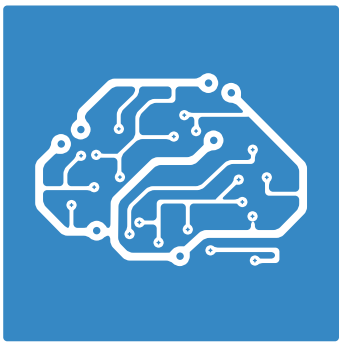


Center for Brains, Minds & Machines

CBMM Memo No. 049 **June 3, 2016**

View-tolerant face recognition and Hebbian learning imply mirror-symmetric neural tuning to head orientation

by

Joel Z. Leibo[1], Qianli Liao[1], Winrich Freiwald[1,2], Fabio Anselmi[1,3], Tomaso Poggio[1]
1: Center for Brains, Minds, and Machines and McGovern Institute for Brain Research at MIT, Cambridge, MA, USA
2: Laboratory of Neural Systems, The Rockefeller University, New York, NY, USA
3: Istituto Italiano di Tecnologia, Genova, Italy

Abstract: 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 relatively robust against identity-preserving transformations like depth-rotations [33, 32, 23, 13]. 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 [46, 8, 44, 29]. While simulations of these models recapitulate the ventral stream's progression from early view-specific to late view-tolerant representations, they fail to generate the most salient property of the intermediate representation for faces found in the brain: mirror-symmetric tuning of the neural population to head orientation [16]. Here we prove that a class of hierarchical architectures and a broad set of biologically plausible learning rules can provide approximate invariance at the top level of the network. While most of the learning rules do not yield mirror-symmetry in the mid-level representations, we characterize a specific biologically-plausible Hebb-type learning rule that is guaranteed to generate mirror-symmetric tuning to faces tuning at intermediate levels of the architecture.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF - 1231216.

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 [52, 53, 28, 2]. Face patches are selectively interconnected to form a face-processing network [36]. Face patches are arranged along an occipito-temporal axis (from the middle lateral (ML) and middle fundus (MF) patches, through the antero-lateral face patch (AL), and culminating in the antero-medial (AM) patch [51] (Fig. 1-A)) along which response latencies increase systematically from ML/MF via AL to AM, suggesting sequential forward-processing [16].

Face patches differ qualitatively in how they represent identity across head orientations [16]. Neurons in the ML/MF patches are view-specific, while neurons in AM approach view-invariance. Furthermore, spatial position and size invariance increase from ML/MF to AL, and further to AM [16]. These properties of the face-processing network replicate the general trend of the ventral stream as summarized in [32, 43, 13] and conform to the concept of a feedforward processing hierarchy.

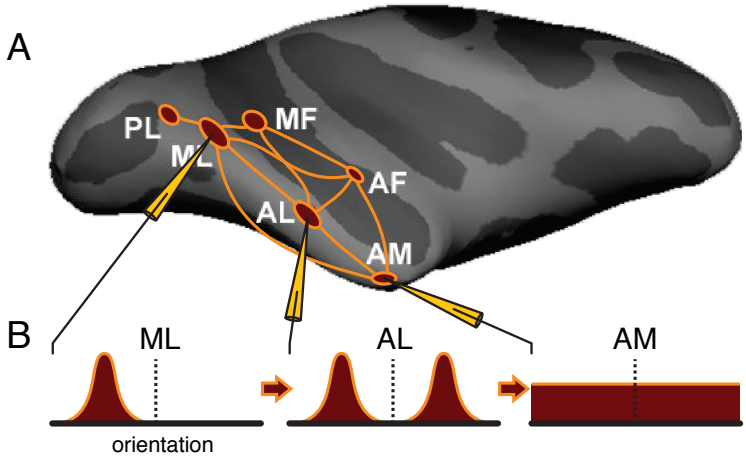


Figure 1: Schematic of the macaque face-patch system [36, 54, 16]. (A) Side view of computer-inflated macaque cortex with six areas of face-selective cortex (red) in the temporal lobe together with connectivity graph (orange). Face areas are named based on their anatomical location: PL, posterior lateral; ML; middle lateral; MF, middle fundus; AL, anterior lateral; AF, anterior fundus ; AM, anterior medial (3), and have been found to be directly connected to each other to form a face-processing network [36]. Recordings from three face areas, ML, AL, 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 to 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 [16].

Several hierarchical models of object recognition [17, 40, 43, 9] and face recognition [8, 29, 14] feature a progression from view-specific early processing stages to view-invariant later processing stages similar to ML/MF and AM, respectively. Simulations have shown that view-based models can achieve an AM-like representation by successively pooling the responses of view-tuned units like those found in the early processing stage ML/MF. The theoretical underpinnings of this property are described in the Appendix (section 1.1).

Neurons in the intermediate face area AL, but not in preceding areas ML/MF, exhibit mirror-symmetric head orientation tuning [16]. That is, an AL neuron tuned to one profile view of the head typically responds similarly to the opposite profile, but not to the front view (Fig. 1-B). This phenomenon is not predicted by simulations of classical and current view-based computational models of the ventral stream

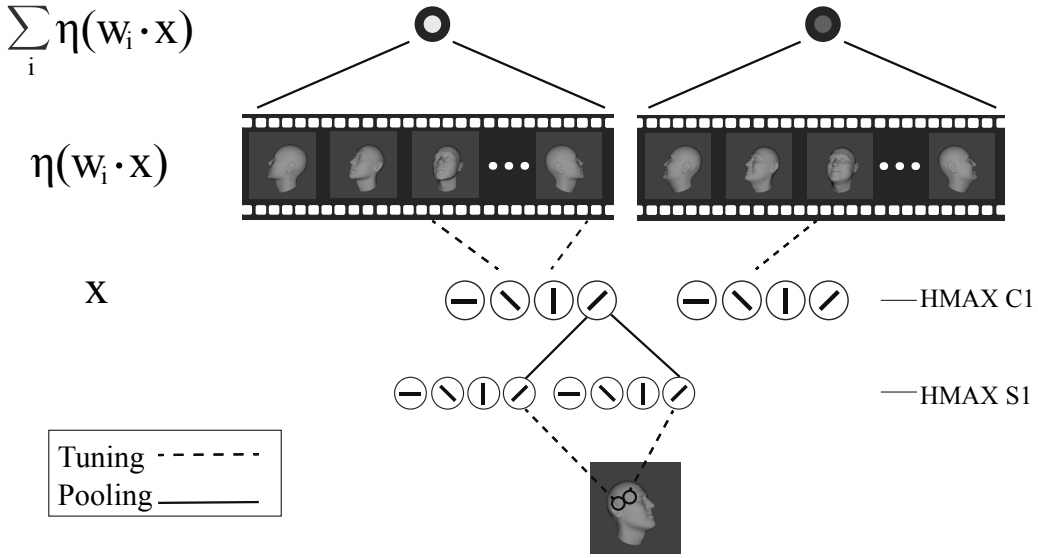


Figure 2: Illustration of the model. Inputs are encoded in HMAX C1 [43], then projected onto w_i. In the view-based model the w_i represent faces at specific views. In the Oja-model, the w_i are principal components. Units in the output layer pool over all the units in the previous layer corresponding to projections onto the same template individual's views (view-based model) or PCs (Oja-model).

[59]. In this paper we ask why the primate brain may compute a mirror symmetric representation as a necessary intermediate step towards invariant face-representation and what this tells us about the brain's mechanisms of learning.

