CBCL, McGovern Institute, Brain Science Department, Computer Science and
Artificial Intelligence Laboratory, Massachusetts Institute of Technology,
Cambridge, MA, USA
Istituto Italiano di Tecnologia, Genova, Italy

DRAFT

Notes on versions and dates

This is a package of a theory under development since 2011.
The main monograph ("the magic memo") is version 3.1; it replaces version 3.0 which appeared on December 30th, 2012, as a CSAIL technical report. The original report was published online in Nature Precedings on July 20, 2011 (npre.2011.6117.1).
Before the "magic memo", which is quite outdated by now, the Magic Materials package contains an up-to-date (Sept 5, 2013) draft of a short paper for PNAS, summarizing some of the main results of the theory. The first version of the Magic Materials package (ver1) was posted online on March 21, 2013. The present version has been substantially updated.

Unsupervised Learning of Invariant Representations in Hierarchical Architectures

Fabio Anselmi [*,†], **Joel Z Leibo** [†], **Lorenzo Rosasco** [*,†], **Jim Mutch** [†], **Andrea Tacchetti** [*,†], **and Tomaso Poggio** [*,†]

[*]Istituto Italiano di Tecnologia, Genova, 16163, and [†]Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02139

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Representations that are invariant to translation, scale and other transformations, can considerably reduce the sample complexity of learning, allowing recognition of new object classes from very few examples—a hallmark of human recognition. Empirical estimates of one-dimensional projections of the distribution induced by a group of affine transformations are proven to represent a unique and invariant signature associated with an image. We show how projections yielding invariant signatures for future images can be learned automatically, and updated continuously, during unsupervised visual experience. A module performing filtering and pooling, like simple and complex cells as proposed by Hubel and Wiesel, can compute such estimates. Under this view, a pooling stage estimates a one-dimensional probability distribution. Invariance from observations through a restricted window is equivalent to a sparsity property w.r.t. to a transformation, which yields templates that are a) Gabor for optimal simultaneous invariance to translation and scale or b) very specific for complex, class-dependent transformations such as rotation in depth of faces. Hierarchical architectures consisting of this basic Hubel-Wiesel module inherit its properties of invariance, stability, and discriminability while capturing the compositional organization of the visual world in terms of wholes and parts, and are invariant to complex transformations that may only be locally affine. The theory applies to several existing deep learning convolutional architectures for image and speech recognition. It also suggests that the main computational goal of the ventral stream of visual cortex is to provide a hierarchical representation of new objects/images which is invariant to transformations, stable, and discriminative for recognition—and that this representation may be continuously learned in an unsupervised way during development and natural visual experience.

Invariance | Hierarchy | Convolutional networks | Visual cortex

We propose a theory of hierarchical architectures and, in particular, of the ventral stream in visual cortex. The initial assumption is that the computational goal of the ventral stream is to compute a representation of objects which is invariant to transformations. The theory shows how a process based on high-dimensional dot products can use stored "movies" of objects transforming, to encode new images in an invariant way. Theorems show that invariance implies several properties of the ventral stream organization and of the tuning of its neurons. Our main contribution is a theoretical framework for the next phase of machine learning beyond supervised learning: the unsupervised learning of representations that reduce the sample complexity of the final supervised learning stage.

It is known that Hubel and Wiesel's original proposal [1] for visual area V1—of a module consisting of complex cells (C-units) combining the outputs of sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct translation-invariant detectors. This is the insight underlying many networks for visual recognition, including HMAX [2] and convolutional neural nets [3, 4]. We show here how the original idea can be expanded into a comprehensive theory of visual recognition relevant for computer vision and possibly for visual cortex. The first step in the theory is the conjecture that a representation of images and image patches, with a feature vector that is invariant to a broad range of transformations—such as translation, scale, expression of a face, pose of a body, and viewpoint—makes it possible to recognize objects from only a few labeled labeled examples, as humans do. The second step is proving that hierarchical architectures of Hubel-Wiesel ('HW') modules (indicated by $\bigwedge$ in Fig. 1) can provide such invariant representations while maintaining discriminative information about the original image. Each $\bigwedge$-module provides a feature vector, which we call a *signature*, for the the part of the visual field that is inside its "receptive field"; the signature is invariant to ($\mathbb{R}^2$) affine transformations within the receptive field. The hierarchical architecture, since it computes a set of signatures for different parts of the image, is invariant to the rather general family of locally affine transformations (which includes globally affine transformations of the whole image). This remarkable invariance of the hierar-

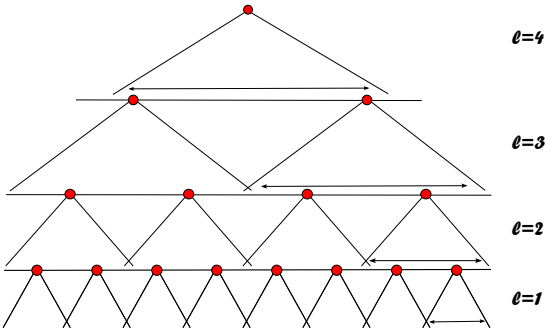


Fig. 1: *A hierarchical architecture built from HW-modules. Each red circle represents the signature vector computed by the associated module (the outputs of complex cells) and double arrows represent its receptive fields – the part of the (neural) image visible to the module (for translations this is also the pooling range). The "image" is at level* 0*, at the bottom. The vector computed at the top of the hierarchy consists of invariant features for the whole image and is usually fed as input to a supervised learning machine such as a classifier; in addition signatures from modules at intermediate layers may also be inputs to classifiers for objects and parts.*

Reserved for Publication Footnotes

[1] At the time of our writing, the working monograph [5] contains the most up-to-date account of the theory. The current monograph evolved from one that first appeared in July 2011 ([5]). Shorter papers describing isolated aspects of the theory have also appeared: [6, 7, 5]. The present paper is the first time the entire argument has been brought together in a short document.

chies we consider, follows from the key property of *covariance* of such architectures for image transformations and from the uniqueness and invariance of the individual module signatures. The basic HW-module is at the core of the properties of the architecture. This paper focuses first on its characterization and then outlines the rest of the theory, including its connections with machine learning, machine vision and neuroscience. Most of the theorems are in the supplementary information, where in the interest of telling a complete story we quote some results which are described more fully elsewhere[1].

Invariant representations and sample complexity

One could argue that the most important aspect of intelligence is the ability to learn. How do present supervised learning algorithms compare with brains? One of the most obvious differences is the ability of people and animals to learn from very few labeled examples. A child, or a monkey, can learn a recognition task from just a few examples. The main motivation of this paper is the conjecture that the key to reducing the sample complexity of object recognition is invariance to transformations. Images of the same object usually differ from each other because of simple transformations such as translation, scale (distance) or more complex deformations such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face).

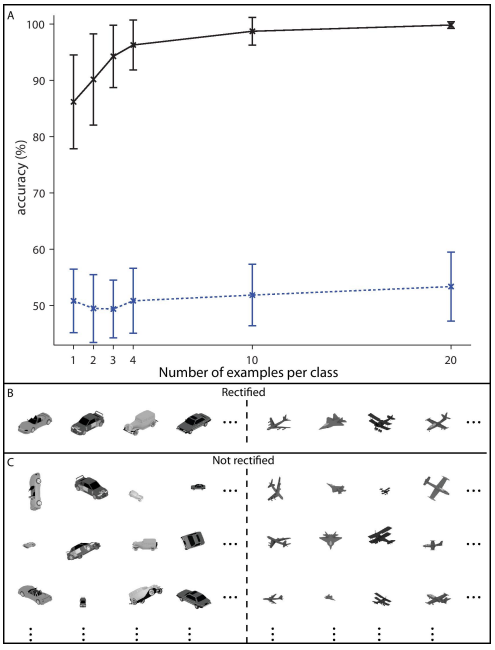


Fig. 2: *Sample complexity for the task of categorizing cars vs airplanes from their raw pixel representations (no preprocessing). A. Performance of a nearest-neighbor classifier (distance metric = 1 - correlation) as a function of the number of examples per class used for training. Each test used 74 randomly chosen images to evaluate the classifier. Error bars represent +/- 1 standard deviation computed over 100 training/testing splits using different images out of the full set of 440 objects × number of transformation conditions. Solid line: The rectified task. Classifier performance for the case where all training and test images are rectified with respect to all transformations; example images shown in B. Dashed line: The unrectified task. Classifier performance for the case where variation in position, scale, direction of illumination, and rotation around any axis (including rotation in depth) is allowed; example images shown in C. The images were created using 3D models from the Digimation model bank and rendered with Blender.*

The conjecture is supported by previous theoretical work showing that *almost all the complexity* in recognition tasks is often due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object [8]. It implies that in many cases, recognition—i.e., both identification, e.g., of a specific car relative to other cars—as well as categorization, e.g., distinguishing between cars and airplanes—would be much easier (only a small number of training examples would be needed to achieve a given level of performance), *if* the images of objects were rectified with respect to all transformations, or equivalently, if the image representation itself were invariant.

The case of identification is obvious since the difficulty in recognizing exactly the same object, e.g., an individual face, is only due to transformations. In the case of categorization, consider the suggestive evidence from the classification task in Fig. 2. The Fig. shows that if an oracle factors out all transformations in images of many different cars and airplanes, providing "rectified" images with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy: it can be done accurately with very few labeled examples. In this case, good performance was obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low.[2] We argue in this paper that the ventral stream in visual cortex tries to approximate such an oracle, providing a quasi-invariant signature for images and image patches.

Invariance and uniqueness

Consider the problem of recognizing an image, or an image patch, independently of whether it has been transformed by the action of a group like the affine group in $\mathbb{R}^2$. We would like to associate to each object/image I a *signature*, i.e. a vector which is *unique* and *invariant* with respect to a group of transformations, but our analysis, as we will see later, is not restricted to the case of groups. In the following, we will consider groups that are compact and, for simplicity, finite (of cardinality $|G|$). We indicate, with slight abuse of notation, a generic group element and its (unitary) representation with the same symbol g, and its action on an image as $gI(x) = I(g^{-1}x)$ (e.g. a translation, $g_\xi I(x) = I(x - \xi)$). A natural mathematical object to consider is the *orbit* O_I—i.e., the set of images gI generated from a single image I under the action of the group. We say that two images are equivalent when they belong to the same orbit: $I \sim I'$ if $\exists g \in G$ such that $I' = gI$. This equivalence relation formalizes the idea that an orbit is invariant and unique. Indeed, if two orbits have a point in common they are identical everywhere. Conversely, two orbits are different if none of the images in one orbit coincide with any image in the other (see also [9]).

How can two orbits be characterized and compared? There are several possible approaches. A distance between orbits can be defined in terms of a metric on images, but its computation is not obvious (especially by neurons). We follow here a different strategy: intuitively two empirical orbits are the same irrespective of the ordering of their points. This suggests that we consider the probability distribution P_I induced by the group's action on images I (gI can be seen as

[2] A similar argument involves estimating the cardinality of the universe of possible images generated by different viewpoints—such as variations in scale, position and rotation in 3D—versus true intraclass variability, e.g. different types of cars. With reasonable assumptions on resolution and size of the visual field, the first number would be several orders of magnitude larger than the, say, 10^3 distinguishable types of cars.

a realization of a random variable). It is possible to prove (see theorem 1 in SI Appendix section 1) that if two orbits coincide then their associated distributions under the group G are identical, that is

$$I \sim I' \iff O_I = O_{I'} \iff P_I = P_{I'}. \qquad [1]$$

The distribution P_I is thus invariant and discriminative but it also inhabits a high-dimensional space and is therefore difficult to estimate. In particular, it is unclear how neurons or neuron-like elements could estimate it.

As argued later, simple operations for neurons are (high-dimensional) inner products, $\langle \cdot, \cdot \rangle$, between inputs and stored "templates" which are neural images. It turns out that classical results (such as the Cramer-Wold theorem [10], see Theorem 2 section 1 in SI Appendix) ensure that a probability distribution P_I can be almost uniquely characterized by K one-dimensional probability distributions $P_{\langle I,t^k \rangle}$ induced by the (one-dimensional) results of projections $\langle I, t^k \rangle$, where $t^k,\ k = 1, ..., K$ are a set of randomly chosen images called templates. A probability function in d variables (the image dimensionality) induces a unique set of 1-D projections which is discriminative; empirically a small number of projections is usually sufficient to discriminate among a finite number of different probability distributions. Theorem 3 in SI Appendix section 1 says (informally) that an approximately invariant and unique signature of an image I can be obtained from the estimates of K 1-D probability distributions $P_{\langle I,t^k \rangle}$ for $k = 1, \cdots, K$. The number K of projections needed to discriminate n orbits, induced by n images, up to precision ϵ (and with confidence $1 - \delta^2$) is $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant.

Thus the discriminability question can be answered positively (up to ϵ) in terms of empirical estimates of the one-dimensional distributions $P_{\langle I,t^k \rangle}$ of projections of the image onto a finite number of templates $t^k, k = 1, ..., K$ under the action of the group.

Memory-based learning of invariance

Notice that the estimation of $P_{\langle I,t^k \rangle}$ requires the observation of the image *and* "all" its transforms gI. Ideally, however, we would like to compute an invariant signature for a new object seen only once (e.g., we can recognize a new face at different distances after just one observation). It is remarkable and almost magical that this is also made possible by the projection step. The key is the observation that $\langle gI, t^k \rangle = \langle I, g^{-1}t^k \rangle$. The same one-dimensional distribution is obtained from the projections of the image and all its transformations onto a fixed template, as from the projections of the image onto all the transformations of the same fixed template. Indeed, the distributions of the variables $\langle I, g^{-1}t^k \rangle$ and $\langle gI, t^k \rangle$ are the same. Thus it is possible for the system to store for each template t^k all its transformations gt^k for all $g \in G$ and later obtain an invariant signature for new images without any explicit understanding of the transformations g or of the group to which they belong. *Implicit knowledge of the transformations*, in the form of the stored templates, allows the system to be *automatically invariant to those transformations for new inputs (see eq.* [7] *in SI Appendix).*

An estimate of the one-dimensional Probability Density Functions (PDFs) $P_{\langle I,t^k \rangle}$ can be written in terms of histograms as $\mu_n^k(I) = 1/|G| \sum_{i=1}^{|G|} \eta_n(\langle I, g_i t^k \rangle)$, where $\eta_n,\ n = 1, \cdots, N$ is a set of nonlinear functions (see SI Appendix section 1). A visual system need not recover the actual probabilities from the empirical estimate in order to compute a unique signature. The set of $\mu_n^k(I)$ values is sufficient, since it identifies the associated orbit (see box 1 in SI Appendix). Crucially, mechanisms capable of computing invariant representations under affine transformations for future objects can be learned and maintained in an unsupervised automatic way by storing and updating sets of transformed templates which are *unrelated to those future objects.*

A theory of pooling

The arguments above make a few predictions. They require an effective normalization of the elements of the inner product (e.g. $\langle I, g_i t^k \rangle \mapsto \frac{\langle I, g_i t^k \rangle}{\|I\| \|g_i t^k\|}$) for the property $\langle gI, t^k \rangle = \langle I, g^{-1}t^k \rangle$ to be valid (see section 0 of SI Appendix). Notice that invariant signatures can be computed in several ways from one-dimensional probability distributions. Instead of the $\mu_n^k(I)$ components representing directly the empirical distribution, they may represent the moments $m_n^k(I) = 1/|G| \sum_{i=1}^{|G|} (\langle I, g_i t^k \rangle)^n$ of the same distribution [11]. Under weak conditions, the set of *all* moments uniquely characterizes the one-dimensional distribution $P_{\langle I,t^k \rangle}$ (and thus P_I). $n = 1$ corresponds to pooling via sum/average (and is the only pooling function that does not require a nonlinearity); $n = 2$ corresponds to "energy models" of complex cells and $n = \infty$ is related to the max-pooling. In our simulations, using just one of these moments seems to usually provide sufficient selectivity to a hierarchical architecture (see SI Appendix section 5). Other nonlinearities are also possible; see [5]. The arguments of this section may begin to provide a theoretical understanding of "pooling", giving insight to the search for the "best" choice in any particular setting—something which is normally done empirically for each application (e.g., [12]). According to this theory, these different pooling functions are all invariant, each one capturing part of the full information contained in the PDFs.

