For learning rate α, input x, weight vector w, and output $y = \langle x, w \rangle$, Oja's rule is

$$\Delta w = \alpha\left(xy - y^2 w\right) = \alpha\left(xx^T w - \left(w^T x x^T w\right) w\right). \qquad \text{(Equation 2)}$$

The weights of a neuron updated according to this rule will converge to the top PC of the neuron's past inputs, that is, to an eigenvector of the input's covariance C [37]. Thus, the synaptic weights correspond to the solution of the eigenvector-eigenvalue equation $Cw = \lambda w$. Plausible modifications of the rule—involving added noise or inhibitory connections with similar neurons—yield additional eigenvectors [38, 40]. This generalized Oja rule works as an online algorithm that computes the principal components of an incoming stream of images.

The outcome of learning is dictated by the underlying covariance of the inputs. Thus, in order for familiar faces to be stored so that the neural response modeled by Equation 1 tolerates rotations in depth of novel faces, we propose that Oja-type plasticity leads to representations for which the synaptic templates are given by PCs of an image sequence depicting the depth rotation of face k. Consider an immature functional unit exposed, while in a plastic state, to all depth rotations of a face. Learning will converge to the eigenvectors corresponding to the top r eigenvalues and thus to the subspace spanned by them. Equation 8 in section 2 of the Supplemental Mathematical Appendix shows that, for each template face k, the signature $\mu^k(x) = \sum_{i=1}^{r} \eta(\langle x, w_i^k \rangle)$ obtained by pooling over all PCs represented by different w_i^k is an invariant. This is analogous to Equation 1 with $g_i\, w^k$ replaced by the i^{th} PC. Section 2 of the Supplemental Mathematical Appendix also shows that other learning rules, for which the solutions are not PCs but a different set of basis functions, generate invariance as well—for instance, independent components. Figure 3B verifies that the signatures obtained by pooling over PCs are view tolerant through a simulation of the task of matching unfamiliar faces despite depth rotations.

Mirror Symmetry

Consider the case where, for each of the templates w^k, the developing organism has been exposed to a sequence of images showing a single face rotating from a left profile to a right profile. Faces are approximately bilaterally symmetric. Thus, for each face view $g_i\, w^k$, its reflection over the vertical midline $g_{-i}\, w^k$ will also be in the training set. It turns out that this property—along with the assumption of Oja plasticity, but not other kinds of plasticity—is sufficient to explain mirror-symmetric tuning curves. The argument is as follows.

Consider a face, x, and a set of its rotations in 3D,

$$O_x = \left(r_{\theta_{-N}} x, \ldots, r_0 x, r_0 x, \ldots r_{\theta_N} x\right),$$

where r_{θ_i} is a rotation matrix in 3D of angle θ_i, with respect to, e.g., the z axis.

Projecting onto 2D, we have

$$P(O_x) = \left(P\left(r_{\theta_{-N}} x\right), \ldots, P(r_0 x), P(r_0 x), \ldots, P\left(r_{\theta_N} x\right)\right).$$

Note now that, due to the bilateral symmetry, $r_{\theta_{-n}} x = RP(r_{\theta_n} x)$, $n = 0, \ldots, N$, where R is the reflection operator around the z axis. Thus, the above set can be written as

$$P(O_x) = (x_0, \ldots, x_N, Rx_0, \ldots, Rx_N),$$

where $x_n = P(r_{\theta_n} x)$, $n = 0, \ldots, N$. Thus, the set consists of a collection of orbits with respect to the group $G = \{e, R\}$ of the templates $\{x_0, \ldots, x_N\}$.

This property of the training set is needed in order to show that the signature $\mu(x)$ computed by pooling over the solutions to any equivariant learning rule, e.g., Hebb, Oja, Foldiak, ICA, or supervised backpropagation learning, is approximately invariant to depth rotation (Supplemental Mathematical Appendix, sections 1 and 2).