Results

Assumptions underlying the model

We consider a feedforward face-processing hierarchy as a model for how the ventral stream rapidly computes invariant representations. Invariant information can be decoded from inferotemporal cortex, and the face areas within it, roughly 100ms after stimulus presentation [23, 34]. This is too fast of a timescale for feedback to play a large role [23, 50, 27]. Thus while the actual face processing system might operate in other modes as well, all indications are that fundamental properties of shape-selectivity and invariance need to be explained as a property of feedforward processing.

The population of neurons in ML/MF is highly face selective [53] and incoming information can be thought of as passing through a face-likeness filter. We thus assume the existence of a functional gate that routes only images of face-like objects at the input of the face system. The existence of large "face-like" templates or filters explains many of the so-called holistic effects of face perception, including face inversion and the composite face [61] effect [49, 14]. This property has one further computational implication: it provides an automatic face-specific gating mechanism to the face-processing system.

We make the standard assumption that a neuron's basic operation is a pooled dot product between inputs x and synaptic weight vectors $\{w_i\}$, yielding complex-like cells as

$$\mu^k(x) = \frac{1}{|G|} \sum_{i=1}^{|G|} \eta(\left\langle x, g_i w^k \right\rangle) \tag{1}$$

where $\eta : \mathbb{R} \to \mathbb{R}$ is a nonlinear function e.g., squaring as in [1]. We suppose that $g_i \in G$ are image plane transformations corresponding to rotations in depth of the face. Note that G is a set of transformations but it is not a group (see Appendix section 1.1). We call $\vec{\mu}(x) \in \mathbb{R}^K$ the signature of image x.

Approximate view invariance

The model of Eq. (1) encodes a novel face by its similarity to a set of stored template faces. For example, the $g_i w^k$ could correspond to views i of each of a set of well-known individuals k from an early developmental period e.g., parents, caretakers, etc. One could regard the acquisition of this set of familiar faces as the algorithm's (unsupervised) training phase. To see why the algorithm works, consider that whenever w_i^k encodes a non-matching orientation to I, the value of $\langle x, g_i w^k \rangle$ will be very low. Among the w^k tuned to the correct orientation, there will be a range of response values since different template faces will have different levels of similarity to I. When the novel face appears at a different orientation, the only effect is to change which specific view-tuned units carry its signature. Since the pooled neural response is computed by summing over these, the large responses carrying the signature will dominate. Thus the pooled neural response will be approximatively unchanged by rotation (see the Appendix section 1). Since these models are based on stored associations of frames, they 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].

The computational insight enabling depth-rotation tolerant representations to be learned from experience is that, due to properties of how objects move in the world, temporally adjacent frames (the $g_i w^k$) almost always depict the same object [22, 48, 15, 58, 10, 26]. 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 temporal-association strategy of this sort [35, 56, 12, 30, 57, 31]. 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 depth-rotation tolerance because the set of rotations in depth approximates the group structure of affine transformations in the plane (see Appendix section 1). For the latter case, there are theorems guaranteeing invariance without loss of selectivity by operations resembling the convolution in space performed by simple cells and the pooling done by complex cells [5].

Furthermore, [29] showed that Eq. 1 is approximately invariant to rotations in depth for x a face, provided the templates w^k also correspond to images of faces. For each template w^k, the rotated views $\{g_i w^k, \ i = 1, \cdots, |G|\}$ must have been observed and stored. The $\eta(\langle x, g_i w^k \rangle)$ can be interpreted as the output of "simple" cells each with tuning $g_i w^k$ when stimulated with image x. In a similar way $\mu^k(x)$ can be interpreted as the activity of the "complex" cell indexed by k.

Biologically plausible learning

The simple-complex algorithm described above can provide an invariant representation but relies on a biologically implausible learning step: storing a set of discrete views observed during development. Instead we propose a more biologically plausible mechanism: Hebb-like learning [21] at the level of simple cells (see Equation (2)). Instead of storing separate frames, cortical neurons exposed to the rotation in depth of a face update their synaptic weights according to a Hebb-like rule, effectively becoming each tuned to one of a set of basis functions corresponding to different combinations of the set of views. Different Hebb-like rules lead to different sets of basis functions such as Independent Components (IC) or Principal Components (PC). Since each of the neurons become tuned to one of these basis functions instead of one of the views, a set of basis functions replaces the $g_i w^k$ (for a given k) in the pooling Equation (1). The question is whether invariance is still present under this new tuning.

The surprising answer is that most unsupervised learning rules will learn approximate invariance to viewpoint when provided with the appropriate training set (see Appendix section 2 for a proof). In fact, unsupervised Hebb-like plasticity rules such as Oja's, Foldiak's trace rule, and ICA provide a basis that when used in the pooling equation provide invariance. Supervised learning rules such as backpropagation also satisfy the requirement as long as the training set is appropriate.

In the following we consider as an example a simple Hebbian learning scheme called Oja's rule [37, 38]. At this point we are concerned only with establishing the model and why it computes a view-tolerant face representation. For this purpose we could use any of the other learning rules—like Foldiak's trace rule or ICA—but we focus on the Oja rule because it will turn out to be of singular relevance for mirror symmetry.

The Oja rule can be derived as the first order expansion of a normalized Hebb rule. The assumption of this normalization is plausible, because normalization mechanisms are widespread in cortex [55].

For learning rate α, Oja's rule is

$$\Delta w = \alpha(xy - y^2 w) = \alpha(xx^\intercal w - (w^\intercal x x^\intercal w)w)\,. \tag{2}$$

The original paper of Oja showed that the weights of a neuron updated according to this rule will converge to the top principal component (PC) of the neuron's past inputs, that is to an eigenvector of the input's covariance C. 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 [45, 38]. This generalized Oja rule can be regarded as an online algorithm to compute the principal components of incoming stream of vectors, in our case, images.

What is learned and how it is stored depends on the choice of a timescale over which learning takes place since learning is dictated by the underlying covariance C of the inputs (see Appendix, section 3). In order for familiar faces to be stored so that the neural response modeled by Eq. (1) tolerates rotations in depth of novel faces, we propose that Oja-type plasticity leads to representations for which the w_i^k are given by principal components (PCs) of an image sequence depicting depth-rotation of face k. Consider an immature functional unit exposed, while in a plastic state, to all depth-rotations of a face. Oja's rule will converge to the eigenvectors corresponding to the top r eigenvalues and thus to the subspace spanned by them. The Appendix, section 2 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 Eq. (1) with $g_i w^k$ replaced by the i-th PC. The 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 (see Appendix section 2).

Empirical evaluation of view-invariant face recognition performance

View-invariance of the two models was assessed by simulating a sequence of same-different pair-matching tasks, each demanding more invariance than the last. In each test, 600 pairs of face images were sampled from the set of faces with orientations in the current testing interval. 300 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 $[-r, r]$ for $r = 5°, \ldots, 95°$. The radius of the testing interval r, dubbed the invariance range, is the abscissa in Fig. 3.

To classify an image pair (a, b) as depicting the same or a different individual, the cosine similarity $(a \cdot b)/(\|a\|\|b\|)$ of the two representations was compared to a threshold. The threshold was varied systematically in order to compute the area under the ROC curve (AUC), reported on the ordinate of Fig. 3. 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 (Fig. 3).