Implementations

There are other interesting and surprising results beyond the core of the theory described above. We sketch some of the main ones – the supplementary information provides the mathematical statements. Here it is important to stress that the theory has strong empirical support from several specific implementations which have been shown to perform well on a number of databases of natural images. The main set of tests is provided by HMAX, an architecture in which pooling is done with a max operation and invariance, to translation and scale, is mostly hardwired (instead of learned). Its performance on a variety of tasks is summarized in SI Appendix section 5. Strong performance is also achieved by other very similar architectures (again special cases of the theory) such as [13]. High performance for non-affine and even non-group transformations allowed by the hierarchical extension of the theory (see below) has been shown on large databases of face images, where our latest system advances the state-of-the-art on several tests [7]. Deep learning convolutional networks are another case of architectures that have achieved very good performance and are probably special cases of the theory even if they do not incorporate all of the possible invariances or their unsupervised learning ([14, 15], but see [16]).

Extensions of the Theory

Invariance Implies Localization and Sparsity. The core of the theory applies without qualification to compact groups such as rotations of the image in the image plane. Translation and scaling are however only locally compact, and in any case, each of the modules of Fig. 1 observes only a part of the transformation's full range. Each $\bigwedge$-module has a finite pooling range, corresponding to a finite "window" over the orbit associated with an image. *Exact invariance* for each module

is equivalent to a condition of *localization/sparsity* of the dot product between image and template (see Theorem 5 and Fig. 2 in section 1 of SI Appendix). In the simple case of a group parametrized by one parameter r the condition is:

$$\left\langle I, g_r t^k \right\rangle = 0 \quad |r| > a. \qquad [2]$$

Since this condition is a form of sparsity of the generic image I w.r.t. to a dictionary of templates t^k (under a group), this results provides a powerful justification for *sparse* encoding in sensory cortex (e.g. [17]).

It turns out that localization yields the following surprising result (Theorem 6 and 7 in SI Appendix): *optimal invariance for translation and scale implies Gabor functions as templates*. Since a frame of Gabor wavelets follows from natural requirements of completeness, this may also provide a general motivation for the Scattering Transform approach of Mallat based on wavelets [18].

The same Equation 2, if relaxed to hold approximately, that is $\left\langle I_C, g_r t^k \right\rangle \approx 0 \quad |r| > a$, becomes a *sparsity condition for the class of I_C wrt the dictionary t^k under the group G* when restricted to a subclass I_C of similar images. This property (see SI Appendix at the end of section 1), which is similar to compressive sensing "incoherence" (but in a group context), requires that I and t^k have a representation with rather sharply peaked autocorrelation (and correlation). When the condition is satisfied, the basic HW-module equipped with such templates can provide approximative invariance to non-group transformations such as rotations in depth of a face or its changes of expression (see Proposition 8, section 1, SI Appendix). In summary, condition Equation 2 can be satisfied in two different *regimes*. The first one, exact and valid for generic I, yields optimal Gabor templates. The second regime, approximate and valid for specific subclasses of I, yields highly tuned templates, specific for the subclass. Note that this arguments suggests generic, Gabor-like templates in the first layers of the hierarchy and highly specific templates at higher levels (note also that incoherence improves with increasing dimensionality).

Hierarchical architectures. We focused so far on the basic HW-module. Architectures consisting of such modules can be single-layer as well as multi-layer (hierarchical) (see Fig. 1). In our theory, the key property of hierarchical architectures of repeated HW-modules—allowing the recursive use of single module properties at all layers—is the property of *covariance*: the neural image at layer n transforms like the neural image at layer $n-1$, that is, calling $\Sigma_\ell(I)$ the signature at the ℓ^{th} layer, $\Sigma_l(g\Sigma_{\ell-1}(I)) = g^{-1}\Sigma_\ell(\Sigma_{\ell-1}(I)), \quad \forall g \in G, I \in \mathcal{X}$ (see Proposition 9 in section 2, SI Appendix).

One-layer networks can achieve invariance to *global* transformations of the whole image (exact invariance if the transformations are a subgroup of the affine group in $\mathbb{R}^2$) while providing a unique global signature which is stable with respect to small perturbations of the image, (see Theorem 4 SI Appendix and [5]). The two main reasons for a hierarchical architecture such as Fig. 1 are a) the need to compute an invariant representation not only for the whole image but especially for all parts of it which may contain objects and object parts and b) invariance to global transformations that are not affine (but are locally affine, that is, affine within the pooling range of some of the modules in the hierarchy)[3] Fig. 3 show examples of invariance and stability for wholes and parts. In the architecture of Fig. 1, each $\bigwedge$-module provides uniqueness, invariance and stability at different levels, over increasing ranges from bottom to top. Thus, in addition to the desired properties of invariance, stability and discriminability, these architectures match the hierarchical structure of the visual world and the need to retrieve items from memory at various levels of size and complexity. The results described

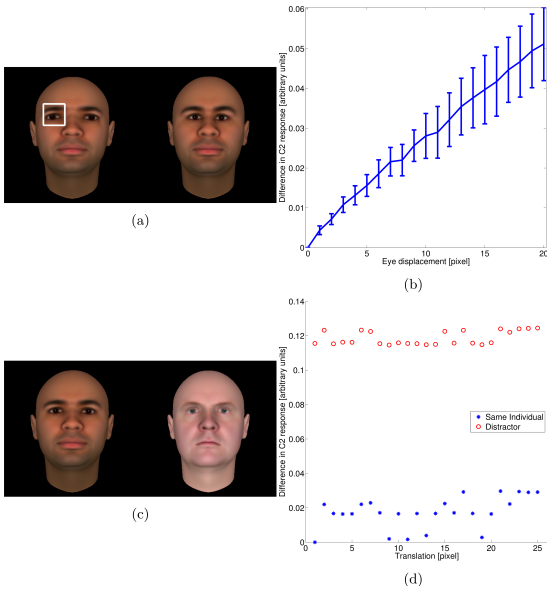


Fig. 3: *Empirical demonstration of the properties of invariance, stability and uniqueness of the hierarchical architecture (see Theorem 12) in a specific 2 layers implementation (HMAX). Inset (a) shows the reference image on the left and a deformation of it (the eyes are closer to each other) on the right; (b) shows an HW-module at layer 2 (c_2) whose receptive fields contain the whole face provides a signature vector which is (Lipschitz) stable with respect to the deformation. In all cases, the Figure shows just the Euclidean norm of the signature vector. Notice that the c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. Two different images (c) are presented at various location in the visual field. The Euclidean distance between the signatures of a set of HW-modules at layer 2 with the same receptive field (the whole image) and a reference vector is shown in (d). The signature vector is invariant to global translation and discriminative (between the two faces). In this example the HW-module represents the top of a hierarchical, convolutional architecture. The images we used were 200×200 pixels*

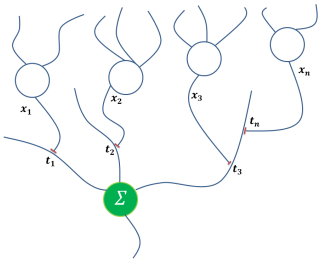


Fig. 4: *A neuron (green) can easily perform high-dimensional inner products between inputs on its dendritic tree and stored synapse weights.*

[3] Of course, one could imagine local and global one-layer architectures used in the same visual system without a hierarchical configuration, but there are further reasons favoring hierarchies including compositionality and reusability of parts. In addition to the issues of sample complexity and connectivity, one-stage architectures are unable to capture the hierarchical organization of the visual world where scenes are composed of objects which are themselves composed of parts. Objects (i.e., parts) can move in a scene relative to each other without changing their identity and often changing only in a minor way the scene (i.e., the object). Thus global and local signatures from all levels of the hierarchy must be able to access memory in order to enable the categorization and identification of whole scenes as well as of patches of the image corresponding to objects and their parts.

here are part of a general theory of hierarchical architectures which is beginning to take form (see [5, 18, 19, 20]) around the basic function of computing invariant representations.

The property of compositionality discussed above is related to the efficacy of hierarchical architectures vs. one-layer architectures in dealing with the problem of partial occlusion and the more difficult problem of clutter in object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks [21], because they provide signatures for image patches of several sizes and locations. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g. recurrent) architectures are likely needed for human-level recognition in clutter (see for instance [22, 23, 24]) and for other aspects of human vision. It is likely that much of the circuitry of visual cortex is required by these recurrent computations, not considered in this paper.

Visual Cortex

The theory described above effectively maps the computation of an invariant signature onto well-known capabilities of cortical neurons. A key difference between the basic elements of our digital computers and neurons is the number of connections: 3 vs. 10^3-10^4 synapses per cortical neuron. Taking into account basic properties of synapses, it follows that a single neuron can compute high-dimensional (10^3-10^4) inner products between input vectors and the stored vector of synaptic weights [25]. A natural scenario is then the following (see also Fig. 4). Consider an HW-module of "simple" and "complex" cells [1] looking at the image through a window defined by their receptive fields (see SI Appendix, section 1). Suppose that images of objects in the visual environment undergo affine transformations. During development—and more generally, during visual experience—a set of $|G|$ simple cells store in their synapses an image patch t^k and its transformations $g_1 t^k, ..., g_{|G|} t^k$—one per simple cell. This is done, possibly at separate times, for K different image patches t^k (templates), $k = 1, \cdots, K$. Each gt^k for $g \in G$ is a sequence of frames, literally a movie of image patch t^k transforming. There is a very *simple, general, and powerful way to learn* such unconstrained transformations. Unsupervised (Hebbian) learning is the main mechanism: for a "complex" cell to pool over several simple cells, the key is an unsupervised Foldiak-type rule: *cells that fire together are wired together.* At the level of complex cells this rule determines *classes of equivalence* among simple cells – reflecting observed *time correlations in the real world, that is transformations* of the image. Time continuity, induced by the Markovian physics of the world, allows associative labeling of stimuli based on their temporal contiguity.

Later, when an image is presented, the simple cells compute $\langle I, g_i t^k \rangle$ for $i = 1, ..., |G|$. The next step, as described above, is to estimate the one-dimensional probability distribution of such a projection, that is the distribution of the outputs of the simple cells. It is generally assumed that complex cells pool the outputs of simple cells. Thus a complex cell could compute $\mu_n^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + n\Delta)$ where σ is a smooth version of the step function ($\sigma(x) = 0$ for $x \leq 0$, $\sigma(x) = 1$ for $x > 0$) and $n = 1, ..., N$. Each of these N complex cells would estimate one bin of an approximated CDF (cumulative distribution function) for $P_{\langle I, t^k \rangle}$. Following the theoretical arguments above, the complex cells could compute, instead of an empirical CDF, one or more of its moments. $n = 1$ is the mean of the dot products, $n = 2$ corresponds to an energy model of complex cells [26]; very large n corresponds to a *max* operation. Conventional wisdom interprets available physiological data to suggest that simple/complex cells in V1 may be described in terms of energy models, but our alternative suggestion of empirical histogramming by sigmoidal nonlinearities with different offsets may fit the diversity of data even better.

As described above, a template and its transformed versions may be learned from unsupervised visual experience through Hebbian plasticity. Remarkably, our analysis and empirical studies[5] show that Hebbian plasticity, as formalized by Oja, can yield *Gabor-like tuning*—i.e., the templates that provide optimal invariance to translation and scale (see SI Appendix section 1)[4].

The localization condition (Equation 2) can also be satisfied by images and templates that are similar to each other. The result is invariance to class-specific transformations. This part of the theory is consistent with the existence of class-specific modules in primate cortex such as a face module and a body module [32, 33, 6]. It is intriguing that *the same localization condition* suggests *general Gabor-like templates for generic images* in the first layers of a hierarchical architectures and *specific, sharply tuned templates* for the last stages of the hierarchy[5]. This theory also fits physiology data concerning Gabor-like tuning in V1 and possibly in V4 (see [5]). It can also be shown that the theory, together with the hypothesis that storage of the templates takes place via Hebbian synapses, also predicts properties of the tuning of neurons in the face patch AL of macaque visual cortex [5, 34].

From the point of view of neuroscience, the theory makes a number of predictions, some obvious, some less so. One of the main predictions is that simple and complex cells should be found in all visual and auditory areas, not only in V1. Our definition of simple cells and complex cells is different from the traditional ones used by physiologists, which do not quite capture the different role in the theory of simple and complex cells. Simple cells represent the result of dot products between image and (transformed) templates: they are therefore linear. Complex cells represent invariant measurements associated with histograms of the outputs of simple cells or of moments of it. Probably the simplest and most useful moment is the average of the simple cells output: the corresponding complex cells are linear (contrary to common classification rules)[6]. The theory implies that invariance to all image transformations can be learned during development and adult life. This is however consistent with the possibility that the basic invariances may be genetically encoded by evolution but also refined and maintained by unsupervised visual experience. Studies on the development of visual invariance in organisms such as mice raised in virtual environments could test these predictions and their boundaries.

Discussion

The goal of this paper is to introduce a new theory of learning invariant representations for object recognition which cuts

[4] There is psychophysical and neurophysiological evidence that the brain employs such learning rules (e.g. [28, 30] and references therein). A second step of Hebbian learning may be responsible for wiring a complex cells to simple cells that are activated in close temporal contiguity and thus correspond to the same patch of image undergoing a transformation in time [27]. Simulations show that the system could be remarkably robust to violations of the learning rule's assumption that temporally adjacent images correspond to the same object [31]. The same simulations also suggest that the theory described here is qualitatively consistent with recent results on plasticity of single IT neurons and with experimentally-induced disruptions of their invariance [30].

[5] These incoherence properties of visual signatures are attractive from the point of view of information processing stages beyond vision, such as memory access.