The same property of the training set, in the specific case of the Oja learning rule, is used to prove that the solutions for the weights (i.e., the PCs) are either even or odd (Supplemental Mathematical Appendix, section 3). This in turn implies that the penultimate stage of the signature computation, the stage where $\eta(\langle x, w \rangle)$ is computed, will have orientation tuning curves that are either even or odd functions of the view angle.

Finally, to get mirror symmetric tuning curves like those in AL, we need one final assumption: the nonlinearity before pooling at the level of the "simple" cells in AL must be an even nonlinearity, such as $\eta(z) = z^2$. This is the same assumption as in the energy model of [25]. This assumption is needed in order to predict mirror-symmetric tuning curves for the neurons corresponding to odd solutions to the Oja equation. The neurons corresponding to even solutions have mirror-symmetric tuning curves regardless of whether η is even or odd.

An orientation tuning curve is obtained by varying the orientation of the test image θ. Figure 3C shows example orientation tuning curves for the model based on a raw pixel representation. It plots $\langle x_\theta, w_i \rangle^2$ as a function of the test face's orientation for five example units tuned to features with different corresponding eigenvalues. All of these tuning curves are symmetric about 0° —i.e., the frontal face orientation. Figure 4A shows how the three model layers represent face view and identity, and Figure 4B shows the same for populations of neurons recorded in ML/MF, AL, and AM. The model is the same one as in Figures 3B and 3C.

These results imply that, if neurons in AL learn according to a broad class of Hebb-like rules, then there will be invariance to viewpoint. Different AM cells would come to represent components of a view-invariant signature—one per neuron. Additionally, if the learning rule is of the Oja type and the output nonlinearity is, at least roughly, squaring, then the model predicts that, on the way to view invariance, mirror-symmetric tuning emerges as a necessary consequence of the intrinsic bilateral symmetry of faces. In contrast to the Oja/principal component analysis (PCA) case, we show through a simulation analogous