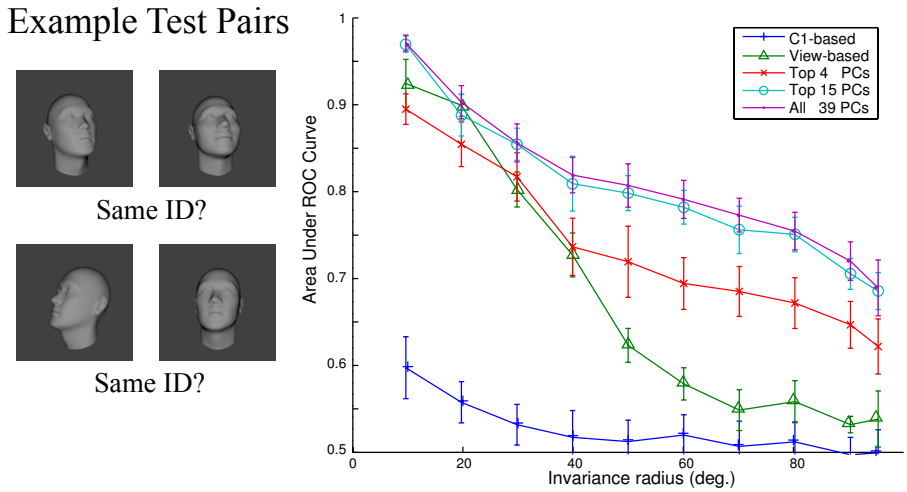


Figure 3: Model performance at the task of same-different pair matching as a function of the extent of depth rotations appearing in the test set (the invariance range of the task). All models were based on HMAX C1 features [46].

Mirror symmetry

Consider the 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 its orbit in $3D$ w.r.t. the rotation group:

$$O_x = (r_0 x, \cdots, r_N x).$$

where r is a rotation matrix in 3D, w.r.t., e.g., the z axis.

Projecting onto $2D$ we have

$$P(O_x) = (P(r_0 x), \cdots, P(r_N x)).$$

Note now that, due to the bilateral symmetry, the above set can be written as:

$$P(O_x) = (x_0, \cdots, x_{\frac{N}{2}}, Rx_1, \cdots, Rx_{\frac{N}{2}}).$$

where $x_n = Pr_n x$, $n = 1, \cdots, N/2$ and R is the reflection operator. Thus the set consists of a collection of orbits w.r.t. the group $G = \{e, R\}$ of the templates $\{x_1, \cdots, x_{N/2}\}$.

This property of the training set is used in the appendix in two ways. First, it 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 (sections 1 – 2).

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

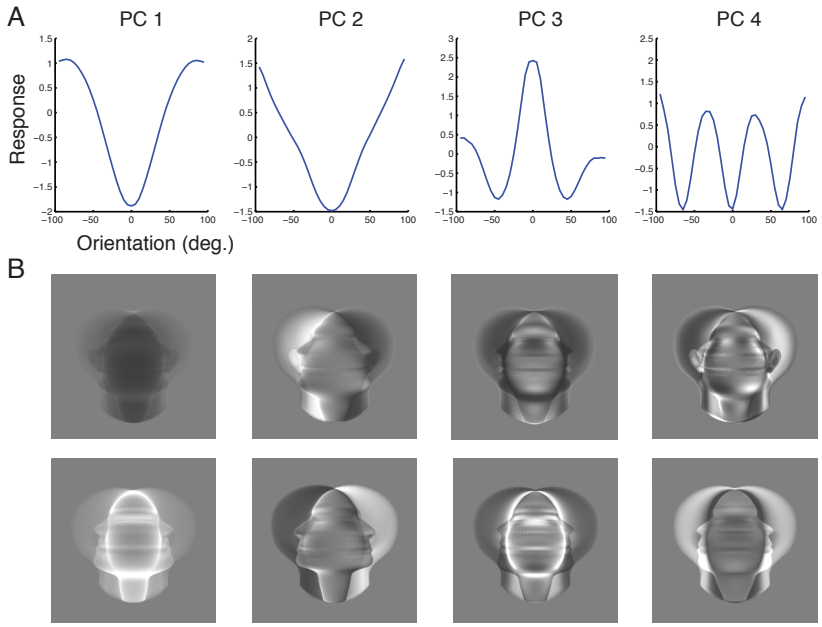


Figure 4: Mirror symmetric orientation tuning of the raw pixels-based model (A) $(w_i \cdot x_\theta)^2$ as a function of the orientation of x_θ. Here each curve represents a different PC. (B) Solutions to the Oja equation (w_i) visualized as images. They are either symmetric or antisymmetric about the vertical midline.

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 [1]. 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. Fig. 4-A 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. Fig. 5-A shows how the three populations in the C1-based model represent face view and identity and Fig. 5-B shows the same for populations of neurons recorded in ML/MF, AL, and AM. The model is the same one as in Fig. 3.

In contrast to the Oja/PCA case, we show through a simulation analogous to Fig. 5 that ICA does not yield mirror symmetric tuning curves (appendix section 4). Though this is an empirical finding for a specific form of ICA, we do not expect, based on our proof technique for the Oja case, that a generic learning rule would predict mirror symmetric tuning curves.

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. Each component can correspond to a single face or to a set of faces, different for each component of the signature. 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.

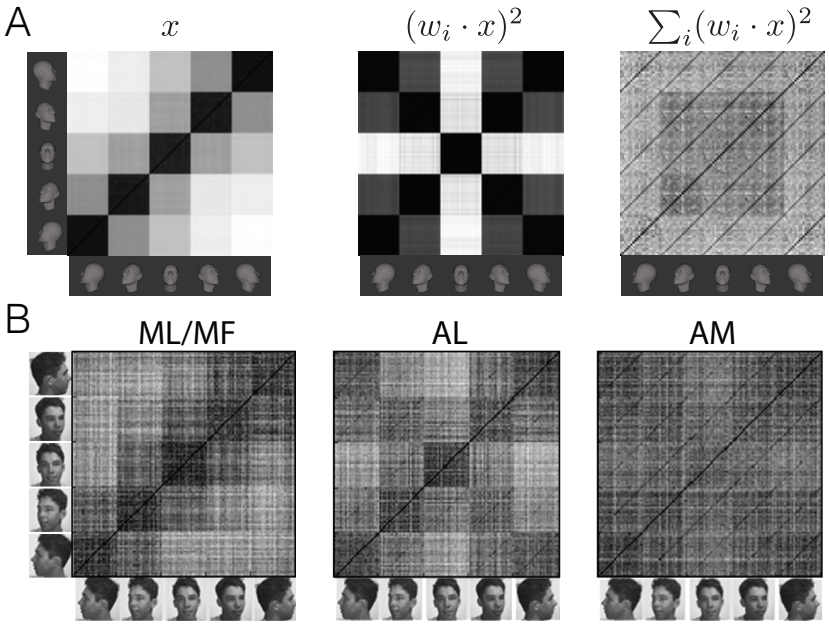


Figure 5: Population representations of face view and identity (A) Model population similarity matrices (B) Neural population similarity matrices from [16].

Discussion

The model discussed here provides a computational account of how experience and evolution may wire up the ventral stream circuitry to achieve the computational goal of view-invariant face recognition. Neurons in top-level face patch AM maintain an explicit representation selective for face identity and tolerant to position, scale, and viewing angle [16] (along with other units tolerant to identity but selective for other variables such as viewing angle). The approach in this paper explains how this property may arise in a feed-forward hierarchy. To the best of our knowledge, it is the first account that provides a computational explanation of why cells in the face network's penultimate processing stage, AL, are tuned symmetrically to head orientation.

Our assumptions about the architecture for invariance conform to i-theory [4, 5] which is a theory of invariant recognition that characterizes and generalizes the convolutional and pooling layers in deep networks. i-theory has recently been shown to predict domain-specific regions in cortex [29] with the function of achieving invariance to class-specific transformations (e.g. for faces) and the specific form of eccentricity-dependent cortical magnification [42]. Our assumption of Hebbian-like plasticity for learning template views is, however, outside the mathematics of i-theory: it links it to biological properties of cortical synapses.