[6] It is also important to note that simple and complex units do not need to always correspond to different cells: it is conceivable that a simple cell may be a cluster of synapses on a dendritic branch of a complex cell with nonlinear operations possibly implemented by active properties in the dendrites.

across levels of analysis [5, 35]. At the computational level, it gives a unified account of *why* a range of seemingly different models have recently achieved impressive results on recognition tasks. HMAX [2, 36, 37], Convolutional Neural Networks [3, 4, 38, 39] and Deep Feedforward Neural Networks [14, 15, 16] are examples of this class of architectures—as is, possibly, the feedforward organization of the ventral steam. In particular, the theoretical framework of this paper may help explain the recent successes of hierarchical architectures of convolutional type on visual and speech recognition tests e.g. [15, 14]). At the algorithmic level, it motivates the development, now underway, of a new class of models for vision and speech which includes the previous models as special cases. At the level of biological implementation, its characterization of the optimal tuning of neurons in the ventral stream is consistent with the available data on Gabor-like tuning in V1 ([5]) and the more specific types of tuning in higher areas such as in faces patches.

Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proven to be elusive. Thus it is interesting that the theory of this paper is directly implied by a simple hypothesis for the main computational function of the ventral stream: the representation of new objects/images in terms of a signature which is invariant to transformations learned during visual experience, thereby allowing recognition from very few labeled examples—in the limit, just one. A main contribution of our work to machine learning is a novel theoretical framework for the next major challenge in learning theory beyond the supervised learning setting which is now relatively mature: the problem of *representation learning*, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

ACKNOWLEDGMENTS. We would like to thank the McGovern Institute for Brain Research for their support. We would also like to thank for detailed and helpful comments on the manuscript Steve Smale, Stephane Mallat, Marco Cuturi, Robert Desimone, Jake Bouvrie, Charles Cadieu, Ryan Rifkin, Andrew Ng, Terry Sejnowski. This research was sponsored by grants from the National Science Foundation, AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by the Eugene McDermott Foundation.

1. D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex The Journal of Physiology 160, 1962.
2. M. Riesenhuber and T. Poggio. Models of object recognition. Nature Neuroscience, 3(11), 2000.
3. K. 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, Apr. 1980.
4. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.
5. F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio. Magic Materials: a theory of deep hierarchical architectures for learning sensory representations CBCL paper, Massachusetts Institute of Technology, Cambridge, MA, April 1, 2013.
6. J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In Advances in Neural Information Processing Systems (NIPS), Granada, Spain, 2011.
7. Q. Liao, J.Z. Leibo, T. Poggio. Learning invariant representations and applications to face verification NIPS to appear, 2013.
8. T. Lee and S. Soatto. Video-based descriptors for object recognition. Image and Vision Computing, 2012.
9. H. Schulz-Mirbach. Constructing invariant features by averaging techniques. In Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International. Conference on, volume 2, pages 387 –390 vol.2, 1994.
10. H. Cramer and H. Wold. Some theorems on distribution functions. J. London Math. Soc., 4:290–294, 1936.
11. A. Koloydenko. Symmetric measures via moments. Bernoulli, 14(2):362–390, 2008.
12. K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y . LeCun. What is the best multi-stage architecture for object recognition? IEEE International Conference on Computer Vision, 2146-2153, 2009.
13. N. Pinto, D. Doukhan, J.J. DiCarlo, D.D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Computational Biology, 5, 2009.
14. O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4277–4280. IEEE, 2012.
15. A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 2012.
16. Q. V. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. CoRR,http://arxiv.org/abs/1112.6209, abs/1112.6209, 2011.
17. B.A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature, 381, 6583, 607–609, 1996.
18. S. Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65(10):1331–1398, 2012.
19. S. Soatto. Steps Towards a Theory of Visual Information: Active Perception, Signal-to-Symbol Conversion and the Interplay Between Sensing and Control. arXiv:1110.2053, pages 0–151, 2011.
20. S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the neural response. Foundations of Computational Mathematics, 10(1):67–91, 2010.
21. T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Paper #259/AI Memo #2005-036, 2005.
22. S. S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A Bayesian inference theory of attention. Vision Research, May 2010.
23. D. George and J. Hawkins. A hierarchical bayesian model of invariant pattern recognition in the visual cortex. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), volume 3, pages 1812–1817, 2005.
24. S. Geman. Invariance and selectivity in the ventral visual pathway. Journal of Physiology-Paris, 100(4):212–224, 2006.
25. W.S. McCulloch, W. Pitts A logical calculus of the ideas immanent in the nervous activity Bull. Math. Biophysics 5, 5115-133, 1943.
26. E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A, 2(2):284–299, 1985.
27. P. Földiák. Learning invariance from transformation sequences. Neural Computation, 3(2):194–200, 1991.
28. 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, Apr. 2001.
29. D. Cox, P. Meier, N. Oertelt, and J. DiCarlo. 'Breaking'position-invariant object recognition. Nature Neuroscience, 8(9):1145–1147, 2005.
30. N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321(5895):1502–7, Sept. 2008.
31. L. Isik, J. Z. Leibo, and T. Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. Frontiers in Computational Neuroscience, 6, 2012.
32. N. Kanwisher, Functional specificity in the human brain: a window into the functional architecture of the mind, Proceedings of the National Academy of Sciences, 107, 25, 11163, 2010.
33. D.Y. Tsao, W.A. Freiwald, Faces and objects in macaque cerebral cortex Nature, 9, 6, 989-995, 2003.
34. J.Z. Leibo, F. Anselmi, J. Mutch, A.F. Ebihara, W. Freiwald, T. Poggio, Viewinvariance and mirror-symmetric tuning in a model of the macaque face-processing system Computational and Systems Neuroscience, I-54, 2013
35. D. Marr, T. Poggio From understanding computation to understanding neural circuitry AIM-357, 1976.
36. J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. Computer Vision and Pattern Recognition 2006, 1:11–18, 2006.
37. T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15):6424–6429, 2007.
38. Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, pages 255–258, 1995.
39. Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–97. IEEE, 2004.
40. C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo. The neural representation benchmark and its evaluation on brain and machine. arXiv preprint arXiv:1301.3530, 2013.

Supporting Information

Fabio Anselmi [*] [†], **Joel Z Leibo** [†], **Lorenzo Rosasco** [*] [†], **Jim Mutch** [†], **Andrea Tacchetti** [*] [†] **and Tomaso Poggio** [*] [†]

[*]Istituto Italiano di Tecnologia, Genova, 16163 , and [†]Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02139

Submitted to Proceedings of the National Academy of Sciences of the United States of America

0. Setup and Definitions

Let $\mathcal{X}$ be a Hilbert space with norm and inner product denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. We can think of $\mathcal{X}$ as the space of images (our images are usually "neural images"). We typically consider $\mathcal{X} = \mathbb{R}^d$, $L^2(\mathbb{R})$, $L^2(\mathbb{R}^2)$. We denote with G a (locally) compact group and with an abuse of notation, we denote by g both a group element in G and its action/representation on $\mathcal{X}$.

When useful we will make the following assumptions which are justified from a biological point of view.

Normalized dot products of signals (e.g. images or "neural activities") are usually assumed throughout the theory, for convenience but also because they provide *the most elementary invariances – to measurement units (origin and scale).* We assume that the dot products are between functions or vectors that are *zero-mean and of unit norm.* Thus $\langle I, t\rangle$ sets $I = \frac{I'-\bar{I'}}{\|I'-\bar{I'}\|}$, $t = \frac{t'-\bar{t'}}{\|t'-\bar{t'}\|}$ with $(\bar{\cdot})$ the mean. This normalization stage before each dot product is consistent with the convention that the empty surround of an isolated image patch has zero value (which can be taken to be the average "grey" value over the ensemble of images). In particular the dot product of a template – in general different from zero – and the "empty" region outside an isolated image patch will be zero. The dot product of two uncorrelated images – for instance of random 2D noise – is also approximately zero.

Remarks:

1. The k-th component of the signature associated with a *simple-complex* module is (see Equation **[10]**) $\mu_n^k(I) = \frac{1}{|G_0|}\sum_{g\in G_0}\eta_n\Big(\langle gI, t^k\rangle\Big)$ where the functions η_n are such that $Ker(\eta_n) = \{0\}$: in words, the empirical histogram estimated for $\langle gI, t^k\rangle$ does not take into account the 0 value, since it does not carry any information about the image patch. The functions η_n are also assumed to be positive and invertible.
2. Images I are inputs to the modules of later one and have a maximum total possible support corresponding to a bounded region $B \subseteq \mathbb{R}^2$, which we refer to as the *visual field*, and which corresponds to the spatial pooling range of the module at the top of the hierarchy of Figure 1 in the main text. *Neuronal images* also written as I are inputs to the modules in higher layers and are usually supported in a higher dimensional space $\mathbb{R}^d$, corresponding to the signature components provided by lower layers modules; *isolated objects* are images with support contained in the pooling range of one of the modules at an intermediate level of the hierarchy. We use the notation $\nu(I), \mu(I)$ respectively for the simple responses $\langle gI, t^k\rangle$ and for the complex response $\mu_n^k(I) = \frac{1}{|G_0|}\sum_{g\in G_0}\eta_n\Big(\langle gI, t^k\rangle\Big)$. To simplify the notation we suppose that the center of the support of $\mu_\ell(I)$ coincides with the center of the pooling range.
3. The domain of the dot products $\langle gI, t^k\rangle$ corresponding to templates and to simple cells is in general different from the domain of the pooling $\sum_{g\in G_0}$. We will continue to use the commonly used term *receptive field* – even if it mixes these two domains.
4. The main part of the theory characterizes properties of the basic HW module – which computes the components of an invariant signature vector from an image patch within its receptive field.
5. It is important to emphasize that the *basic module is always the same* throughout the paper. We use different mathematical tools, including approximations, to study under which conditions (e.g. localization or linearization, see end of section 1) the signature computed by the module is invariant or approximatively invariant.
6. The pooling $\sum_{g\in G_0}$ is effectively over a *pooling window* in the group parameters. In the case of 1D scaling and 1D translations, the pooling window corresponds to an interval, e.g. $[a^j, a^{j+k}]$, of scales and an interval, e.g. $[-\bar{x}, \bar{x}]$, of x translations, respectively.
7. All the results in this paper are valid in the case of a discrete or a continuous compact group: in the first case we have a sum over the transformations, in the second an integral over the Haar measure of the group. In the following, for convenience, the theorems are proved in the continuous setting.
8. Normalized dot products also eliminate the need of the explicit computation of the determinant of the Jacobian for affine transformations (which is a constant and is simplified dividing by the norms) assuring that $\langle AI, At\rangle = \langle I, t\rangle$, where A is an affine transformation.

1. Basic Module

Compact Groups (fully observable). Given an image $I \in \mathcal{X}$ and a group representation g, the orbit $O_I = \{I' \in \mathcal{X} s.t. I' = gI, g \in G\}$ is uniquely associated to an image and all its transformations. The orbit provides an invariant representation of I, i.e. $O_I = O_{gI}$ for all $g \in G$. Indeed, we can view an orbit as all the possible realizations of a random variable with distribution P_I induced by the group action. From this observation, a signature $\Sigma(I)$ can be derived for compact groups, by using results characterizing probability distributions via their one dimensional projections.

In this section we study the signature given by

$$\Sigma(I) = (\mu^1(I), \ldots, \mu^k(I)) = (\mu_1^1(I), .., \mu_N^1, .., \mu_1^K, .., \mu_N^K(I)),$$

Reserved for Publication Footnotes

where each component $\mu^k(I) \in \mathbb{R}^N$ is a histogram corresponding to a one dimensional projection defined by a template $t^k \in \mathcal{X}$. In the following we let $\mathcal{X} = \mathbb{R}^d$.

Orbits and probability distributions. If G is a compact group, the associated Haar measure dg can be normalized to be a probability measure, so that, for any $I \in \mathbb{R}^d$, we can define the random variable,

$$Z_I : G \to \mathbb{R}^d, \qquad Z_I(g) = gI.$$

The corresponding distribution P_I is defined as $P_I(A) = dg(Z_I^{-1}(A))$ for any Borel set $A \subset \mathbb{R}^d$ (with some abuse of notation we let dg be the normalized Haar measure).
Recall that we define two images, $I, I' \in \mathcal{X}$ to be equivalent (and we indicate it with $I \sim I'$) if there exists $g \in G$ s.t. $I = gI'$. We have the following theorem:

Theorem 1. *The distribution P_I is invariant and unique i.e. $I \sim I' \Leftrightarrow P_I = P_{I'}$.*

Proof:
We first prove that $I \sim I' \Rightarrow P_I = P_{I'}$. By definition $P_I = P_{I'}$ iff $\int_A dP_I(s) = \int_A dP_{I'}(s)$, $\forall\ A \subseteq \mathcal{X}$, that is $\int_{Z_I^{-1}(A)} dg = \int_{Z_{I'}^{-1}(A)} dg$, where,

$$Z_I^{-1}(A) = \{g \in G \ s.t.\ gI \subseteq A\}$$
$$Z_{I'}^{-1}(A) = \{g \in G \ s.t.\ gI' \in A\} = \{g \in G \ s.t.\ g\bar{g}I \subseteq A\},$$

$\forall\ A \subseteq \mathcal{X}$. Note that $\forall\ A \subseteq \mathcal{X}$ if $gI \in A \Rightarrow g\bar{g}^{-1}\bar{g}I = g\bar{g}^{-1}I' \in A$, so that $g \in Z_I^{-1}(A) \Rightarrow g\bar{g}^{-1} \in Z_{I'}^{-1}(A)$, i.e. $Z_I^{-1}(A) \subseteq Z_{I'}^{-1}(A)$. Conversely $g \in Z_{I'}^{-1}(A) \Rightarrow g\bar{g} \in Z_I^{-1}(A)$, so that $Z_I^{-1}(A) = Z_{I'}^{-1}(A)\bar{g}$, $\forall A$. Using this observation we have,

$$\int_{Z_I^{-1}(A)} dg = \int_{(Z_{I'}^{-1}(A))\bar{g}} dg = \int_{Z_{I'}^{-1}(A)} d\hat{g}$$

where in the last integral we used the change of variable $\hat{g} = g\bar{g}^{-1}$ and the invariance property of the Haar measure: this proves the implication.
To prove that $P_I = P_{I'} \Rightarrow I \sim I'$, note that $P_I(A) - P_{I'}(A) = 0$, $\forall\ A \subseteq \mathcal{X}$, is equivalent to

$$\int_{Z_{I'}^{-1}(A)} dg - \int_{Z_I^{-1}(A)} dg = \int_{Z_I^{-1}(A) \triangle Z_{I'}^{-1}(A)} dg = 0,\ \forall A \in \mathcal{X}$$

where $\triangle$ denotes the symmetric difference. This implies $Z_I^{-1}(A) \triangle Z_{I'}^{-1}(A) = \emptyset$ or equivalently

$$Z_I^{-1}(A) = Z_{I'}^{-1}(A),\ \forall\ A \in \mathcal{X}$$

In other words of any element in A there exist $g', g'' \in G$ such that $g'I = g''I'$. This implies $I = g'^{-1}g''I' = \bar{g}I'$, $\bar{g} = g'^{-1}g''$, i.e. $I \sim I'$. Q.E.D.