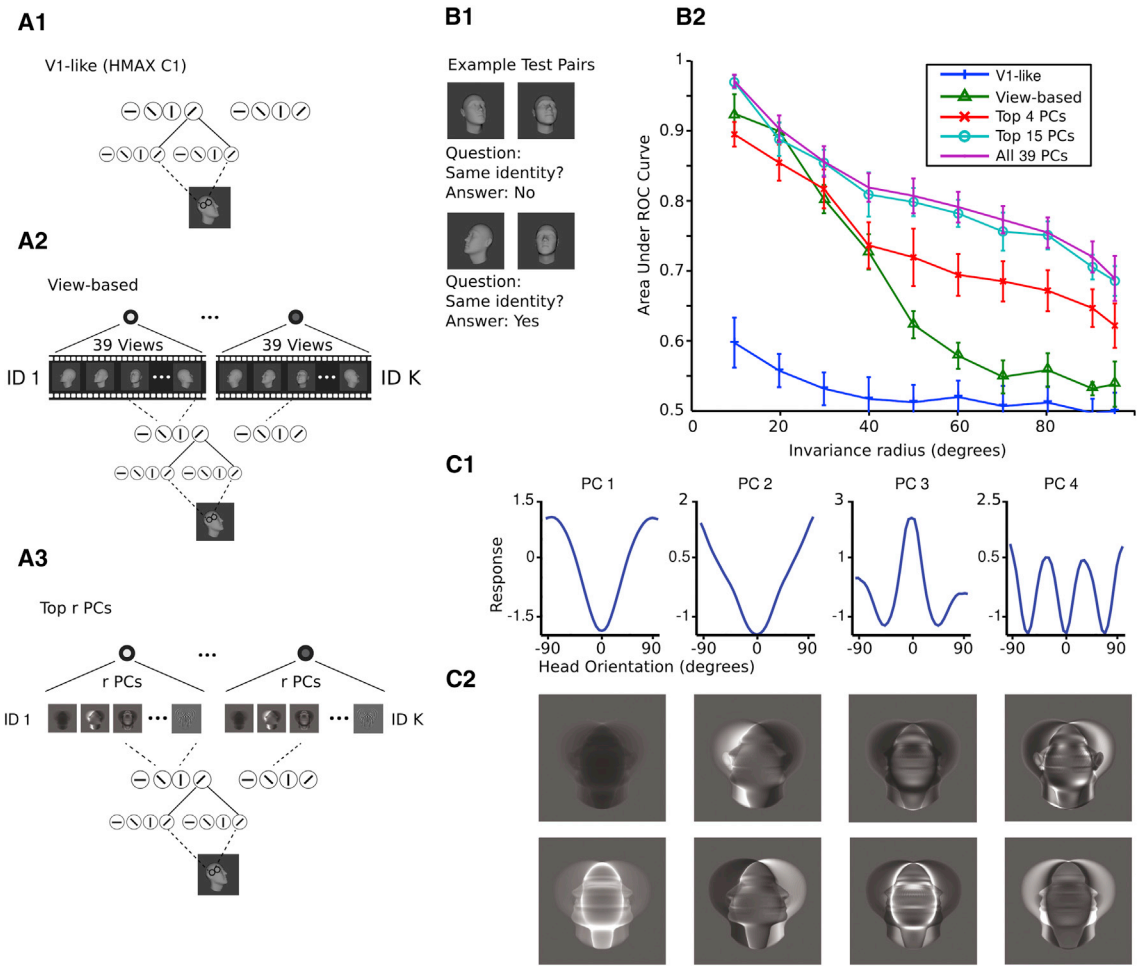

Figure 3. Model Performance on the Task of Same-Different Face Pair Matching

(A1–A3) The structure of the models tested in (B). (A1) The V1-like model encodes an input image in the C1 layer of HMAX, which models complex cells in V1 [14]. (A2) The view-based model encodes an input image as $\mu^k(x) = \sum_{i=1}^{|G|} \langle x, g_i w^k \rangle^2$, where x is the V1-like encoding. (A3) The top r PCs model encodes an input image as $\mu^k(x) = \sum_{i=1}^{r} \langle x, w_i^k \rangle^2$, where x is the V1-like encoding.

(B1 and B2) The test of depth-rotation invariance required discriminating unfamiliar faces. That is, the template faces did not appear in the test set, so this is a test of depth-rotation invariance from a single example view. (B1) In each trial, two face images appear and the task is to indicate whether they depict the same or different faces. They may appear at different orientations from each other. For classification of an image pair (a, b) as depicting the same or a different individual, the cosine similarity of the two representations was compared to a threshold. The threshold was varied systematically in order to compute the area under the receiver operating characteristic (ROC) curve (AUC). (B2) In each test, 600 pairs of face images were sampled from the set of faces with orientations in the current testing interval. Three hundred pairs depicted the same individual, and 300 pairs depicted different individuals. Testing intervals were ordered by inclusion and were always symmetric about 0° the set of frontal faces; i.e., they were $[-x, x]$ for $x = 5°, \ldots, 95°$. The radius of the testing interval x, dubbed the invariance radius, is the abscissa. AUC declines as the range of testing orientations is widened. As long as enough PCs are used, the proposed model performs on par with the view-based model. It even exceeds its performance if the complete set of PCs is used. Both models outperform the baseline HMAX C1 representation. The error bars were computed over repetitions of the experiment with different template and test sets; see the Supplemental Experimental Procedures.

(C1 and C2) Mirror-symmetric orientation tuning of the raw pixels-based model. $\langle x_\theta, w_i \rangle^2$ is shown as a function of the orientation of x_θ. Here, each curve represents a different PC. Below are shown the PCs w_i^k visualized as images. They are either symmetric (even) or antisymmetric (odd) about the vertical midline. See also Figure S1.

to Figure 3 that ICA does not generally yield mirror-symmetric tuning curves (Figure S1).