This argument of this paper has been made, as nearly as possible, from first principles. It begins with a claim about the computational problem faced by a part of the brain: the need to compute view-tolerant representations for faces. Yet it seeks to explain properties of single neurons in a specific brain region, AL, far from the sensory periphery. The argument proceeds by considering which of the various biologically-plausible learning rules satisfy requirements coming from the theory while also yielding non-trivial predictions for AL neurons in qualitative accord with the available data. It seems significant then that the argument only works in the case of Oja-like plasticity; it may suggest the hypothesis that such plasticity may indeed be driving learning in AL.

The class of learning rules yielding invariance includes those that emerge from principles such as sparsity and the efficient coding hypothesis [6, 7, 39]. However, explaining the mirror symmetric tuning of AL

neurons apparently requires the Oja rule. An interesting direction for future work in this area could be to investigate the role of sparsity in the face processing system. Perhaps a learning algorithm derived from the efficient coding perspective that also explains AL's mirror symmetry could be found.

Our model is designed to account only for the feed-forward processing in the ventral stream. Back-projections between visual areas—and of course within each area—are well known to exist in the ventral stream and probably also exist in the face patch network. They are likely to play a major role in visual recognition after ~ 80 ms from image onset. Representations computed in the first feeforward sweep are likely used to provide information about a few basic questions such as the identity or pose of a face. Additional processing is likely to require iterations and even top-down computations involving shifts of fixation and generative models. An example for face recognition is recent work [60] which combines a feedforward network like ours—also showing mirror-symmetric tuning of cell populations—with a probabilistic generative model. Thus our feedforward model, which succeeds in explaining the main tuning and invariance properties of the macaque face-processing system, may serve as a building block for future object-recognition models addressing brain areas such as prefrontal cortex, hippocampus and superior colliculus, integrating feed-forward processing with subsequent computational steps that involve eye-movements and their planning, together with task dependency and interactions with memory.

Materials

Stimuli

40 face models were rendered with perspective projection. Each face was rendered (using Blender [11]) at each orientation in 5° increments from -95° to 95°. The untextured face models were generated using Facegen [47]. All faces appeared on a uniform gray background.

View-invariant Same-different Pair Matching Task

For each of the 5 repetitions of the same-different pair matching task, 20 template and 20 test individuals were randomly selected from the full set of 40 individuals. The template and test sets were chosen independently and were always disjoint. 50% of the 600 test pairs sampled from each testing interval depicted the same two individuals. Each testing interval was symmetric about 0° (frontal) and testing intervals were ordered by inclusion. The smallest was $[-10^\circ, 10^\circ]$ and the largest was $[-95^\circ, 95^\circ]$ (left and right profile views). The classifier compared the Cosine similarity of the two zero-mean, and unit-standard deviation representations to a threshold. The threshold was integrated over to compute the area under the ROC curve (AUC). The abscissa of Fig. 3 is the radius of the testing interval from which test pairs were sampled. The ordinate of Fig. 3 is the mean AUC $\pm$ the standard deviation computed over the 5 repetitions of the experiment.

A similarity matrix in Figure 5 was obtained by computing Pearson's linear correlation coefficient between each test sample pair. The same matrix was computed 10 times with different training/test splits and the average was reported. Same procedures were repeated for features from area MLMF, AL and AM to get corresponding matrices.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFOSR-THRL (FA8650-05-C-7262). Additional support was provided by the Eugene McDermott Foundation.

References

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985. ISSN 1084-7529. URL http://www.opticsinfobase.org/abstract.cfm?URI=josaa-2-2-284.

[2] Arash Afraz, Edward S Boyden, and James J DiCarlo. Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, 112(21):6730–6735, 2015.

[3] Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv:1311.4158*, 2013.

[4] Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations. *Theoretical Computer Science*, 2015. doi: http://dx.doi.org/10.1016/j.tcs.2015.06.048.

[5] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. *arXiv preprint arXiv:1503.05938*, 2015.

[6] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[7] Horace B Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communication*, pages 217–234, 1961.

[8] E. Bart and S. Ullman. Class-based feature matching across unrestricted transformations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1618–1631, 2008. ISSN 0162-8828. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4378342.

[9] E Bart, E Byvatov, and S Ullman. View-invariant recognition using corresponding object fragments. In *European Conference on Computer Vision (ECCV)*, volume 3024, pages 152–165, Prague, Czech Republic, 2004. Springer. URL http://www.springerlink.com/index/GGBDRQ3WQFGB9LDN.pdf.

[10] Pietro Berkes, Richard E. Turner, and Maneesh Sahani. A structured model of video reproduces primary visual cortical organisation. *PLoS Computational Biology*, 5(9):e1000495, 10 2009. doi: 10.1371/journal.pcbi.1000495.

[11] Blender.org. Blender 2.6, 2013.

[12] DD Cox, P Meier, N Oertelt, and James J. DiCarlo. 'Breaking' position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–1147, 2005. URL http://www.nature.com/neuro/journal/v8/n9/abs/nn1519.html.

[13] James J. DiCarlo, D Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. URL http://www.sciencedirect.com/science/article/pii/S089662731200092X.

[14] Amirhossein Farzmahdi, Karim Rajaei, Masoud Ghodrati, Reza Ebrahimpour, and Seyed-Mahdi Khaligh-Razavi. A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Scientific Reports*, 6, 2016.

[15] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. URL http://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.2.194.

[16] Winrich A. Freiwald and D.Y. Tsao. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, 330(6005):845, 2010. ISSN 0036-8075. URL http://www.sciencemag.org/cgi/content/abstract/330/6005/845.

[17] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. ISSN 0340-1200. doi: 10.1007/BF00344251. URL http://www.springerlink.com/content/r6g5w3tt54528137.

[18] M. Golubitsky and I. Stewart. The symmetry perspective; from equilibrium to chaos in phase space and physical space. 2002.

[19] Singer Hardt, Recht. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv:1509.01240*, 2016.

[20] M. H. Hassoun. Fundamentals of artificial neural networks. *MIT Press*, 1995.

[21] D. O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949. URL http://books.google.com/books?hl=en&lr=&id=gUtwMochAI8C&oi=fnd&pg=PP1&dq=Hebb&ots=w1kQ2jqppz&sig=QmaxGp399apRC1HQccm_nu9WnU8.

[22] Geoffrey E Hinton and Suzanna Becker. An unsupervised learning procedure that discovers surfaces in random-dot stereograms. In *Proceedings of the International Joint Conference on Neural Networks, Washington DC*, pages 218–222, 1990.

[23] Chou P. Hung, Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866, November 2005. doi: 10.1126/science.1117593.

[24] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

[25] Aapo Hyvärinen and Erkki Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[26] Leyla Isik, Joel Z. Leibo, and Tomaso Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. *Frontiers in Computational Neuroscience*, 6(37), 2012. doi: 10.3389/fncom.2012.00037. URL http://www.frontiersin.org/Computational_Neuroscience/10.3389/fncom.2012.00037/abstract.

[27] Leyla Isik, Ethan M Meyers, Joel Z Leibo, and Tomaso Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, 111(1):91–102, 2014.

[28] S.P. Ku, A.S. Tolias, N.K. Logothetis, and J. Goense. fMRI of the Face-Processing Network in the Ventral Temporal Lobe of Awake and Anesthetized Macaques. *Neuron*, 70(2):352–362, 2011. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627311002054.