Random Projections for Probability Distributions.. Given the above discussion, a *signature* may be associated to I by constructing a histogram approximation of P_I, but this would require dealing with high dimensional histograms. The following classic theorem gives a way around this problem.
For a *template* $t \in \mathbb{S}(\mathbb{R}^d)$, where $\mathbb{S}(\mathbb{R}^d)$ is unit sphere in $\mathbb{R}^d$, let $I \mapsto \langle I, t\rangle$ be the associated projection. Moreover, let $P_{\langle I,t\rangle}$ be the distribution associated to the random variable $g \mapsto \langle gI, t\rangle$ (or equivalently $g \mapsto \langle I, g^{-1}t\rangle$, if g is unitary). Let $\mathcal{E} = [t \in \mathbb{S}(\mathbb{R}^d),\ s.t.\ P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}]$.

Theorem 2. *(Cramer-Wold, [1]) For any pair P, Q of probability distributions on $\mathbb{R}^d$, we have that $P = Q$ if and only if $\mathcal{E} = \mathbb{S}(\mathbb{R}^d)$.*

In words, two probability distributions are equal if and only if their projections on any of the unit sphere directions is equal. The above result can be equivalently stated as saying that the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 1 if and only if $P = Q$ and the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 0 if and only if $P \neq Q$ (see Theorem 3.4 in [2]). The theorem suggests a way to define a metric on distributions (orbits) in terms of

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle}) d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_0 is any metric on one dimensional probability distributions and $d\lambda(t)$ is a distribution measure on the projections. Indeed, it is easy to check that d is a metric. In particular note that, in view of the Cramer Wold Theorem, $d(P, Q) = 0$ if and only if $P = Q$. As mentioned in the main text, each one dimensional distribution $P_{\langle I,t\rangle}$ can be approximated by a suitable histogram $\mu^t(I) = (\mu_n^t(I))_{n=1,\ldots,N} \in R^N$, so that, in the limit in which the histogram approximation is accurate

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I')) d\lambda(t), \quad \forall I, I' \in \mathcal{X}, \qquad \mathbf{[1]}$$

where d_μ is a metric on histograms induced by d_0.

A natural question is whether there are situations in which a finite number of projections suffice to discriminate any two probability distributions, that is $P_I \neq P'_I \Leftrightarrow d(P_I, P_{I'}) \neq 0$. Empirical results show that this is often the case with a small number of templates (see [3] and HMAX experiments, section 5). The problem of mathematically characterizing the situations in which a finite number of (one-dimensional) projections are sufficient is challenging. Here we provide a partial answer to this question.
We start by observing that the metric **[1]** can be approximated by uniformly sampling K templates and considering

$$\hat{d}_K(P_I, P_{I'}) = \frac{1}{K} \sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')), \qquad \mathbf{[2]}$$

where $\mu^k = \mu^{t^k}$. The following result shows that a finite number K of templates is sufficient to obtain an approximation within a given precision ϵ. Towards this end let

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\| \mu^k(I) - \mu^k(I) \right\|_{\mathbb{R}^N}. \qquad \mathbf{[3]}$$

where $\|\cdot\|_{\mathbb{R}^N}$ is the Euclidean norm in $\mathbb{R}^N$. The following theorem holds:

Theorem 3. *Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then*

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon, \qquad \mathbf{[4]}$$

with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

Proof:
The proof follows from an application of Höeffding inequality and a union bound.
Fix $I, I' \in \mathcal{X}_n$. Define the real random variable $Z : \mathbb{S}(\mathbb{R}^d) \to \mathbb{R}$,

$$Z(t^k) = \left\| \mu^k(I) - \mu^k(I) \right\|_{\mathbb{R}^N}, \quad k = 1, \ldots, K.$$

From the definitions it follows that $\|Z\| \leq c$ and $\mathbb{E}(Z) = d(P_I, P_{I'})$. Then Höeffding inequality implies

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| = |\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}(Z) - Z(t^k)| \geq \epsilon,$$

with probability at most $e^{-c\epsilon^2 k}$. A union bound implies a result holding uniformly on $\mathcal{X}_n$; the probability becomes at most $n^2 e^{-c\epsilon^2 K}$. The desired result is obtained noting that this probability is less than δ^2 as soon as $n^2 e^{-c\epsilon^2 K} < \delta^2$ that is $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$. Q.E.D.

The above result shows that the discriminability question can be answered in terms of empirical estimates of the one-dimensional distributions of projections of the image and transformations induced by the group on a number of templates $t^k, k = 1, ..., K$.
Theorem 3 can be compared to a version of the Cramer Wold Theorem for discrete probability distributions. Theorem 1 in [4] shows that for a probability distribution consisting of k atoms in $\mathbb{R}^d$, we see that at most $k+1$ directions ($d_1 = d_2 = ... = d_{k+1} = 1$) are enough to characterize the distribution, thus a finite – albeit large – number of one-dimensional projections.

The signature $\Sigma(I) = (\mu_1^1(I), \ldots, \mu_N^K(I))$ is obviously invariant (and unique) since it is associated to an image and all its transformations (an orbit). Each component of the signature is also invariant – it corresponds to a group average. Indeed, each measurement can be defined as

$$\mu_n^k(I) = \frac{1}{|G|}\sum_{g\in G} \eta_n\Big(\big\langle gI, t^k\big\rangle\Big), \qquad [5]$$

for G finite group, or equivalently

$$\mu_n^k(I) = \int_G dg\ \eta_n\Big(\big\langle gI, t^k\big\rangle\Big) = \int_G dg\ \eta_n\Big(\big\langle I, g^{-1}t^k\big\rangle\Big), \quad [6]$$

when G is a (locally) compact group. Here, the non linearity η_n is chosen to define an histogram approximation. Then, it is clear that from the properties of the Haar measure we have

$$\mu_n^k(\bar{g}I) = \mu_n^k(I), \quad \forall \bar{g} \in G, I \in \mathcal{X}. \qquad [7]$$

Box 1: computing an invariant signature $\mu(I)$

procedure SIGNATURE(I)
Given K templates $\{gt^k | \forall g \in G\}$.
  for $k = 1, \ldots, K$ **do**
    Compute $\big\langle I, gt^k\big\rangle$, the normalized dot products of the image with all the transformed templates (all $g \in G$).
    Pool the results: POOL($\{\big\langle I, gt^k\big\rangle | \forall g \in G\}$).
  end for
  return $\mu(I)$ = the pooled results for all k.
▷ $\mu(I)$ is unique and invariant if there are enough templates.
end procedure

Stability. With $\Sigma(I) \in \mathbb{R}^{NK}$ denoting as usual the signature of an image, and $d(\Sigma(I), \Sigma(I'))$, $I, I' \in \mathcal{X}$, a metric, we say that a *signature* Σ *is stable if it is Lipschitz continuous* (see [7]), that is

$$d(\Sigma(I), \Sigma(I')) \leq L\left\|I - I'\right\|_2, \quad L > 0, \ \forall I, I' \in \mathcal{X}. \qquad [8]$$

In our setting we let

$$d(\Sigma(I), \Sigma(I')) = \frac{1}{K}\sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')),$$

and assume that $\mu_n^k(I) = \int dg\ \eta_n(\big\langle gI, t^k\big\rangle)$ for $n = 1, \ldots, N$ and $k = 1, \ldots, K$. If $L < 1$ we call the signature map contractive. The following theorem holds.

Theorem 4. *Assume the templates to be normalized and $L_\eta = \max_n(L_{\eta_n})$ s.t. $NL_\eta < 1$, where L_{η_n} is the Lipschitz constant of the function η_n. Then*

$$d(\Sigma(I), \Sigma(I')) \leq \left\|I - I'\right\|_2, \qquad [9]$$

for all $I, I' \in \mathcal{X}$.

Proof:
By definition, if the non linearities η_n are Lipschitz continuous, for all $n = 1, \ldots, N$, with Lipschitz constant L_{η_n}, it follows that for each k component of the signature we have

$$\begin{aligned} &\left\|\Sigma^k(I) - \Sigma^k(I')\right\|_{\mathbb{R}^N} \\ \leq\ & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N}\Big(\sum_{g\in G} L_{\eta_n} |\ \langle gI, t^k\rangle - \langle gI', t^k\rangle\ |\Big)^2} \\ \leq\ & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N} L_{\eta_n}^2 \sum_{g\in G} (|\ \langle g(I - I'), t^k\rangle\ |)^2}, \end{aligned}$$

where we used the linearity of the inner product and Jensen's inequality. Applying Schwartz's inequality we obtain

$$\left\|\Sigma^k(I) - \Sigma^k(I')\right\|_{\mathbb{R}^N} \leq \frac{L_\eta}{|G|}\sqrt{\sum_{n=1}^{N}\sum_{g\in G} \|I - I'\|^2 \|g^{-1}t^k\|^2}$$

where $L_\eta = \max_n(L_{\eta_n})$. If we assume the templates and their transformations to be normalized to unity then we finally have,

$$\left\|\Sigma^k(I) - \Sigma^k(I')\right\|_{\mathbb{R}^N} \leq NL_\eta \left\|I - I'\right\|_2.$$

from which we obtain [8] summing over all K components and dividing by $1/K$. In particular if $NL_\eta \leq 1$ the map is non expansive and summing each component we have eq. [9]. Q.E.D.

The above result shows that the stability of the empirical signature

$$\Sigma(I) = (\mu_1^1(I), \ldots, \mu_N^K(I)) \in \mathbb{R}^{NK},$$

provided with the metric [2] (together with [3]) holds for nonlinearities with Lipschitz constants L_{η_n} such that $Nmax_n(L_{\eta_n}) < 1$.

Partially Observable Groups. This section outlines invariance, uniqueness and stability properties of the signature obtained in the case in which transformations of a group are observable only within a *window* "over" the orbit. The term POG (Partially Observable Groups) emphasizes the properties of the group – in particular associated invariants – as seen by an observer (e.g. a neuron) looking through a window at a

part of the orbit. Let G be a finite group and $G_0 \subseteq G$ a subset (note: G_0 is not usually a subgroup). The subset of transformations G_0 can be seen as the set of transformations that can be *observed* by a window on the orbit that is the transformations that correspond to a part of the orbit. A *local* signature associated to the partial observation of G can be defined considering

$$\mu_n^k(I) = \frac{1}{|G_0|} \sum_{g \in G_0} \eta_n\Big(\big\langle gI, t^k \big\rangle \Big), \qquad [10]$$

and $\Sigma_{G_0}(I) = (\mu_n^k(I))_{n,k}$. This definition can be generalized to any locally compact group considering,

$$\mu_n^k(I) = \frac{1}{V_0} \int_{G_0} \eta_n\Big(\big\langle gI, t^k \big\rangle \Big) dg, \quad V_0 = \int_{G_0} dg. \qquad [11]$$

Note that the constant V_0 normalizes the Haar measure, restricted to G_0, so that it defines a probability distribution. The latter is the distribution of the images subject to the group transformations which are observable, that is in G_0. The above definitions can be compared to definitions **[5]** and **[6]** in the fully observable groups case. In the next sections we discuss the properties of the above signature. While stability and uniqueness follow essentially from the analysis of the previous section, invariance requires developing a new analysis.

POG: Stability and Uniqueness. A direct consequence of Theorem 1 is that *any two orbits with a common point are identical*. This follows from the fact that if $gI, g'I'$ is a common point of the orbits, then

$$g'I' = gI \;\Rightarrow\; I' = (g')^{-1}gI.$$

Thus the two images are transformed versions of one another and $O_I = O_{I'}$.
Suppose now that only a fragment of the orbits – the part within the window – is observable; the reasoning above is still valid since if the orbits are different or equal so must be any of their "corresponding" parts.
Regarding the stability of POG signatures, note that the reasoning in the previous section can be repeated without any significant change. In fact, only the normalization over the transformations is modified accordingly.

POG: Partial Invariance and Localization. Since the group is only partially observable we introducethe notion of *partial invariance* for images and transformations G_0 that are within the observation window. Partial invariance is defined in terms of invariance of

$$\mu_n^k(I) = \frac{1}{V_0} \int_{G_0} dg\; \eta_n\Big(\big\langle gI, t^k \big\rangle \Big). \qquad [12]$$

We recall that when gI and t^k do not share any common support on the plane or I and t are uncorrelated, then $\langle gI, t^k \rangle = 0$. The following theorem, where G_0 corresponds to the pooling range states a sufficient and necessary condition for partial invariance:

Theorem 5. Invariance and Localization. *Let $I, t \in H$ a Hilbert space, $\eta_n : \mathbb{R} \to \mathbb{R}^+$ a set of bijective (positive) functions and G a locally compact group. Let $G_0 \subseteq G$ and suppose $supp(\langle gI, t^k \rangle) \subseteq G_0$. Then for any given $\bar{g} \in G$, $t^k, I \in \mathcal{X}$*

$$\mu_n^k(I) = \mu_n^k(\bar{g}I) \;\Leftrightarrow\; \begin{aligned} &\big\langle gI, t^k \big\rangle = 0, \;\forall g \in G/(G_0 \cap \bar{g}G_0), \\ &\big\langle gI^k, t \big\rangle \neq 0, \;\forall g \in G_0 \cap \bar{g}G_0. \end{aligned} \qquad [13]$$

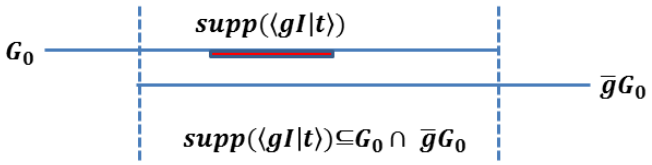


Fig. 1: *Necessary and sufficient condition for local invariance: if the support of $\langle gI, t \rangle$ is sufficiently localized it will be completely contained in the pooling interval even if the image is group shifted, or, equivalently (as shown in the Figure), if the pooling interval is group shifted by the same amount.*

Proof:
If $\mu_n^k(I) - \mu_n^k(\bar{g}I) = 0$ by definition we have

$$\begin{aligned} 0 &= \int_{G_0} dg\; \eta_n\Big(\big(\big\langle gI, t^k \big\rangle \big) - \eta_n\big(\big\langle g\bar{g}I, t^k \big\rangle \big) \Big) \\ &= \int_{G_0 \Delta \bar{g} G_0} dg\; \eta_n\big(\big\langle gI, t^k \big\rangle \big) \qquad [14] \\ &= \int_{G/(G_0 \cap \bar{g}G_0)} dg\; \eta_n\big(\big\langle gI, t^k \big\rangle \big) \end{aligned}$$

where Δ is the symbol for symmetric difference ($A \Delta B = (A \cup B)/(A \cap B)$ A, B *sets*) and the last equality holds if $supp(\langle gI, t^k \rangle) \subseteq G_0$. Since the functions η_n are positive and bijective, eq. **[14]** implies $\langle gI, t^k \rangle = 0$, $g \in G/(G_0 \cap \bar{g}G_0)$. See Fig. above for a visual explanation. Q.E.D.
Condition in eq. **[14]** is a *localization* condition on the product of the transformed image and the template (see Fig. below for a pictorial intuitive example in the case of translation group). In the next paragraph we will see how localization conditions for scale and translation transformations implies a specific form of the templates.