DISCUSSION

Neurons in the face network's penultimate processing stage (AL) are tuned symmetrically to head orientation [7]. In the model proposed here, AL's mirror-symmetric tuning is explained as a necessary step along the way to computing a view-tolerant representation in the final face patch: AM. The argument begins with a theoretical characterization of how an idealized temporal association learning scheme could learn view-tolerant face representations like those in AM. It then proceeds by considering which of the various biologically plausible learning rules satisfy requirements coming from the theory while also predicting mirror-symmetric representation in the computation's penultimate stage. It turns out that the Oja-like plasticity is the only biologically plausible rule that fits.

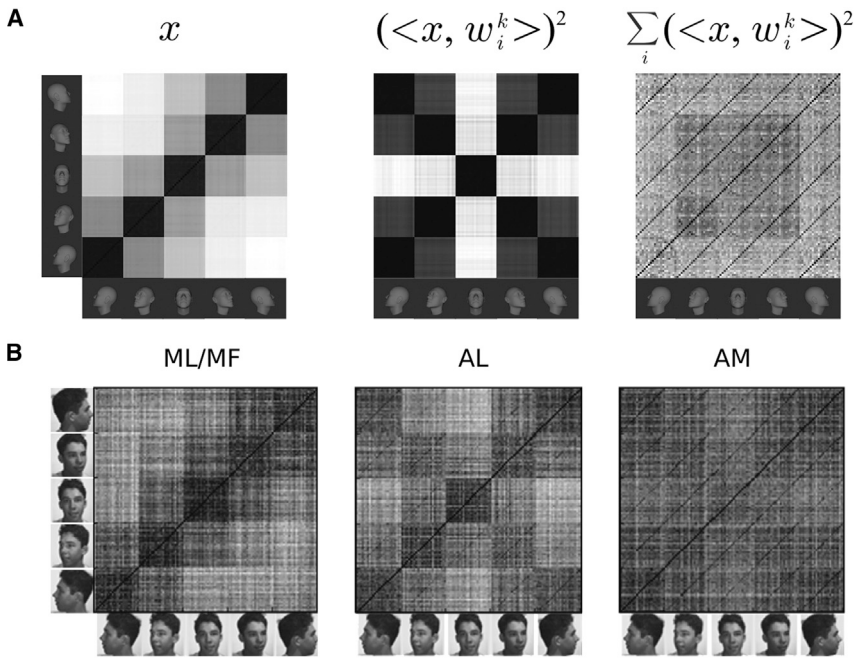

Figure 4. Population Representations of Face View and Identity

(A) Model population similarity matrices corresponding to the simulation of Figure 3B were obtained by computing Pearson's linear correlation coefficient between each test sample pair. Left: the similarity matrix for the V1-like representation, the C1 layer of HMAX [14]. Middle: the similarity matrix for the penultimate layer of the PCA-model of Figure 3B. It models AL. Right: the similarity matrix for the final layer of the PCA-model of Figure 3B. It models AM.

(B) Compare the results in (A) to the corresponding neuronal similarity matrices from [7].

See also Figure S1.

This argument suggests that Oja-like plasticity may indeed be driving learning in AL. We now discuss potential implications of this hypothesis. Whereas Oja's learning rule was originally motivated on computational grounds [37], it is now believed to be a widespread mechanism in cortex because it describes the interaction of long-term potentiation (LTP), long-term depression (LTD), and synaptic scaling (a form of homeostatic plasticity) [39]. Biophysical mechanisms underlying homeostatic plasticity include activity-dependent scaling of AMPA receptor recycling rates [39]. It has been observed in mouse visual cortex in vivo under bi/monocular deprivation paradigms [41, 42] and is believed to be particularly important in developmental critical periods [43]. Such synaptic scaling has not yet been described in the face patch network; however, the present work can be seen as lending support to the hypothesis that it may act there. Just as monocular deprivation paradigms are used to study homeostatic plasticity in visual cortex, analogous "face deprivation" paradigms may reveal homeostatic plasticity in the developing face patch network.