[29] Joel Z. Leibo, Qianli Liao, Fabio Anselmi, and Tomaso Poggio. The invariance hypothesis implies domain-specific regions in visual cortex. *PLoS Computational Biology*, 11(10):e1004390, 10 2015. doi: 10.1371/journal.pcbi.1004390.

[30] Nuo Li and James J DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–7, September 2008. ISSN 1095-9203. doi: 10.1126/science.1160028. URL http://www.sciencemag.org/cgi/content/abstract/321/5895/1502.

[31] Nuo Li and James J DiCarlo. Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. *Neuron*, 67(6):1062–1075, 2010. URL http://www.cell.com/neuron/fulltext/S0896-6273(10)00639-2.

[32] NK Logothetis and DL Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19(1): 577–621, 1996. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.ne.19.030196.003045.

[33] NK Logothetis, J Pauls, and T Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563, 1995. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982295001084.

[34] Ethan M Meyers, Mia Borzello, Winrich A Freiwald, and Doris Tsao. Intelligent information loss: The coding of facial identity, head pose, and non-face information in the macaque face patch system. *The Journal of Neuroscience*, 35(18):7069–7081, 2015.

[35] Y. Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988. URL http://hebb.mit.edu/courses/9.641/readings/Miyashita88.pdf.

[36] S. Moeller, Winrich A. Freiwald, and D.Y. Tsao. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, 320(5881):1355, 2008. URL http://www.sciencemag.org/content/320/5881/1355.short.

[37] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. URL http://www.springerlink.com/index/u9u6120r003825u1.pdf.

[38] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992. URL http://www.sciencedirect.com/science/article/pii/S0893608005800899.

[39] Bruno A Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. URL http://redwood.psych.cornell.edu/papers/olshausen_field_nature_1996.pdf.

[40] T Poggio and S Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990. URL http://cbcl.mit.edu/people/poggio-new/journals/poggio-edelman-nature-1990.pdf.

[41] T. Poggio, T. Vetter, and H. Bulthoff. 3D Object Recognition: Symmetry and Virtual Views, 1992. URL http://www.stormingmedia.us/83/8379/A837952.pdf.

[42] Tomaso Poggio, Jim Mutch, and Leyla Isik. Computational role of eccentricity dependent cortical magnification. *CBMM Memo No. 017. arXiv preprint arXiv:1406.1770*, 2014.

[43] M Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999. ISSN 1097-6256. doi: 10.1038/14819.

[44] ET Rolls. Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience*, 6, 2012.

[45] T.D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989. URL http://www.sciencedirect.com/science/article/pii/0893608089900440.

[46] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007. URL http://portal.acm.org/citation.cfm?id=1263421&dl=.

[47] Singular Inversions. FaceGen Modeller 3, 2003.

[48] Michael P. Stryker. Temporal associations. *Nature*, (6349):108–109, 1991. URL http://www.nature.com/nature/journal/v354/n6349/abs/354108d0.html.

[49] C. Tan and T. Poggio. Neural tuning size in a model of primate visual processing accounts for three key markers of holistic face processing. *Public Library of Science | PLoS ONE*, 1(3): e0150980, 2016.

[50] S Thorpe, D Fize, and C Marlot. Speed of processing in the human visual system. *Nature*, 381 (6582):520–522, 1996.

[51] Doris Y Tsao, Sebastian Moeller, and Winrich A Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49):19514–19519, 2008.

[52] D.Y. Tsao, Winrich A. Freiwald, T.A. Knutsen, J.B. Mandeville, and R.B.H. Tootell. Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995, 2003. URL http://www.nature.com/neuro/journal/v6/n9/abs/nn1111.html.

[53] D.Y. Tsao, Winrich A. Freiwald, R.B.H. Tootell, and M.S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670, 2006. URL http://www.sciencemag.org/content/311/5761/670.short.

[54] D.Y. Tsao, S. Moeller, and Winrich A. Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49):19514, 2008. URL http://www.pnas.org/content/105/49/19514.short.

[55] Gina G. Turrigiano and Sacha B. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2):97–107, 2004. URL http://www.nature.com/nrn/journal/v5/n2/abs/nrn1327.html.

[56] G Wallis and H H Bülthoff. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4800–4, April 2001. ISSN 0027-8424. doi: 10.1073/pnas.071028598. URL http://www.pnas.org/cgi/content/abstract/98/8/4800.

[57] G. Wallis, B.T. Backus, M. Langer, G. Huebner, and H. Bülthoff. Learning illumination-and orientation-invariant representations of objects through temporal association. *Journal of vision*, 9(7), 2009.

[58] L. Wiskott and T.J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. URL http://www.mitpressjournals.org/doi/abs/10.1162/089976602317318938.

[59] D.L.K. Yamins and J.D. Dicarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19,3:356–365, 2016.

[60] I. Yildirim, T. Kulkarni, and J. B. Freiwald, W.and Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. *Annual Conference of the Cognitive Science Society*, 2015.

[61] Andrew W Young, Deborah Hellawell, and Dennis C Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

Appendix

The key results in this appendix can be informally stated as follows:

- We prove than a number of learning rules, supervised and unsupervised, are equivariant with respect to the symmetries of the training data. We use this result in the case of training data consisting of images of faces for all view angles obtaining equivariance of the solutions of the learning rules with respect to the reflection group and the group of rotations. The implications that we use in the paper are
  - the solutions of all learning rules can be used as templates in the computation of an invariant signature. The algorithm consists of performing dot products of the input image with each template, transforming nonlinearly (for instance using a rectifier nonlinearity or a square) the result and then pooling over *all* templates, i.e., the solutions of the learning rule. The result is approximately invariant to rotation in depth.
  - in the case of the Oja rule we prove that the solutions are even or odd functions of the view angle; a square nonlinearity provides even functions, which are mirror-symmetric. We were not able to prove such a property for any of the other learning rules.
- in the case of the ICA rule we show empirical evidence that the solutions are neither odd nor even. This suggests that most learning rules do not lead to even or odd solutions.

The appendix is divided into four sections:

1. In section A we show how recent theorems on invariance under group transformations could be extended to nongroups and under which conditions. We show how an *approximately invariant* signature can be computed in this setting. In particular we analyze the case of rotation in depth and mirror symmetry transformations of bilateral symmetric objects such as faces.
2. In section B we describe how the group symmetry properties of the set of images to which neurons are exposed (the "unsupervised" training set) determine the symmetries of the learned weights. In particular we show how the weight symmetries gives a simple way of computing an invariant signature.
3. In section C we prove that the solutions of the Oja equation, given that the input vectors that are reflections of each other (like a face's view at θ degrees and its view at $-\theta$ degrees), must be odd or even.
4. In section D we provide empirical evidence that there are solutions of ICA algorithms—on the same data as above—that do not show any symmetry.

In the following we indicate with $x \in R^d$ an image, with $w \in R^d$ a filter or neural weight and with G a locally compact group.