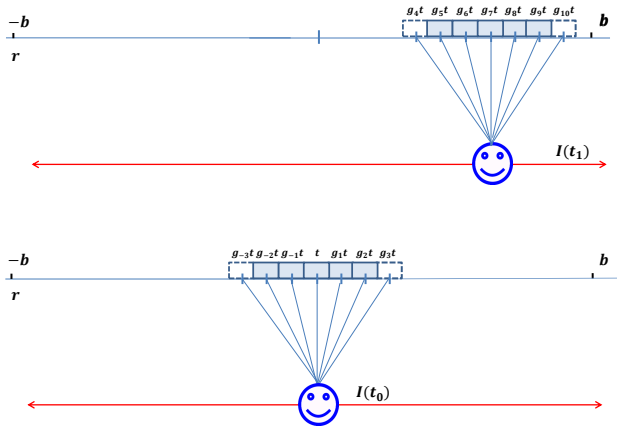


Fig. 2: *An HW-module pooling the dot products of transformed templates with the image. The input image I is shown centered on the template t; the same module is shown above for a group shift of the input image, which now localizes around the transformed template $g_7 t$. Images and templates satisfy the localization condition $\langle I, T_x t \rangle \neq 0$, $|x| > a$ with $a = 3$. The interval $[-b, b]$ indicates the pooling window. The shift in x shown in the Figure is a special case: the reader should consider the case in which the transformation parameter, instead of x, is for instance rotation in depth.*

The Localization condition: Translation and Scale In this section we identify G_0 with subsets of the affine group. In particular, we study separately the case of scale and translations (in 1D for simplicity).

In the following it is helpful to assume that all images I and templates t are strictly contained in the range of translation or scale pooling, P, since image components outside it are not measured. We will consider images I *restricted to* P: for translation this means that the support of I is contained in P, for scaling, since $g_s I = I(sx)$ and $\widehat{I(sx)} = (1/s)\hat{I}(\omega/s)$ (where $\hat{\cdot}$ indicates the Fourier transform), assuming a scale pooling range of $[s_m, s_M]$, implies a range $[\omega^I_m, \omega^I_M]$, $[\omega^t_m, \omega^t_M]$ (m and M indicates maximum and minimum) of spatial frequencies for the maximum support of I and t. As we will see because of Theorem 5 *invariance to translation requires spatial localization of images and templates* and less obviously *invariance to scale requires bandpass properties of images and templates*. Thus images and templates are assumed to be localized from the outset in either space or frequency. The corollaries below show that a stricter localization condition is needed for invariance and that this condition determines the form of the template. Notice that in our framework images and templates are bandpass because of being zero-mean. Notice that, in addition, neural "images" which are input to the hierarchical architecture are spatially bandpass because of retinal processing.

We now state the result of Theorem 5 for one dimensional signals under the translation group and – separately – under the dilation group.

Let $I, t \in L^2(\mathbb{R})$, $(\mathbb{R}, +)$ the one dimensional locally compact group of translations and $T_x : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the translation operator. Let, e.g., $G_0 = [-b, b]$, $b > 0$ and suppose $supp(I) \subseteq [-b, b]$. Further suppose $supp(\langle T_x I, t\rangle) \subseteq [-b, b]$. Then eq. **[13]** specializes to

Corollary 1: *Localization in the spatial domain is necessary and sufficient for translation invariance.* For any fixed $t, I \in \mathcal{X}$ we have:

$$\mu^k_n(I) = \mu^k_n(T_x I),\ \forall x \in [0, \bar{x}] \ \Leftrightarrow\ \langle T_x I, t\rangle \neq 0,\ \forall x \in [-b+\bar{x}, b]. \quad \textbf{[15]}$$

with $\bar{x} > 0$.

Similarly let $G = (\mathbb{R}^+, \cdot)$ be the one dimensional locally compact group of dilations and denote with $D_s : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the dilation operator. Let $G_0 = [1/S, S]$, $S > 1$ and suppose $supp(\langle D_s I, t\rangle) \subseteq [1/S, S]$. Then eq. **[13]** gives

Corollary 2: *Localization in the spatial frequency domain is necessary and sufficient for scale invariance.* For any fixed $t, I \in \mathcal{X}$ we have:

$$\mu^k_n(I) = \mu^k_n(D_s I), s \in [1, \bar{s}] \ \Leftrightarrow\ \langle D_s I, t\rangle \neq 0,\ \forall s \in [\frac{\bar{s}}{S}, S]. \quad \textbf{[16]}$$

with $S > 1$.

Localization conditions of the support of the dot product for translation and scale are depicted in Fig. 3,a),b).

As shown by the following Lemma 1 Eq. **[15]** and **[16]** gives interesting conditions on the supports of t and its Fourier transform $\hat{t}$. For translation, the corollary is equivalent to zero overlap of the compact supports of I and t. In particular using Theorem 5, for $I = t$, the maximal invariance implies the following localization conditions on t

$$\langle gt, t\rangle = 0 \quad g \notin G_L \subseteq G \quad \textbf{[17]}$$

which we call self-localization. For $1D$ translations it has the simple form $\langle T_x t, t\rangle = 0\ |x| > a,\ a > 0$.

For scaling we consider the support of the Fourier transforms of I and t. The Parseval theorem allows to rewrite the dot product $\langle D_s I, t\rangle$ which is in $L^2(\mathbb{R}^2)$ as $\left\langle \widehat{D_s I}, \hat{t} \right\rangle$ in the Fourier domain.

In the following we suppose that the support of $\hat{t}$ and $\hat{I}$ is respectively $[\omega^t_m, \omega^t_M]$ and $[\omega^I_m, \omega^I_M]$ where $\omega^{t,I}_m$ could be very close to zero (images and templates are supposed to be zero-mean) but usually are bigger then zero.

Note that the effect of scaling I with (typically $s = 2^j$ with $j \leq 0$) is to change the support as $supp(\widehat{D_s I}) = s(supp(\hat{I}))$. This change of the support of $\hat{I}$ in the dot product $\left\langle \widehat{D_s I}, \hat{t} \right\rangle$ gives non trivial conditions on the intersection with the support of $\hat{t}$ and therefore on the localization w.r.t. the scale invariance. We have the following Lemma:

Lemma 1. *Invariance to translation in the range* $[0, \bar{x}]$, $\bar{x} > 0$ *is equivalent to the following localization condition of* t *in space*

$$supp(t) \subseteq [-b - \bar{x}, b] - supp(I),\ I \in \mathcal{X}. \quad \textbf{[18]}$$

Separately, invariance to dilations in the range $[1, \bar{s}]$, $\bar{s} > 1$ *is equivalent to the following localization condition of* $\hat{t}$ *in frequency* ω

$$supp(\hat{t}) \subseteq [-\omega_t - \Delta^*_t, -\omega_t + \Delta^*_t] \cup [\omega_t - \Delta^*_t, \omega_t + \Delta^*_t]$$
$$\Delta^*_t = S\omega^I_m - \omega^I_M \frac{\bar{s}}{S},\ \ \omega_t = \frac{\omega^t_M - \omega^t_m}{2}. \quad \textbf{[19]}$$

Proof:

To prove that $supp(t) \subseteq [-b + \bar{x}, b] - supp(I)$ note that eq. **[15]** implies that $supp(\langle T_x I, t\rangle) \subseteq [-b + \bar{x}, b]$ (see Fig. 3, a)). Being $supp(\langle T_x I, t\rangle) = supp(I * t) \subseteq supp(I) + supp(t)$ we have $supp(t) \subseteq [-b - \bar{x}, b] - supp(I)$.

To prove the condition in eq. **[19]** note that eq. **[16]** is equivalent in the Fourier domain to

$$\langle D_s I, t\rangle = \left\langle \widehat{D_s I}, \hat{t} \right\rangle = \frac{1}{s} \int d\omega\ \hat{I}\left(\frac{\omega}{s}\right)\hat{t}(\omega) \neq 0\ \ \forall s \in [\frac{\bar{s}}{S}, S] \quad \textbf{[20]}$$

The situation is depicted in Fig. 3 b') for S big enough: in this case in fact we can suppose the support of $\widehat{D_{\bar{s}/S} I}$ to be on an interval on the left of that of $supp(\hat{t})$ and $\widehat{D_S I}$ on the

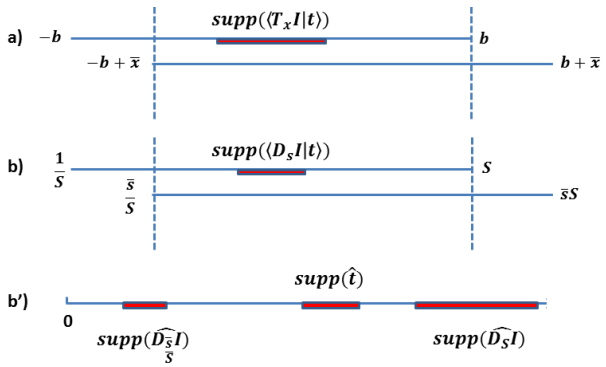


Fig. 3: *a), b): if the support of the dot product between the image and the template is contained in the intersection between the pooling range and the group translated (a) or dilated (b) pooling range the signature is invariant. In frequency condition b) becomes b'): when the Fourier supports of the dilated image and the template do not intersect their dot product is zero.*

right; the condition $supp(\langle \widehat{D_s I}, \hat{t} \rangle) \subseteq [\bar{s}/S, S]$ is in this case equivalent to

$$\omega_M^I \frac{\bar{s}}{S} < \omega_m^t, \;\; \omega_M^t < \omega_m^I S \qquad [21]$$

which gives

$$\Delta_t^* = Max(\Delta_t) \equiv Max\Big(\frac{\omega_M^t - \omega_m^t}{2}\Big) = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S} \qquad [22]$$

and therefore eq. [19].Q.E.D.
Note that for some $s \in [\bar{s}/S, S]$ the condition that the Fourier supports are disjoint is only sufficient and not necessary for the dot product to be zero since cancelations can occur. However to have $\langle \widehat{D_s I}, \hat{t} \rangle = 0$ on a continuous interval of scales, (unless some pathological examples of the function I) implies disjointness of the supports, since $\hat{I}(\omega/s) \neq \hat{I}(\omega/s'), \; s \neq s'$ unless I has constant spectrum in the interval $[\bar{s}/S, S]$. A similar reasoning is valid for the translation case.

The results above lead to a statement connecting *invariance* with *localization* of the templates:

Theorem 6. *Maximum translation invariance implies a template with minimum support in the space domain (x); maximum scale invariance implies a template with minimum support in the Fourier domain (ω).*

Proof:
We illustrate the statement of the theorem with a simple example. In the case of translations suppose, e.g., $supp(I) = [-b, b']$, $supp(t) = [-a, a]$, $a \leq b' \leq b$. Eq. [18] reads

$$[-a, a] \subseteq [-b + \bar{x} + b', b - b']$$

which gives the condition $-a \geq -b + b' + \bar{x}$, i.e. $\bar{x}^{max} = b - b' - a$; thus, for any fixed b, b' the smaller the template support $2a$ in space, the greater is translation invariance.
Similarly, in the case of dilations, increasing the range of invariance $[1, \bar{s}]$, $\bar{s} > 1$ implies a decrease in the support of $\hat{t}$ as shown by eq. [22]; in fact noting that $|supp(\hat{t})| = 2\Delta_t$ we have

$$\frac{d|supp(\hat{t})|}{d\bar{s}} = -\frac{2\omega_M^I}{S} < 0$$

i.e. the measure, $|\cdot|$, of the support of $\hat{t}$ is a decreasing function w.r.t. the measure of the invariance range $[1, \bar{s}]$. Q.E.D.

Because of the assumption of maximum possible support of all I being finite there is always localization for any choice of I and t under spatial shift. Of course if the localization support is larger than the pooling range there is no invariance.
For a complex cell with pooling range $[-b, b]$ in space only templates with self-localization smaller than the pooling range make sense. An extreme case of self-localization is $t(x) = \delta(x)$, corresponding to maximum localization of tuning of the simple cells.

Invariance, Localization and Wavelets. The conditions equivalent to optimal translation and scale invariance – maximum localization in space and frequency – cannot be simultaneously satisfied because of the classical *uncertainty principle*: if a function $t(x)$ is essentially zero outside an interval of length Δx and its Fourier transform $\hat{I}(\omega)$ is essentially zero outside an interval of length $\Delta\omega$ then

$$\Delta x \cdot \Delta\omega \geq 1. \qquad [23]$$

In other words a function and its Fourier transform cannot both be highly concentrated. Interestingly for our setup the uncertainty principle also applies to sequences (see [5]).

It is well known that the equality sign in the uncertainty principle above is achieved by Gabor functions (see [6]) of the form

$$\psi_{x_0,\omega_0}(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}, \; \sigma_x \in \mathbb{R}^+, \; \omega_0 \in \mathbb{R} \qquad [24]$$

The uncertainty principle leads to the concept of "optimal localization" instead of exact localization. In a similar way, it is natural to relax our definition of strict invariance (e.g. $\mu_n^k(I) = \mu_n^k(g'I)$) and to introduce ϵ-invariance as $\mu_n^k(I) - \mu_n^k(g'I) \leq \epsilon$. In particular if we suppose, e.g., the following localization condition

$$\langle T_x I, t \rangle = e^{-\frac{x^2}{\sigma_x^2}}, \quad \langle D_s I, t \rangle = e^{-\frac{s^2}{\sigma_s^2}}, \; \sigma_x, \sigma_s \in \mathbb{R} \qquad [25]$$

we have

$$\mu_n^k(T_{\bar{x}} I) - \mu_n^k(I) = \frac{1}{2}\sqrt{\sigma_x}\Big(\mathrm{erf}\big([-b, b]\Delta[-b + \bar{x}, b + \bar{x}]\big)\Big)$$
$$\mu_n^k(D_{\bar{s}} I) - \mu_n^k(I) = \frac{1}{2}\sqrt{\sigma_s}\Big(\mathrm{erf}\big([-1/S, S]\Delta[\bar{s}/S, S\bar{s}]\big)\Big).$$

where erf is the error function. The differences above, with an opportune choice of the localization ranges σ_s, σ_x can be made as small as wanted.
We end this paragraph by a conjecture: the optimal $\epsilon-$invariance is satisfied by templates with non compact support which decays exponentially such as a Gaussian or a Gabor wavelet. We can then speak of *optimal invariance* meaning "optimal ϵ-invariance". The reasonings above lead to the theorem:

Theorem 7. *Assume invariants are computed from pooling within a pooling window with a set of linear filters. Then the optimal templates (e.g. filters) for maximum simultaneous invariance to translation and scale are Gabor functions*

$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}. \qquad [26]$$

Remarks

1. The Gabor function $\psi_{x_0,\omega_0}(x)$ corresponds to a *Heisenberg box* which has a x-spread $\sigma_x^2 = \int x^2 |g(x)| dx$ and a ω spread $\sigma_\omega^2 = \int \omega^2 |\hat{g}(\omega)| d\omega$ with area $\sigma_x \sigma_\omega$. Gabor wavelets arise under the action on $\psi(x)$ of the translation and scaling groups as follows. The function $\psi(x)$, as defined, is zero-mean and normalized that is

$$\int \psi(x) dx = 0 \qquad [27]$$

and

$$||\psi(x)|| = 1. \qquad [28]$$

A family of Gabor wavelets is obtained by translating and scaling ψ:

$$\psi_{u,s}(x) = \frac{1}{s^{\frac{1}{2}}} \psi(\frac{x - u}{s}). \qquad [29]$$

Under certain conditions (in particular, the Heisenberg boxes associated with each wavelet must together cover the space-frequency plane) the Gabor wavelet family becomes a Gabor wavelet frame.