In particular, to the best of our knowledge, this is the first account that explains why cells in the face network's penultimate processing stage, AL, are tuned symmetrically to head orientation. This shows that feedforward processing hierarchies can capture the main progression of face representations observed in the macaque brain. The mirror symmetry result is especially significant because it shows how a time-contiguous learning rule operating within this architecture can give rise to a counter-intuitive property that is not intrinsic to the temporal sequence of the stream of visual images impinging on the eyes. Rather, it arises as an interaction between an intrinsic property of the geometry of an object, here bilateral symmetry, and the computational strategy of the information-processing hierarchy, that is view invariance.

Our model is designed to account only for the feedforward processing in the face patch network ($\lesssim$80 ms from image onset). The representations computed in the first feedforward sweep are most likely used to provide information about a few basic questions, such as the identity or pose of a face. Feedback processing is most likely more important at longer timescales. What computations might be implemented by the feedback processing occurring on longer timescales? One hypothesis addressed by the recent work of [44] combines a feedforward network like ours—also showing mirror-symmetric tuning of cell populations—with a probabilistic generative model. Thus, our feedforward model may serve as a building block for future object-recognition models addressing brain areas, such as prefrontal cortex, hippocampus, and superior colliculus, integrating feedforward processing with subsequent computational steps that involve eye movements and their planning, together with task dependency and interactions with memory.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, Supplemental Mathematical Appendix, and one figure and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.10.015.

AUTHOR CONTRIBUTIONS

J.Z.L., Q.L., F.A., W.A.F., and T.P. designed the experiments and analyses and wrote the paper. J.Z.L., Q.L., and F.A. conducted the experiments and performed the analyses.

ACKNOWLEDGMENTS

This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by National Science Foundation (NSF) STC award CCF-1231216. This research was also sponsored by grants from the NSF (NSF-0640097 and NSF-0827427), the National Eye Institute (R01 EY021594-01A1), and AFOSR-THRL (FA8650-05-C-7262). T.P. is supported by the Eugene McDermott chair. W.A.F. is a New York Stem Cell Foundation-Robertson Investigator.

Received: June 13, 2016
Revised: August 30, 2016
Accepted: October 10, 2016
Published: December 1, 2016