A Approximate Invariance for non-group transformations

In this section we analyze the problem of getting an approximately invariant signature for image transformations that do not have a group structure. In fact, clearly, not all image transformations have a group structure. However assuming that the object transformation defines a smooth manifold we have (by the

theory of Lie manifolds) that locally a Lie group is defined by the generators on the tangent space. We illustrate this in a simple example. Let $x \in \mathbb{R}^d$. Let $s : \mathbb{R}^d \times \mathbb{R}^Q \to \mathbb{R}^d$ a C^∞ transformation depending on $\Theta = (\theta_1, \cdots, \theta_Q)$ parameters. For any fixed $x \in \mathbb{R}^d$ the set $M = (s(x, \Theta),\ \Theta \in \mathbb{R}^Q)$ describe a differentiable manifold. If we expand the transformation around e.g. $\vec{0}$ we have:

$$s(x, \Theta) = s(x, \vec{0}) + \sum_{i=1}^{Q} \frac{\partial s(x, \Theta)}{\partial \theta_i} \theta_i + o(\|\Theta\|^2) = x + \sum_{i=1}^{Q} \theta_i L_{\theta_i}(x) + o(\|\Theta\|^2) \tag{3}$$

where L_{θ_i} are the infinitesimal generators of the transformation in the i^{th} direction.
Therefore locally (when the term $o(\|\Theta\|^2)$ can be neglected) the associated group transformation can be expressed by exponentiation as:

$$g(\Theta) = \exp(\theta_1 L_{\theta_1} + \theta_2 L_{\theta_2} + \cdots + \theta_Q L_{\theta_Q}).$$

Note that the above expansion is valid only locally. In other words instead of a global group structure of the transformation we will have a collection of local transformations that obey a group structure. The results derived in section B will then say that the local learned weights will be orbits w.r.t. the local group approximating the non-group global transformation.

A.1 Invariance under rotations in depth

The 3D "views" of an object undergoing a 3D rotation are group transformations but the 2D projections of an object undergoing a 3D rotation are not group transformations. However for any fixed angle θ_0 and for small rotations the projected images approximately follow a group structure. This can be easily seen making the substitution in eq. (3) $s(x, \Theta) = P(r_\theta x)$ where P is the 2D projection. Let $\eta : \mathbb{R} \to \mathbb{R}$ be a nonlinear function, e.g., squaring or rectification. For small values of θ we have therefore that the signature:

$$\mu_w(x) = \int_{-\theta_0}^{\theta_0} d\theta\ \eta(\langle Px, Pr_\theta w \rangle)$$

or its discrete version

$$\mu_w(x) = \sum_i\ \eta(\langle Px, Pr_{\theta_i} w \rangle) = \sum_i\ \eta(\langle Px, g(\theta_i) Pw \rangle)$$

is invariant under 3D rotation of x of an angle $\bar{\theta}$ up to a factor proportional to $O(\|\bar{\theta}\|)$. Alternatively if the following property holds:

$$\langle Px, Pr_\theta w \rangle = 0 \quad \theta > \bar{\theta} \tag{4}$$

the invariance will be exact (see [3, 29]); this is the case e.g. when both w and x are faces.

The locality of the group structure (eq. (4)) means that we have invariance of the signature only within each local neighborhood but not over all viewpoints. A reasonable scenario could be that each local neighborhood may consist of, say, ± 30 degrees (depending on the universe of distractors). Almost complete view invariance can be obtained from a single view at $+30$ degrees. In fact the view, together with the associated virtual view at -30 degrees because of mirror symmetry, provides invariance over $-60, +60$ degrees [41].

A.2 Rotation in depth and mirror symmetry.

As explained on the previous paragraph, projected rotations in depth are not group transformations. However in the case of a bilateral symmetric objects, as we will see below, projected rotations in depth are a collection of orbits of the mirror symmetry group. Section B will clarify why this property is important

proving that it forces the set of solutions of a variety of learning rules to be a collection of orbits w.r.t. the mirror symmetry group.
Consider e.g. a face, x, which is a bilateral symmetric object and its orbit in $3D$ w.r.t. the rotation group:

$$O_x = (r_0 x, \cdots, r_{2\pi} x).$$

where r is a rotation matrix in 3D, e.g. w.r.t. the z axis.
Projecting onto $2D$ we have

$$P(O_x) = (P(r_0 x), \cdots, P(r_{2\pi} x)).$$

Note now that, due to the bilateral symmetry, the above set can be written as:

$$P(O_x) = (x_0, \cdots, x_{\frac{N}{2}}, Rx_1, \cdots, Rx_{\frac{N}{2}}).$$

where $x_n = Pr_{\theta_n} x$, $n = 1, \cdots, N/2$ and R is the reflection operator. The set consists of a collection of orbits w.r.t. the group $G = \{e, R\}$. This is due to the relation

$$x_n = P(r_{\theta_n} x) = Rx_{\frac{N}{2}+n} = R(Pr_{-\theta} x).$$

i.e. a face rotated by an angle θ and then projected is equal to the reflection of the same face rotated by an angle $-\theta$ and projected.
The reasoning generalizes to multiple faces. In summary in the specific case of bilateral symmetric objects rotating in depth, a projection onto a plane parallel to the rotation axis creates images which are transformations w.r.t. the group of reflection, thus falling in the group case described in the above paragraphs.

B Unsupervised and supervised learning and data symmetries

In the following we show how symmetry properties on the neuronal inputs affect the learned weights. We model different unsupervised (Hebbian, Oja, Foldiak, ICA) or supervised learning (SGD) rules as dynamical systems coming from the requirement of minimization of some target function. We see how these dynamical systems are equivariant (in the sense specified below) and how equivariance determines the symmetry properties of their solutions.

This gives a simple way to generate an invariant signature by averaging over all solutions.

B.1 Equivariant dynamical systems and their solutions.

We make the general assumption that the dynamical system can be described in terms of trying to minimize a non-linear functional of the form:

$$\underset{w \in X}{\arg\min} \;\; \mathcal{L}(w, x), \quad \mathcal{L}(w, x) = h(w, x), \quad x, w \in \mathbb{R}^d \tag{5}$$

The associated dynamical system reads as:

$$\dot{w} = f(w) = \dot{h}(w, x). \tag{6}$$

A general result holds for equivariant dynamical systems. A dynamical system is called *equivariant* w.r.t. a group G if f in eq. (6) commutes with any transformation $g \in G$ i.e.

$$f(gw) = gf(w), \quad \forall g \in G. \tag{7}$$

In this case we have:

Theorem 1. *If an equivariant dynamical system has a solution w, then the whole group orbit of w will also be a set of solutions (see [18]).*

In the following we are going to analyze different cases of updating rules for neuronal weights showing, under the hypothesis that the training set is a (scrambled) collection of the orbits i.e. we specialize the set X to be of the form:

$$X = G\mathcal{T}, \quad \mathcal{T} \in \mathbb{R}^{d\times N}, \quad X = \{x_1, \cdots, x_N\}, \tag{8}$$

that the dynamical system is equivariant.
We will see that the following variant of the equivariance holds for many dynamical systems:

$$f(gw, x) = gf(w, \pi_g(x)), \quad \forall g \in G,\ x \in X. \tag{9}$$

where $\pi_g(x)$ is permutation of the set X that depends on g. The derivation stands on the simple observation:

$$\langle x, gw\rangle = \left\langle g^{-1}x, w\right\rangle$$

and the hypothesis that the training set is a collection of orbits. In fact in this case

$$gX = \pi_g(X).$$

In general if the training set X is large enough the dynamical system will be equivalent to the unpermuted one due to the stability of the stochastic gradient descent method [19]. Since the dynamical systems associated with the Oja and the ICA rules minimize statistical moments they are clearly independent of training data permutations. The fact that the set of solutions is a collections of orbits, $S = \bigcup_i O_i$ implies that any average operator over them is invariant. In our case the operator is the signature:

$$\mu(x) = \sum_{ij} \eta(\langle x, O_{ij}\rangle)$$

where O_{ij} is the element j of the orbit i and $\eta : \mathbb{R} \to \mathbb{R}$ is a non-linear function.