2. Optimal self-localization of the templates (which follows from localization), when valid simultaneously for space and scale, is also equivalent to Gabor wavelets. If they are a frame, full information can be preserved in an optimal quasi invariant way.

Approximate Invariance and Localization. In the previous section we analyzed the relation between localization and invariance in the case of group transformations. By relaxing the requirement of exact invariance and exact localization we show how the same strategy for computing invariants can still be applied even in the case of non-group transformations if certain localization properties of $\langle TI, t\rangle$ holds, where T is a smooth transformation.

We first notice that the localization condition of theorems 5 and 7 – when relaxed to approximate localization – takes the (e.g. for the $1D$ translations group) form $\left\langle I, T_x t^k \right\rangle < \delta \quad \forall x \ s.t. \ |x| > a$, where δ is small in the order of $1/\sqrt{n}$ (where n is the dimension of the space) and $\left\langle gI, t^k \right\rangle \approx 1 \quad \forall x \ s.t. \ |x| < a$.
We call this property *sparsity of I in the dictionary t^k under G*. This condition can be satisfied by templates that are similar to images in the set and are sufficiently "rich" to be incoherent for "small" transformations. Note that form the reasoning above the sparsity of I in t^k under G is expected to improve with increasing n and with noise-like encoding of I and t^k by the architecture.
Another important property of sparsity of I in t^k (in addition to allowing local approximate invariance to arbitrary transformations, see later) is *clutter-tolerance* in the sense that if n_1, n_2 are additive uncorrelated spatial noisy clutter $\left\langle I + n_1, gt^k + n_2 \right\rangle \approx \langle I, gt\rangle$.
Interestingly the *sparsity condition under the group* is related to associative memories for instance of the holographic type (see [8] and [9]). If the sparsity condition holds only for $I = t^k$ and for very small set of $g \in G$, that is, it has the form $\left\langle I, gt^k \right\rangle = \delta(g)\delta_{I,t^k}$ it implies strict memory-based recognition (see non-interpolating look-up table in the description of [10]) with inability to generalize beyond stored templates or views.

While the first regime – exact (or $\epsilon-$) invariance for generic images, yielding universal Gabor templates – applies to the first layer of the hierarchy, this second regime (sparsity) – approximate invariance for a class of images, yielding class-specific templates – is important for dealing with non-group transformations at the top levels of the hierarchy where receptive fields may be as large as the visual field.

Several interesting transformations do not have the group structure, for instance the change of expression of a face or the change of pose of a body. We show here that approximate invariance to transformations that are not groups can be obtained if the approximate localization condition above holds, and if the transformation can be locally approximated by a linear transformation, e.g. a combination of translations, rotations and non-homogeneous scalings, which corresponds to a locally compact group admitting a Haar measure.

Suppose, for simplicity, that the smooth transformation T, at least twice differentiable, is parametrized by the parameter $r \in \mathbb{R}$. We approximate its action on an image I with a Taylor series (around e.g. $r = 0$) as:

$$\begin{aligned} T_r(I) &= T_0(I) + \Big(\frac{dT}{dr}\Big)_{r=0}(I)r + R(I) \qquad [30] \\ &= I + \Big(\frac{dT}{dr}\Big)_{r=0}(I)r + R(I) \\ &= I + J^I(I)r + R(I) = [e + rJ^I](I) + R(I) \\ &= L^I_r(I) + R(I) \end{aligned}$$

where $R(I)$ is the reminder, e is the identity operator, J^I the Jacobian and $L^I_r = e + J^I r$ is a linear operator.
Let R be the range of the parameter r where we can approximately neglect the remainder term $R(I)$. Let L be the range of the parameter r where the scalar product $\langle T_r I, t\rangle$ is localized i.e. $\langle T_r I, t\rangle = 0, \ \forall r \notin L$. If $L \subseteq R$ we have

$$\langle T_r I, t\rangle \approx \left\langle L^I_r I, t \right\rangle, \qquad [31]$$

If the above linearization holds, we have the following:

Proposition 8. *Let $I, t \in H$ a Hilbert space, $\eta_n : \mathbb{R} \to \mathbb{R}^+$ a set of bijective (positive) functions and T a smooth transformation (at least twice differentiable) parametrized by $r \in \mathbb{R}$. Let $L = supp(\langle T_r I, t\rangle)$, P the pooling interval in the r parameter and $R \subseteq \mathbb{R}$ defined as above. If $L \subseteq P \subseteq R$ and*

$$\langle T_r I, t\rangle = 0, \ \forall r \in \mathbb{R}/(T_{\bar{r}} P \cap P)$$

then $\mu^k_n(T_{\bar{r}} I) = \mu^k_n(I)$.
Proof:
We have

$$\begin{aligned} \mu^k_n(T_{\bar{r}} I) &= \int_P dr \ \eta_n(\langle T_r T_{\bar{r}} I, t\rangle) = \int_P dr \ \eta_n(\left\langle L^I_r L^I_{\bar{r}} I, t \right\rangle) \\ &= \int_P dr \ \eta_n(\left\langle L^I_{r+\bar{r}} I, t \right\rangle) = \mu^k_n(I) \end{aligned}$$

where the last equality is true if $\langle T_r I, t\rangle = \left\langle L^I_r I, t\right\rangle = 0, \ r \in \mathbb{R}/(T_{\bar{r}} P \cap P)$. Q.E.D.

As an example, consider the transformation induced on the image plane by rotation in depth of a face: it can be decomposed into piecewise linear approximations around a small number of key templates, each one corresponding to a specific $3D$ rotation of a template face. Each key template corresponds to a complex cell containing as (simple cells) a number of observed transformations of the key template within a small range of rotations. Each key template corresponds to a different signature which is invariant only for rotations around its center. Notice that the form of the linear approximation or the number of key templates needed does not affect the algorithm or its implementation. The templates learned are used in the standard dot-product-and-pooling module. The choice of the key templates – each one corresponding to a complex cell, and thus to a signature component – is not critical, as long as there are enough of them. For one parameter groups, the key templates correspond to the knots of a piecewise linear spline approximation. Optimal placement of the centers – if desired – is a separate problem that we leave aside for now.

Summary of the argument: Different transformations can be classified in terms of invariance and localization.

Compact Groups: consider the case of a compact group transformation such as rotation in the image plane. A complex cell is invariant when pooling over all the templates which span the full group $\theta \in [-\pi, +\pi]$. In this case there is no restriction on which images can be used as templates: any template yields perfect invariance over the whole range of transformations (apart from mild regularity assumptions) and a

single complex cell pooling over all templates can provide a globally invariant signature.

Locally Compact Groups and Partially Observable Compact Groups: consider now the POG situation in which the pooling is over a subset of the group: (the POG case always applies to Locally Compact groups (LCG) such as translations). As shown before, a complex cell is partially invariant **if** the value of the dot-product between a template and its shifted template under the group falls to zero fast enough with the size of the shift relative to the extent of pooling. In the POG and LCG case, such partial invariance holds over a restricted range of transformations if the templates and the inputs have a *localization* property that implies wavelets for transformations that include translation and scaling.

General (non-group) transformations: consider the case of a smooth transformation which may not be a group. Smoothness implies that the transformation can be approximated by piecewise linear transformations, each centered around a template (the local linear operator corresponds to the first term of the Taylor series expansion around the chosen template). Assume – as in the POG case – a special form of *sparsity* – the dot-product between the template and its transformation fall to zero with increasing size of the transformation. Assume also that the templates transform as the input image. For instance, the transformation induced on the image plane by rotation in depth of a face may have piecewise linear approximations around a small number of key templates corresponding to a small number of rotations of a given template face (say at $\pm 30^o, \pm 90^o, \pm 120^o$). Each key template and its transformed templates within a range of rotations corresponds to complex cells (centered in $\pm 30^o, \pm 90^o, \pm 120^o$). Each key template, e.g. complex cell, corresponds to a different signature which is invariant only for that part of rotation. The strongest hypothesis is that there exist input images that are sparse w.r.t. templates of the same class – these are the images for which local invariance holds.

Remarks:

1. We are interested in two main cases of POG invariance:
   - partial invariance *simultaneously* to translations in x, y, scaling and possibly rotation in the image plane. This should apply to "generic" images. The signatures should ideally preserve full, locally invariant information. This first regime is ideal for the first layers of the multilayer network and may be related to Mallat's scattering transform, [7]. We call the sufficient condition for for LCG invariance here, *localization*, and in particular, *self-localization* given by Equation [**17**].
   - partial invariance to linear transformations for a subset of all images. This second regime applies to high-level modules in the multilayer network specialized for specific classes of objects and non-group transformations. The condition that is sufficient here for LCG invariance is given by Theorem 5 which applies only to a specific class of I. We prefer to call it *sparsity* of the images with respect to a set of templates.
2. For classes of images that are sparse with respect to a set of templates, the localization condition does not imply wavelets. Instead it implies templates that are
   - similar to a class of images so that $\langle I, g_0 t^k \rangle \approx 1$ and
   - complex enough to be "noise-like" in the sense that $\langle I, g t^k \rangle \approx 0$ for $g \neq g_0$.
3. Templates must transform similarly to the input for approximate invariance to hold. This corresponds to the assumption of a class-specific module and of a *nice object class* [11, 12].
4. For the localization property to hold, the image must be similar to the key template or contain it as a diagnostic feature (a sparsity property). It must be also quasi-orthogonal (highly localized) under the action of the local group.
5. For a general, non-group, transformation it may be impossible to obtain invariance over the full range with a single signature; in general several are needed.
6. It would be desirable to derive a formal characterization of the error in local invariance by using the standard module of dot-product-and-pooling, equivalent to a complex cell. The above arguments provide the outline of a proof based on local linear approximation of the transformation and on the fact that a local linear transformation is a LCG.

2. Hierarchical Architectures

So far we have studied the invariance, uniqueness and stability properties of signatures, both in the case when a whole group of transformations is observable (see [**5**] and [**6**]), and in the case in which it is only partially observable (see [**10**] and [**11**]). We now discuss how the above ideas can be iterated to define a multilayer architecture. Consider first the case when G is finite. Given a subset $G_0 \subset G$, we can associate a *window* gG_0 to each $g \in G$. Then, we can use definition [**10**] to define for each window a signature $\Sigma(I)(g)$ given by the measurements,

$$\mu_n^k(I)(g) = \frac{1}{|G_0|} \sum_{\bar{g} \in gG_0} \eta_n\Big(\big\langle I, \bar{g} t^k \big\rangle \Big).$$

Note that, for reasons that will be clear later, the average in the integral is done for transformed templates and not on transformed images. We will keep this form as the definition of signature. For fixed n, k, a set of measurements corresponding to different windows can be seen as a $|G|$ dimensional vector. A signature $\Sigma(I)$ for the whole image is obtained as *a signature of signatures*, that is, a collection of signatures $(\Sigma(I)(g_1), \ldots, \Sigma(I)(g_{|G|})$ associated to each window.
Since we assume that the output of each module is made zero-mean and normalized before further processing at the next layer, *conservation of information from one layer to the next requires saving the mean and the norm* at the output of each module at each level of the hierarchy.
We *conjecture* that the neural image at the first layer is uniquely represented by the final signature at the top of the hierarchy and the means and norms at each layer.

The above discussion can be easily extended to continuous (locally compact) groups considering,

$$\mu_n^k(I)(g) = \frac{1}{V_0} \int_{gG_0} d\bar{g} \eta_n\Big(\big\langle I, \bar{g} t^k \big\rangle \Big), \quad V_0 = \int_{G_0} d\bar{g},$$

where, for fixed n, k, $\mu_n^k(I) : G \to \mathbb{R}$ can now be seen as a function on the group. In particular, if we denote by $K_0 : G \to \mathbb{R}$ the indicator function on G_0, then we can write

$$\mu_n^k(I)(g) = \frac{1}{V_0} \int_G d\bar{g} K_0(\bar{g}^{-1} g) \eta_n\Big(\big\langle I, \bar{g} t^k \big\rangle \Big).$$

The signature for an image can again be seen as a collection of signatures corresponding to different windows, but in this case it is a function $\Sigma(I) : G \to \mathbb{R}^{NK}$, where $\Sigma(I)(g) \in \mathbb{R}^{NK}$,

is a signature corresponding to the window G_0 "centered" at $g \in G$.

The above construction can be iterated to define a hierarchy of signatures. Consider a sequence $G_1 \subset G_2, \ldots, \subset G_L = G$. For $h : G \to \mathbb{R}^p$, $p \in \mathbb{N}$ with an abuse of notion we let $gh(\bar{g}) = h(g^{-1}\bar{g})$. Then we can consider the following construction.

We call *complex cell operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $\mu_\ell(I) : G \to \mathbb{R}^{NK}$ where

$$\mu_\ell^{n,k}(I)(g) = \frac{1}{|G_\ell|} \sum_{\bar{g} \in gG_\ell} \eta_n \left(\nu_\ell^k(I)(\bar{g}) \right), \qquad [32]$$

and *simple cell operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $\nu_\ell(I) : G \to \mathbb{R}^K$

$$\nu_\ell^k(I)(g) = \left\langle \mu_{\ell-1}(I), g t_\ell^k \right\rangle \qquad [33]$$

with t_ℓ^k the k^{th} template at layer ℓ and $\mu_0(I) = I$. Several comments are in order:

- beside the first layer, the inner product defining the simple cell operator is that in $L^2(G) = \{h : G \to \mathbb{R}^{NK}, |\int dg |h(g)|^2 < \infty\}$;
- The index ℓ corresponds to different layers, corresponding to different subsets G_ℓ.
- At each layer a (finite) set of templates $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ ($\mathcal{T}_0 \subset \mathcal{X}$) is assumed to be available. For simplicity, in the above discussion we have assumed that $|\mathcal{T}_\ell| = K$, for all $\ell = 1, \ldots, L$. The templates at layer ℓ can be thought of as *compactly supported functions*, with support much smaller than the corresponding set G_ℓ. Typically templates can be seen as image patches in the space of complex operator responses, that is $t_\ell = \mu_{\ell-1}(\bar{t})$ for some $\bar{t} \in \mathcal{X}$.
- Similarly we have assumed that the number of non linearities η_n, considered at every layer, is the same.