REFERENCES

1. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.
2. Anselmi, F., Leibo, J.Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T. (2016). Unsupervised learning of invariant representations. Theor. Comput. Sci. *633*, 112–121.
3. Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA *104*, 6424–6429.
4. Bart, E., and Ullman, S. (2008). Class-based feature matching across unrestricted transformations. IEEE Trans. Pattern Anal. Mach. Intell. *30*, 1618–1631.
5. Rolls, E.T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. Front. Comput. Neurosci. *6*, 35.
6. Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (NIPS 2012), 1106–1114.
7. Freiwald, W.A., and Tsao, D.Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science *330*, 845–851.
8. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B. (2003). Faces and objects in macaque cerebral cortex. Nat. Neurosci. *6*, 989–995.
9. Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674.
10. Ku, S.P., Tolias, A.S., Logothetis, N.K., and Goense, J. (2011). fMRI of the face-processing network in the ventral temporal lobe of awake and anesthetized macaques. Neuron *70*, 352–362.
11. Afraz, A., Boyden, E.S., and DiCarlo, J.J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. Proc. Natl. Acad. Sci. USA *112*, 6730–6735.
12. Moeller, S., Freiwald, W.A., and Tsao, D.Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. Science *320*, 1355–1359.
13. Tsao, D.Y., Moeller, S., and Freiwald, W.A. (2008). Comparing face patch systems in macaques and humans. Proc. Natl. Acad. Sci. USA *105*, 19514–19519.
14. Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.
15. Poggio, T., and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. Nature *343*, 263–266.
16. Bart, E., Byvatov, E., and Ullman, S. (2004). View-invariant recognition using corresponding object fragments. In Computer Vision – ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 2004 Proceedings, Part II (Springer), pp. 152–165.
17. Leibo, J.Z., Liao, Q., Anselmi, F., and Poggio, T. (2015). The invariance hypothesis implies domain-specific regions in visual cortex. PLoS Comput. Biol. *11*, e1004390.
18. Farzmahdi, A., Rajaei, K., Ghodrati, M., Ebrahimpour, R., and Khaligh-Razavi, S.M. (2016). A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. Sci. Rep. *6*, 25025.
19. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866.
20. Meyers, E.M., Borzello, M., Freiwald, W.A., and Tsao, D. (2015). Intelligent information loss: the coding of facial identity, head pose, and non-face information in the macaque face patch system. J. Neurosci. *35*, 7069–7081.
21. Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. Nature *381*, 520–522.
22. Isik, L., Meyers, E.M., Leibo, J.Z., and Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. J. Neurophysiol. *111*, 91–102.
23. Bruce, V., and Young, A. (1986). Understanding face recognition. Br. J. Psychol. *77*, 305–327.
24. Tan, C., and Poggio, T. (2016). Neural tuning size in a model of primate visual processing accounts for three key markers of holistic face processing. PLoS ONE *11*, e0150980.
25. Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A *2*, 284–299.
26. Isik, L., Leibo, J.Z., and Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. Front. Comput. Neurosci. *6*, 37.
27. Hinton, G.E., and Becker, S. (1990). An unsupervised learning procedure that discovers surfaces in random-dot stereograms. In Proceedings of the IEEE/INNS International Joint Conference on Neural Networks *1*, 218–222.
28. Földiák, P. (1991). Learning invariance from transformation sequences. Neural Comput. *3*, 194–200.
29. Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput. *14*, 715–770.
30. Berkes, P., Turner, R.E., and Sahani, M. (2009). A structured model of video reproduces primary visual cortical organisation. PLoS Comput. Biol. *5*, e1000495.
31. Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature *335*, 817–820.
32. Wallis, G., and Bülthoff, H.H. (2001). Effects of temporal association on recognition memory. Proc. Natl. Acad. Sci. USA *98*, 4800–4804.
33. Cox, D.D., Meier, P., Oertelt, N., and DiCarlo, J.J. (2005). 'Breaking' position-invariant object recognition. Nat. Neurosci. *8*, 1145–1147.
34. Li, N., and DiCarlo, J.J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. Neuron *67*, 1062–1075.
35. Anselmi, F., Rosasco, L., and Poggio, T. (2015). On invariance and selectivity in representation learning. arXiv, arXiv:1503.05938, http://arxiv.org/abs/1503.05938.
36. Hassoun, M.H. (1995). Fundamentals of Artificial Neural Networks (MIT Press).
37. Oja, E. (1982). A simplified neuron model as a principal component analyzer. J. Math. Biol. *15*, 267–273.
38. Oja, E. (1992). Principal components, minor components, and linear neural networks. Neural Netw. *5*, 927–935.
39. Abbott, L.F., and Nelson, S.B. (2000). Synaptic plasticity: taming the beast. Nat. Neurosci. *3*, 1178–1183.
40. Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw. *2*, 459–473.
41. Keck, T., Keller, G.B., Jacobsen, R.I., Eysel, U.T., Bonhoeffer, T., and Hübener, M. (2013). Synaptic scaling and homeostatic plasticity in the mouse visual cortex in vivo. Neuron *80*, 327–334.
42. Hengen, K.B., Lambo, M.E., Van Hooser, S.D., Katz, D.B., and Turrigiano, G.G. (2013). Firing rate homeostasis in visual cortex of freely behaving rodents. Neuron *80*, 335–342.
43. Turrigiano, G.G., and Nelson, S.B. (2004). Homeostatic plasticity in the developing nervous system. Nat. Rev. Neurosci. *5*, 97–107.
44. Yildirim, I., Kulkarni, T.D., Freiwald, W.A., and Tenenbaum, J.B. (2015). Efficient and robust analysis-by-synthesis in vision: a computational framework, behavioral tests, and modeling neuronal representations. In Proceedings of the Annual Conference of the Cognitive Science Society.