In the following we prove equivariance of a few learning rules.

1. **Unsupervised learning** rules[20]:

   In the following $x \in X$ and $\alpha > 0$ and with the notation $\pi_g(x)$ we indicate the permutation of the element x in the training set X due to the transformation g.

   - **Hebbian learning**. Choosing
   $$\mathcal{L}(w, x) = \frac{\alpha}{2}y^2 \tag{10}$$
   where $y = \langle x, w\rangle$ is the neuron's response, we have the associated dynamical system is:
   $$\dot{w} = f(x, w) = \alpha \langle x, w\rangle x. \tag{11}$$
   The system is equivariant. In fact:
   $$f(x, gw) = \alpha \langle x, gw\rangle x = g\alpha \left\langle g^{-1}x, w\right\rangle g^{-1}x = g\alpha \langle \pi_g(x), w\rangle \pi_g(x) = gf(\pi_g(x), w).$$

   - **Oja learning**. Choosing
   $$\mathcal{L}(w, x) = \frac{\alpha}{2\left\|w\right\|_2} \langle x, w\rangle^2 \tag{12}$$
   we obtain by differentiation:
   $$\dot{w} = f(w, x) = \alpha\frac{y}{\left\|w\right\|_2}(x - y\frac{w}{\left\|w\right\|_2}). \tag{13}$$
   The obtained dynamical system is that of Oja's for the choice $\left\|w\right\|_2 = 1$. The system is equivariant (note that $\left\|gw\right\|_2 = \left\|w\right\|_2$). In fact:
   $$\begin{aligned} f(gw, x) &= \alpha \langle x, gw\rangle (x - \langle x, gw\rangle gw) = \alpha \left\langle g^{-1}x, w\right\rangle g(g^{-1}x - \left\langle g^{-1}x, w\right\rangle w) \\ &= \alpha \langle \pi_g(x), w\rangle g(\pi_g(x) - \langle \pi_g(x), w\rangle w) = gf(w, \pi_g(x)) \end{aligned}$$

- **ICA**. Choosing

$$\mathcal{L}(w,x) = \alpha\frac{\langle x,w\rangle^4}{4} + \frac{\|w\|_2^2}{2} \tag{14}$$

we obtain the dynamical system:

$$\dot{w} = \alpha(\langle x,w\rangle^3 x - w) \tag{15}$$

which can be shown to extract one ICA component [25]. The system is equivariant. In fact:

$$f(x,gw) = \alpha(\langle x,gw\rangle^3 x - gw) = g\alpha(\langle g^{-1}x,w\rangle^3 g^{-1}x - w) = gf(\pi_g(x),w).$$

- **Foldiak.** Choosing:

$$\mathcal{L}(x,w) = \frac{\alpha}{2}\bar{y}^2, \quad \bar{y} = \int_{t_0}^{t} d\tau \ \langle w,x\rangle(\tau) \tag{16}$$

the associated dynamical system is:

$$\dot{w}(t) = \alpha\Big(\int_{t_0}^{t} d\tau \ \langle w,x\rangle(\tau)\Big)x(t) = \alpha\bar{y}x(t) \tag{17}$$

which is the so called Foldiak updating rule. The system is equivariant. In fact:

$$\begin{aligned} f(x,gw) &= \alpha\Big(\int_{t_0}^{t} d\tau \ \langle gw,x\rangle(\tau)\Big)x = g\alpha\Big(\int_{t_0}^{t} d\tau \ \langle w,g^{-1}x\rangle(\tau)\Big)g^{-1}x \\ &= \alpha g\bar{y}(w,\pi_g(x))\pi_g(x) = gf(w,\pi_g(x)) \end{aligned}$$

2. **Supervised learning in deep convolutional networks**. The reasoning above can be extended to supervised problems of the form:

$$\arg\min_{W} \ \mathcal{L}(X,\ell,W), \quad X = (x_1,\ldots,x_N) \tag{18}$$

where $\mathcal{L}(X,\ell,W) = Loss(X,\ell,W)$. The term $Loss(X,\ell,W)$ is a function defined using the loss of representing a set of observations X, their labels ℓ, and a the set of the network weights W. The updating rule for each weight w_l is given by the backpropagation algorithm:

$$\dot{w}_l = \frac{\partial\mathcal{L}}{\partial w_l}. \tag{19}$$

If the equation above is equivariant the same results of the previous section will hold, i.e., if there exists a solution the whole orbit will be a set of solutions. In the following we analyze the case of deep networks showing that equivariance holds if the output at each layer l, o_l is covariant w.r.t. the transformation, i.e.:

$$o_l(gx) = go_l(x), \quad \forall\, g \in G \tag{20}$$

We analyze the case of **deep convolutional networks** with pooling layers between each convolutional layer. In this case the response at each layer is covariant w.r.t. to the input transformation: the output at layer l is of the form:

$$o_l(X,W_{l-1})(g) = \int_{gG_l} d\hat{g} \ \eta(\langle o_{l-1}(X,W_{l-2}),\hat{g}w_l\rangle) = \int_{gG_l} d\hat{g} \ \eta(o_{l-1}(X,W_{l-2}) * w_l)(\hat{g}) \tag{21}$$

i.e. it is an average of a group convolution where o_{l-1} is the output of layer $l-1$ and W_{l-1} is the collection of weights up to layer $l-1$. Using the property that the group convolution commutes with group shift i.e. $[(T_{\bar{g}}f) * h](g) = T_{\bar{g}}[f * h](g)$ we have:

$$\begin{aligned} o_l(\bar{g}X,W_{l-1})(g) &= \int_{gG_l} d\hat{g} \ \eta(\bar{g}o_{l-1}(X,W_{l-2}) * w_l)(\hat{g}) = \int_{gG_l} d\hat{g} \ \eta(o_{l-1}(X,W_{l-2}) * w_l)(\bar{g}\hat{g}) \\ &= \int_{\bar{g}gG_l} d\hat{g} \ \eta(o_{l-1}(X,W_{l-2}) * w_l)(\hat{g}) = o_l(X,W_{l-1})(\bar{g}g) = \bar{g}o_l(X,W_{l-1})(g). \end{aligned}$$

where we used the property $o_{l-1}(\bar{g}X, W) = \bar{g}o_{l-1}(X, W)$. This can be seen to hold using an inductive reasoning up to the first layer where:

$$o_2(\bar{g}x, W_1)(g) = \int_{gG_1} d\hat{g}\ \eta((\bar{g}x) * w_1)(\hat{g}) = \int_{\bar{g}gG_1} d\hat{g}\ \eta(x * w_1)(\hat{g}) = \bar{g}o_1(x, W_1)(g).$$

In the following we prove that the dynamical systems (updating rules for the weights) associated to a deep convolutional network are equivariant. We consider e.g. the square loss function (the same reasoning can be extended to many commonly used loss functions):

$$\mathcal{L}(\phi_L(X, W), \ell) = \sum_{\ell}(1 - y_\ell \phi(X, W))^2.$$

where

$$\phi_L(X, W) = \phi_L(\cdots, \phi_3(\phi_2(X, w_1), w_3), \cdots, w_l \cdots, w_L)$$

being L the layers number and ℓ is a set of labels. The associated dynamical system reads as:

$$\frac{\partial \mathcal{L}(\phi_L(X, W), \ell)}{\partial w_l} = \dot{\mathcal{L}}(\phi_L(X, W), \ell)\frac{\partial \phi_L(X, W)}{\partial w_l} = 2\sum_{\ell}(1 - y_\ell \phi_L(X, W))\frac{\partial \phi_L(X, W)}{\partial w_l}$$