Following the above discussion, the extension to continuous (locally compact) groups is straightforward. We collect it in the following definition.

Definition 1. (Simple and complex response) *For $\ell = 1, \ldots, L$, let $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ (and $\mathcal{T}_0 \subset \mathcal{X}$) be a sequence of template sets. The complex cell operator at layer ℓ maps an image $I \in \mathcal{X}$ to a function $c_\ell(I) : G \to \mathbb{R}^{NK}$; in components*

$$\mu_\ell^{n,k}(I)(g) = \frac{1}{V_\ell} \int d\bar{g} K_\ell(\bar{g}^{-1} g) \eta_n \left(\nu_\ell^k(I)(\bar{g}) \right), \; g \in G \qquad [34]$$

where K_ℓ is the indicator function on G_ℓ, $V_\ell = \int_{G_\ell} d\bar{g}$ and where

$$\nu_\ell^k(I)(g) = \left\langle \mu_{\ell-1}(I), g t_\ell^k \right\rangle, \quad g \in G \qquad [35]$$

($\mu_0(I) = I$) is the simple cell operator at layer ℓ that maps an image $I \in \mathcal{X}$ to a function $\nu_\ell(I) : G \to \mathbb{R}^K$.

Fig. 4: *Covariance: the response for an image I at position g is equal to the response of the group shifted image at the shifted position.*

Remark Note that eq. **[34]** can be written as:

$$\mu_\ell^{n,k}(I) = K_\ell * \eta_n(\nu_\ell^k(I)) \qquad [36]$$

where $*$ is the group convolution.

Property 1: covariance. We call the map Σ covariant iff

$$\Sigma(gI) = g^{-1}\Sigma(I), \quad \forall g \in G, I \in \mathcal{X}.$$

In the following we show the covariance property for the $\mu_1^{n,k}$ response (see Fig. 4). An inductive reasoning then can be applied for higher order responses. We assume that the architecture is isotropic in the relevant covariance dimension (this implies that all the modules in each layer should be identical with identical templates) and that there is a continuum of modules in each layer.

Proposition 9. *Let G a locally compact group and $\bar{g} \in G$. Let $\mu_1^{n,k}$ as defined in 34. Then $\mu_1^{n,k}(\tilde{g}I)(g) = \mu_1^{n,k}(I)(\tilde{g}^{-1}g)$.*

Proof:
Using the definition 34 we have

$$\begin{aligned}
\mu_1^{n,k}(\tilde{g}I)(g) &= \frac{1}{V_1} \int_G d\bar{g} K_1(\bar{g}^{-1}g) \eta_n \left(\left\langle \tilde{g}I, \bar{g}t^k \right\rangle \right) \\
&= \frac{1}{V_1} \int_G d\bar{g} K_1(\bar{g}^{-1}g) \eta_n \left(\left\langle I, \tilde{g}^{-1}\bar{g}t^k \right\rangle \right) \\
&= \frac{1}{V_1} \int_G d\hat{g} K_1(\hat{g}^{-1}\tilde{g}^{-1}g) \eta_n \left(\left\langle I, \hat{g}t^k \right\rangle \right) \\
&= \mu_1^{n,k}(I)(\tilde{g}^{-1}g)
\end{aligned}$$

where in the third line we used the change of variable $\hat{g} = \tilde{g}^{-1}\bar{g}$ and the invariance of the Haar measure. Q.E.D.

Remarks

1. The covariance property described in proposition 9 can be stated equivalently as $\mu_1^{n,k}(I)(g) = \mu_1^{n,k}(\bar{g}I)(\bar{g}g)$. This last expression has a more intuitive meaning as shown in Fig. 4.
2. The covariance property described in proposition 9 holds both for abelian and non-abelian groups. However the group average on templates transformations in definition of eq. 34 is crucial. In fact, if we define the signature averaging on the images we do not have a covariant response:

$$\begin{aligned}
\mu_1^{n,k}(\tilde{g}I)(g) &= \frac{1}{V_1} \int_G d\bar{g} K_1(\bar{g}^{-1}g) \eta_n \left(\left\langle \bar{g}\tilde{g}I, t^k \right\rangle \right) \\
&= \int_G d\hat{g} K_1(\tilde{g}\hat{g}^{-1}g) \eta_n \left(\left\langle \hat{g}I, t^k \right\rangle \right)
\end{aligned}$$

   where in the second line we used the change of variable $\hat{g} = \tilde{g}^{-1}\bar{g}$ and the invariance of the Haar measure. The last expression cannot be written as $\mu_1^{n,k}(I)(g'g)$ for any $g' \in G$.

3. With respect to the range of invariance, the following property holds for multilayer architectures in which the output of a layer is defined as covariant if it transforms in the same way as the input: for a given transformation of an image or part of it, the signature from complex cells at a certain layer is either invariant or covariant with respect to the group of transformations; if it is covariant there will be a higher layer in the network at which it is invariant

(more formal details are given in theorem 11), assuming that the image is contained in the visual field. This property predicts a *stratification* of ranges of invariance in the ventral stream: invariances should appear in a sequential order meaning that smaller transformations will be invariant before larger ones, in earlier layers of the hierarchy (see [13]).

Property 2: partial and global invariance (whole and parts). We now find the conditions under which the functions μ_ℓ are locally invariant, i.e. invariant within the restricted range of the pooling. We further prove that the range of invariance increases from layer to layer in the hierarchical architecture. The fact that for an image, in general, no more global invariance is guaranteed allows, as we will see, a novel definition of "parts" of an image.
The local invariance conditions are a simple reformulation of Theorem 5 in the context of a hierarchical architecture. In the following, for sake of simplicity we suppose that at each layer we only have a template t and a non linear function η.

Proposition 10. Invariance and Localization: hierarchy. *Let $I, t \in H$ a Hilbert space, $\eta : \mathbb{R} \to \mathbb{R}^+$ a bijective (positive) functions and G a locally compact group. Let $G_\ell \subseteq G$ and suppose $supp(\langle g\mu_{\ell-1}(I), t\rangle) \subseteq G_\ell$. Then for any given $\bar{g} \in G$*

$$\mu_\ell(I) = \mu_\ell(\bar{g}I) \;\Leftrightarrow\; \begin{array}{l} \langle g\mu_{\ell-1}(I), t\rangle = 0,\ g \in G/(G_\ell \cap \bar{g}G_\ell), \\ \langle g\mu_{\ell-1}(I), t\rangle \neq 0,\ g \in G_\ell \cap \bar{g}G_\ell. \end{array} \quad [37]$$

The proof follows the reasoning done in Theorem 5 with I substituted by $\mu_{\ell-1}(I)$ using the covariance property $\mu_{\ell-1}(gI) = g\mu_{\ell-1}(I)$. Q.E.D.
We can give now a formal definition of *object part* as the subset of the signal I whose complex response, at layer ℓ, is invariant under transformations in the range of the pooling at that layer.
This definition is consistent since the invariance is increasing from layer to layer (as formally proved below) therefore allowing bigger and bigger parts. Consequently for each transformation there will exists a layer $\bar{\ell}$ such that any signal subset will be a part at that layer. We can now state the following:

Theorem 11. Whole and parts. *Let $I \in \mathcal{X}$ (an image or a subset of it) and μ_ℓ the complex response at layer ℓ. Let $G_0 \subseteq \cdots \subset G_\ell \subseteq \cdots \subseteq G_L = G$ a set of nested subsets of the group G. Suppose η is a bijective (positive) function and that the template t and the complex response at each layer has finite support. Then $\forall \bar{g} \in G$, $\mu_\ell(I)$ is invariant for some $\ell = \bar{\ell}$, i.e.*

$$\mu_m(\bar{g}I) = \mu_m(I),\ \ \exists\ \bar{\ell}\ \ s.t.\ \ \forall m \geq \bar{\ell}.$$

The proof follows from the observation that the pooling range over the group is a bigger and bigger subset of G with growing layer number, in other words, there exists a layer such that the image and its transformations are within the pooling range at that layer (see Fig. 5). This is clear since for any $\bar{g} \in G$ the nested sequence

$$G_0 \cap \bar{g}G_0 \subseteq ... \subseteq G_\ell \cap \bar{g}G_\ell \subseteq ... \subseteq G_L \cap \bar{g}G_L = G.$$

will include a set $G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}}$ such that

$$\langle g\mu_{\bar{\ell}-1}(I), t\rangle \neq 0\ \ \forall g \in G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}}$$

being $supp(\langle g\mu_{\bar{\ell}-1}(I), t\rangle) \subseteq G$. Details are reported in [14].

Property 3: stability. Using the definition of stability given in [**9**], we can formulate the following theorem characterizing stability for the complex response:

Theorem 12. Stability. *Let $I, I' \in \mathcal{X}$ and μ_ℓ the complex response at layer l. Let the nonlinearity η a Lipschitz function with Lipschitz constant $L_\eta \leq 1$. Then*

$$\left\|\mu_\ell(I) - \mu_\ell(I')\right\| \leq \left\|I - I'\right\|,\ \forall\ \ell,\ \forall\ I, I' \in \mathcal{X}. \quad [38]$$

The proof follows from a repeated application of the reasoning done in Theorem 9. See details in [14].

Comparison with stability defined by Mallat [7]. The same definition of stability we use (Lipschitz continuity) was recently given by [7], in a related context. Let $I, I' \in L^2(\mathbb{R}^2)$ and $\Phi : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ a representation. Φ is stable if it is Lipschitz continuous with Lipschitz constant $L \leq 1$, i.e., is a non expansive map:

$$\left\|\Phi(I) - \Phi(I')\right\|_2 \leq \left\|I - I'\right\|_2,\ \ \forall\ I, I' \in L^2(\mathbb{R}^2). \quad [39]$$

In particular in [7] the author is interested in stability of group invariant scattering representations to the action of small diffeomorphisms close to translations. Consider transformations of the form $I'(\mathbf{x}) = \mathbf{L}_\tau \mathbf{I}(\mathbf{x}) = \mathbf{I}(\mathbf{x} - \tau(\mathbf{x}))$ (which can be though as small diffeomorphic transformations close to translations implemented by a displacement field $\tau : \mathbb{R}^2 \to \mathbb{R}^2$). A translation invariant operator Φ is said to be Lipschitz continuous to the action of a $C^2(\mathbb{R}^2)$ diffeomorphisms if for any compact $\Omega \subseteq \mathbb{R}^2$ there exists C such that for all $I \in L^2(\mathbb{R}^2)$ supported in $\Omega \subseteq \mathbb{R}^2$ and $\tau \in C^2(\mathbb{R}^2)$

$$\begin{array}{l} \left\|\Phi(I) - \Phi(L_\tau I)\right\|_2 \leq \\ \leq C \left\|I\right\|_2 \left(sup_{\mathbf{x}\in\mathbb{R}^2}|\nabla\tau(\mathbf{x})| + sup_{\mathbf{x}\in\mathbb{R}^2}|H\tau(\mathbf{x})|\right) \end{array} \quad [40]$$

where H is the Hessian and C a positive constant.
Condition [**40**] is a different condition then that in eq. [**38**] since it gives a Lipschitz bound for a diffeomorphic transformation at each layer of the scattering representation.
Our approach differs in the assumption that small (close to identity) diffeomorphic transformations can be well approximated, at the first layer, as locally affine transformations or, in the limit, as local translations which therefore falls in the POG case. This assumption is substantiated by the following reasoning in which any smooth transformation is seen as parametrized by the parameter t (the r parameter of the T_r transformation in section 1), which can be thought as time.
Let $T \subseteq \mathbb{R}$ be a bounded interval and $\Omega \subseteq \mathbb{R}^N$ an open set and let $\Phi = (\Phi_1, ..., \Phi_N) : T \times \Omega \to \mathbb{R}^N$ be $\mathcal{C}_2$ (twice differentiable), where $\Phi(0, .)$ is the identity map. Here $\mathbb{R}^N$ is

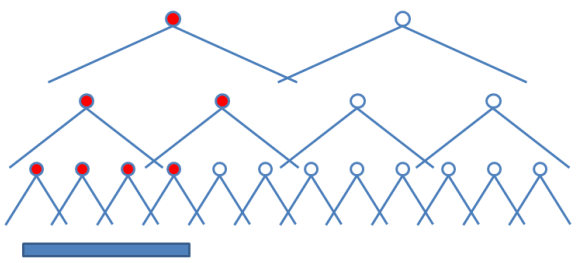


Fig. 5: *An image I with a finite support may or may not be fully included in the receptive field of a single complex cell at layer n (more in general the transformed image may not be included in the pooling range of the complex cell). However there will be a higher layer such that the support of its neural response is included in the pooling range of a single complex cell.*

assumed to model the image plane, intuitively we should take $N = 2$, but general values of N allow our result to apply in subsequent, more complex processing stages, for example continuous wavelet expansions, where the image is also parameterized in scale and orientation, in which case we should take $N = 4$. We write (t, x) for points in $T \times \Omega$, and interpret $\Phi(t, x)$ as the position in the image at time t of an observed surface feature which is mapped to $x = \Phi(0, x)$ at time zero. The map Φ results from the (not necessarily rigid) motions of the observed object, the motions of the observer and the properties of the imaging apparatus. The implicit assumption here is that no surface features which are visible in Ω at time zero are lost within the time interval T. The assumption that Φ is twice differentiable reflects assumed smoothness properties of the surface manifold, the fact that object and observer are assumed massive, and corresponding smoothness properties of the imaging apparatus, including eventual processing. Now consider a closed ball $B \subset \Omega$ of radius $\delta > 0$ which models the aperture of observation. We may assume B to be centered at zero, and we may equally take the time of observation to be $t_0 = 0 \in T$. Let

$$K_t = \sup_{(t,x)\in T\times B} \left\| \frac{\partial^2}{\partial t^2}\Phi(t,x) \right\|_{\mathbb{R}^N}, \; K_x = \sup_{x\in B} \left\| \frac{\partial^2}{\partial x \partial t}\Phi(0,x) \right\|_{\mathbb{R}^{N\times N}}$$

Here $(\partial/\partial x)$ is the spatial gradient in $\mathbb{R}^M$, so that the last expression is spelled out as

$$K_x = \sup_{x\in B} \left(\sum_{l=1}^{N} \sum_{i=1}^{N} \left(\frac{\partial^2}{\partial x_i \partial t}\Phi_l(0,x) \right)^2 \right)^{1/2}.$$

Of course, by compactness of $T \times B$ and the $\mathcal{C}_2$-assumption, both K_t and K_x are finite. The following theorem is due to Maurer and Poggio:

Theorem 13. *There exists $V \in \mathbb{R}^N$ such that for all $(t, x) \in T \times B$*

$$\|\Phi(t,x) - [x + tV]\|_{\mathbb{R}^N} \le K_x \delta |t| + K_t \frac{t^2}{2}.$$

The proof reveals this to be just a special case of Taylor's theorem.