Substituting w_l with $\bar{g}w_l$ we have, by the covariance property, that the first factor of the r.h.s. of the equation above becomes $\sum_\ell(1 - y_\ell \phi_L(\pi_{\bar{g}}(X), W))$. We are then left to prove the equivariance of the second factor.
Using the chain rule, we have:

$$\begin{aligned} \dot{w}_l &= \frac{\partial \phi_L(\cdots, \phi_3(\phi_2(x, w_1), w_2) \cdots, w_L)}{\partial w_l} \\ &= \dot{\phi}_L[o_L(W_{L-1}, x)]\ \dot{\phi}_{L-1}[o_{L-1}(W_{L-2}, x)]\ \cdots\ \dot{\phi}_l(o_{l-1}(x, W_{l-2}), w_l), \cdots, w_L) \end{aligned}$$

where $o_j(W_{l-1}, x) = \phi_j(\cdots \phi_l(o_{l+1}(x, W_{l-1}), w_l), \cdots, w_L)$, $l < j < L$, being the output at layer j.
Notice that, in the case of covariant layer outputs, we have:

$$\begin{aligned} \phi_j(\cdots \phi_{l+1}(o_l(X, W_{l-1}), \bar{g}w_l), \cdots, w_L) &= \phi_j(\cdots \phi_{l+1}(\bar{g}^{-1}o_l(X, W_{l-1}), w_l), \cdots, w_L) \\ &= \phi_j(\cdots \phi_{l+1}(o_l(\bar{g}^{-1}X, W_{l-1})), w_l), \cdots, w_L) \\ &= \phi_j(\cdots \phi_{l+1}(o_l(\pi_{\bar{g}}(X), W_{l-1})), w_l), \cdots, w_L) \end{aligned}$$

where we used the covariance property in eq. (20) and the fact that the training set is a collection of orbits w.r.t. the group G.
Finally we have:

$$\frac{\partial \mathcal{L}(X, \{w_1, \cdots, \bar{g}w_l, \cdots, w_L\}, \ell)}{\partial w_l} = \bar{g}\frac{\partial \mathcal{L}(\pi_g(X), \{w_1, \cdots, w_l, \cdots, w_L\}, \ell)}{\partial w_l}$$

where the $\bar{g}$ comes from the derivative of $\bar{g}w_l$ w.r.t. w_l.
Summarizing we have the following result

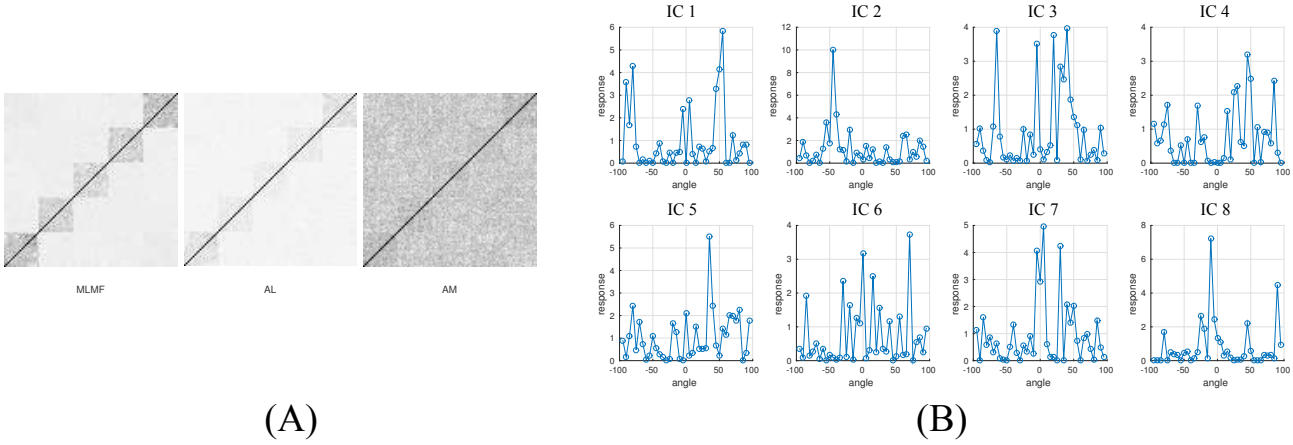
Theorem 2. *For $i = 1, \ldots, L$, let $\phi_i : \mathbb{R}^{d_i} \to R^{d_{i+1}}$ depend on a set of weights w_i. Consider a deep convolutional network with output of the form*

$$\phi_L(X, W) = \phi_L(\cdots, \phi_2(\phi_1(X, w_1), w_2), \cdots, w_l) \cdots, w_L). \tag{22}$$

and a differentiable square loss $\mathcal{L}(\phi_L(X, W), \ell)$, being ℓ a set of labels.
If X is a collection of orbits and and each ϕ_i is covariant, then the associated dynamical systems for each layer's weights' evolution in time

$$\dot{w}_l = \frac{\partial \mathcal{L}(\phi_L(X, W), \ell)}{\partial w_l}$$

are equivariant w.r.t. the group G.



(A) (B)

Figure 6: Experiments with ICA: we adopt the same pipeline as shown in the main text but replaced PCA with ICA [24] (which includes a ZCA-whitening preprocessing step). Similar to the original pipeline, for each training identity, ICA is performed to get 39 independent component directions. A testing image is projected to these directions. A square nonlinearity and a pooling are performed on the results. We show (A) the model population similarity matrices of different stages (similar to Fig 5A) and (B) some single cell responses in stage AL (similar to Fig 4A). Unlike PCA, the order of the independent components are arbitrary.

C Proof that the Oja equation's solutions are odd or even.

So far we have shown how biologically plausible learning dynamics in conjunction with appropriate training sets lead to solutions capable of supporting the computation of a view-invariant face signature (Sections A – B). We showed that several different learning rules satisfied these requirements: Hebb, Oja, Foldiak, ICA, and supervised backpropagation (Section B.1). Now we use properties specific to the Oja rule to address the question of why mirror symmetric responses arise in an intermediate step along the brain's circuit for computing view-invariant face representations.

We now use the following well-known property of Oja's learning rule: that it implements an online algorithm for principal component extraction [38]. More specifically, we use that the Oja dynamics converge to an eigenfunction of the training set's covariance $C(X)$.

Recall from section A.2 that in order to guarantee approximate view-invariance for bilaterally symmetric objects like faces, the training set X must consist of a collection of orbits of faces w.r.t. the reflection group $G = (e, R)$. We now show that this implies the eigenfunctions of $C(X)$ (equivalently, the principal components (PCs) of X) must be odd or even.

Under this hypothesis the covariance matrix $C(X)$ can be written as

$$C(X) = XX^\intercal = \mathcal{T}\mathcal{T}^\intercal + R\mathcal{T}\mathcal{T}^\intercal R^\intercal$$

where $\mathcal{T}$ is the set of the orbit representatives (untransformed vectors).
It is immediate to see that the above implies $[C(X), R] = 0$ (they commute). Thus $C(X)$ and R must share the same eigenfunctions. Finally, since the eigenfunctions of the reflection operator R are odd or even, this implies the eigenfunctions of $C(X)$ must also be odd or even.

Finally, we note that in the specific case of a frontal view, even basis functions (w.r.t. the zero view) are mirror symmetric.

D Empirical ICA solutions do not show any symmetry

Fig. 6 shows results from the analogous experiment to main Fig. 4. but with ICA instead of PCA. Note that the ICA result is not mirror symmetric.