Proof: Denote $V(t,x) = (V_1, ..., V_l)(t,x) = (\partial/\partial t)\Phi(t,x)$, $\dot{V}(t,x) = \left(\dot{V}_1, ..., \dot{V}_l\right)(t,x) = \left(\partial^2/\partial t^2\right)\Phi(t,x)$, and set $V := V(0,0)$. For $s \in [0,1]$ we have with Cauchy-Schwartz

$$\begin{aligned} \left\| \frac{d}{ds} V(0,sx) \right\|^2_{\mathbb{R}^N} &= \sum_{l=1}^{N} \sum_{i=1}^{N} \left(\left(\frac{\partial^2}{\partial x_i \partial t}\Phi_l(0,sx) \right) x_i \right)^2 \\ &\le K_x^2 \|x\|^2 \le K_x^2 \delta^2, \end{aligned}$$

whence

$$\begin{aligned} & \|\Phi(t,x) - [x + tV]\| \\ = & \left\| \int_0^t V(s,x)\, ds - tV(0,0) \right\| \\ = & \left\| \int_0^t \left[\int_0^s \dot{V}(r,x)\, dr + V(0,x) \right] ds - tV(0,0) \right\| \\ = & \left\| \int_0^t \int_0^s \frac{\partial^2}{\partial t^2}\Phi(r,x)\, drds + t \int_0^1 \frac{d}{ds} V(0,sx)\, ds \right\| \\ \le & \int_0^t \int_0^s \left\| \frac{\partial^2}{\partial t^2}\Phi(r,x) \right\| drds + |t| \int_0^1 \left\| \frac{d}{ds} V(0,sx) \right\| ds \\ \le & K_t \frac{t^2}{2} + K_x |t| \delta. \end{aligned}$$

Q.E.D.

Of course we are more interested in the visible features themselves, than in the underlying point transformation. If $I : \mathbb{R}^N \to \mathbb{R}$ represents these features, for example as a spatial distribution of gray values observed at time $t = 0$, then we would like to estimate the evolved image $I(\Phi(t,x))$ by a translate $I(x + tV)$ of the original I. It is clear that this is possible only under some regularity assumption on I. The simplest one is that I is globally Lipschitz. We immediately obtain the following

Corollary 14. *Under the above assumptions suppose that $I : \mathbb{R}^N \to \mathbb{R}$ satisfies*

$$|I(x) - I(y)| \le c \|x - y\|$$

for some $c > 0$ and all $x, y \in \mathbb{R}^N$. Then there exists $V \in \mathbb{R}^N$ such that for all $(t, x) \in I \times B$

$$|f(\Phi(t,x)) - f(x + tV)| \le c \left(K_x |t| \delta + K_t \frac{t^2}{2} \right).$$

Theorem 13 and corollary 14 gives a precise mathematical motivation for the assumption that any sufficiently smooth (at least twice differentiable) transformation can be approximated in an enough small compact set with a group transformation (e.g. translation), thus allowing, based on eq. 9, stability w.r.t. small diffeomorphic transformations.

Approximate Factorization: hierarchy. In the first version of [14] we conjectured that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g. the subgroup of translations of the affine group) and then adding the invariance w.r.t. another subgroup (e.g. rotations). While factorization of invariance ranges is possible in a hierarchical architecture (theorem 11), it can be shown that in general the factorization in successive layers for instance of invariance to translation followed by invariance to rotation (by subgroups) is impossible (see [14]). However, approximate factorization is possible under the same conditions of the previous section. In fact, a transformation that can be linearized piecewise can always be performed in higher layers, on top of other transformations, since the global group structure is not required but weaker smoothness properties are sufficient.

Why Hierarchical architectures: a summary.

1. *Optimization of local connections* and optimal reuse of computational elements. Despite the high number of synapses on each neuron it would be impossible for a complex cell to pool information across all the simple cells needed to cover an entire image.
2. *Compositionality.* A hierarchical architecture provides signatures of larger and larger patches of the image in terms of lower level signatures. Because of this, it can access memory in a way that matches naturally with the linguistic ability to describe a scene as a whole and as a hierarchy of parts.
3. *Approximate factorization.* In architectures such as the network sketched in Fig. 1 in the main text, approximate invariance to transformations specific for an object class can be learned and computed in different stages. This property may provide an advantage in terms of the sample complexity of multistage learning [15]. For instance, approximate class-specific invariance to pose (e.g. for faces) can be computed on top of a translation-and-scale-invariant representation [12]. Thus the implementation of invariance can, in some cases, be "factorized" into

different steps corresponding to different transformations. (see also [16, 17] for related ideas).

Probably all three properties together are the reason evolution developed hierarchies.

3. Synopsis of Mathematical Results

List of Theorems

- Orbits are equivalent to P_I and both are invariant and unique.

  Theorem 1. *The distribution P_I is invariant and unique i.e. $I \sim I' \Leftrightarrow P_I = P_{I'}$.*

- P_I can be estimated within ϵ in terms of 1D probability distributions of gI, t^k.

  Theorem 2. *Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then*

  $$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon,$$

  with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

- Invariance from a single image based on memory of template transformations. The simple property

  $$\left\langle gI, t^k \right\rangle = \left\langle I, g^{-1} t^k \right\rangle$$

  implies (for compact groups without any additional property) that the signature components $\mu_n^k(I) = \frac{1}{|G|} \sum_{g \in G} \eta_n \Big(\left\langle I, gt^k \right\rangle \Big)$, calculated on templates transformations are invariant that is $\mu_n^k(I) = \mu_n^k(\bar{g} I)$.
- Invariance for Partially Observable Groups (observed through a window) is equivalent to condition in eq. **[17]** on the dot product between image and template)

  Theorem 3. *Let $I, t \in H$ a Hilbert space, $\eta : \mathbb{R} \to \mathbb{R}^+$ a bijective (positive) function and G a locally compact group. Let $G_0 \subseteq G$ and suppose supp$(\langle gI, t \rangle) \subseteq G_0$. Then*

  $$\mu^t(I) = \mu^t(\bar{g} I) \Leftrightarrow \quad \begin{array}{l} \langle gI, t \rangle = 0, \ g \in G/(G_0 \cap \bar{g} G_0) \\ \langle gI, t \rangle \neq 0, \ g \in G_0 \cap \bar{g} G_0 \end{array}$$

- Condition in **[17]** is equivalent to a localization or sparsity property of the dot product between image and template ($\langle I, gt \rangle = 0$ for $g \notin G_L$). In particular

  Proposition 4. *Localization is necessary and sufficient for translation and scale invariance. Localization for translation (respectively scale) invariance is equivalent to the support of t being small in x (respectively in ω).*

- Optimal simultaneous invariance to translation and scale can be achieved by Gabor templates.

  Theorem 5. *Assume invariants are computed from pooling within a pooling window a set of linear filters. Then the optimal templates of filters for maximum simultaneous invariance to translation and scale are Gabor functions $t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$.*

- Approximate invariance can be obtained if there is approximate sparsity of the image in the dictionary of templates. Approximate localization (defined as $\langle t, gt \rangle < \delta$ for $g \notin G_L$, where δ is small in the order of $\approx \frac{1}{\sqrt{d}}$ and $\langle t, gt \rangle \approx 1$ for $g \in G_L$) is satisfied by templates (vectors of dimensionality n) that are similar to images in the set and are sufficiently "large" to be incoherent for "small" transformations.
- Approximate invariance for smooth (non group) transformations.

  Proposition 6. *$\mu^k(I)$ is locally invariant* **if**

  - *I is sparse in the dictionary t^k;*
  - *I and t^k transform in the same way (belong to the same class);*
  - *the transformation is sufficiently smooth.*

- Sparsity of I in the dictionary t^k under G increases with size of the neural images and provides invariance to clutter. The definition is $\langle I, gt \rangle < \delta$ for $g \notin G_L$, where δ is small in the order of $\approx \frac{1}{\sqrt{n}}$ and $\langle I, gt \rangle \approx 1$ for $g \in G_L$.
  Sparsity of I in t^k under G improves with dimensionality of the space n and with noise-like encoding of I and t.
  If n_1, n_2 are additive uncorrelated spatial noisy clutter $\langle I + n_1, gt + n_2 \rangle \approx \langle I, gt \rangle$.
- Covariance of the hierarchical architecture.

  Proposition 7. *The operator μ_ℓ is covariant with respect to a non abelian (in general) group transformation, that is*

  $$\mu_\ell(T_g I) = T_g \mu_\ell(I).$$

- Factorization:invariance to separate subgroups of affine group cannot be obtained in a sequence of layers while factorization of the ranges of invariance can (because of covariance). Invariance to a smooth (non group) transformation can always be performed in higher layers, on top of other transformations, since the global group structure is not required.
- Uniqueness of signature. **Conjecture:***the neural image at the first layer is uniquely represented by the final signature at the top of the hierarchy and the means and norms at each layer.*

4. General Remarks on the Theory

1. The second regime of localization (sparsity) can be considered as a way to deal with situations that do not fall under the general rules (group transformations) by creating a series of exceptions, one for each object class.
2. Whereas the first regime "predicts" Gabor tuning of neurons in the first layers of sensory systems, the second regime predicts cells that are tuned to much more complex features, perhaps similar to neurons in inferotemporal cortex.
3. The *sparsity condition under the group* is related to properties used in associative memories for instance of the holographic type (see [8]). If the sparsity condition holds only for $I = t^k$ and for very small a then it implies strictly memory-based recognition.
4. The theory is memory-based. It also view-based. Even assuming 3D images (for instance by using stereo information) the various stages will be based on the use of 3D views and on stored sequences of 3D views.

5. The mathematics of the class-specific modules at the top of the hierarchy – with the underlying localization condition – justifies old models of viewpoint-invariant recognition (see [18]).
6. The remark on factorization of general transformations implies that layers dealing with general transformations can be on top of each other. It is possible – as empirical results by Leibo and Li indicate – that a second layer can improve the invariance to a specific transformation of a lower layer.
7. The theory developed here for vision also applies to other sensory modalities, in particular speech.
8. The theory represents a general framework for using representations that are invariant to transformations that are learned in an unsupervised way in order to reduce the sample complexity of the supervised learning step.
9. Simple cells (e.g. templates) under the action of the affine group span a set of positions and scales and orientations. The size of their receptive fields therefore spans a range. The pooling window can be arbitrarily large – and this does not affect selectivity when the CDF is used for pooling. A large pooling window implies that the signature is given to large patches and the signature is invariant to uniform affine transformations of the patches within the window. A hierarchy of pooling windows provides signature to patches and subpatches and more invariance (to more complex transformations).
10. Connections with the *Scattering Transform*.
    - Our theorems about optimal invariance to scale and translation implying Gabor functions (first regime) may provide a justification for the use of Gabor wavelets by Mallat [7], that does not depend on the specific use of the modulus as a pooling mechanism.
    - Our theory justifies several different kinds of pooling of which Mallat's seems to be a special case.
    - With the choice of the modulo as a pooling mechanisms, Mallat proves a nice property of Lipschitz continuity on diffeomorphisms. Such a property is not valid *in general* for our scheme where it is replaced by a hierarchical *parts and wholes* property which can be regarded as an approximation, as refined as desired, of the continuity w.r.t. diffeomorphisms.
    - Our second regime does not have an obvious corresponding notion in the scattering transform theory.
11. The theory characterizes under which conditions the signature provided by a HW module at some level of the hierarchy is invariant and therefore could be used for retrieving information (such as the label of the image patch) from memory. The simplest scenario is that signatures from modules at all levels of the hierarchy (possibly not the lowest ones) will be checked against the memory. Since there are of course many cases in which the signature will not be invariant (for instance when the relevant image patch is larger than the receptive field of the module) this scenario implies that the step of memory retrieval/classification is selective enough to discard efficiently the "wrong" signatures that do not have a match in memory. This is a non-trivial constraint. It probably implies that signatures at the top level should be matched first (since they are the most likely to be invariant and they are fewer) and lower level signatures will be matched next possibly constrained by the results of the top-level matches – in a way similar to *reverse hierarchies* ideas. It also has interesting implications for appropriate encoding of signatures to make them optimally quasi-orthogonal e.g. incoherent, in order to minimize memory interference. These properties of the representation depend on memory constraints and will be object of a future paper on memory modules for recognition.

5. Empirical support for the theory

Several computational vision models in recent literature can be considered instances of the theory described here. HMAX, trained convolutional networks, and the feedforward networks

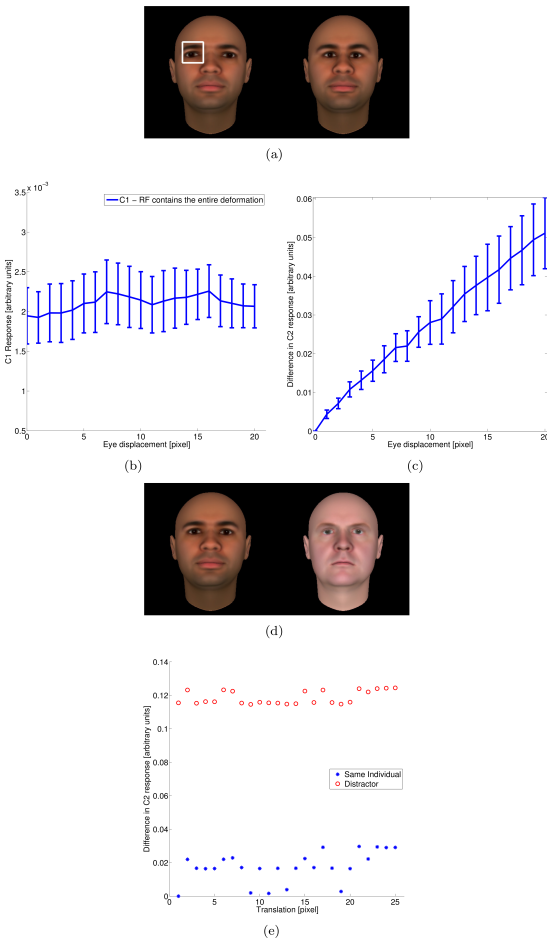


Fig. 6: *Empirical demonstration of the properties of invariance, stability and uniqueness of the hierarchical architecture (see Theorem 12) in a specific 2 layers implementation (HMAX). Inset (a) shows the reference image on the left and a deformation of it (the eyes are closer to each other) on the right; (b) shows that an HW-module in layer 1 whose receptive fields covers the left eye provides a signature vector (c_1) which is invariant to the deformation; in (c) an HW-module at layer 2 (c_2) whose receptive fields contain the whole face provides a signature vector which is (Lipschitz) stable with respect to the deformation. In all cases, the Figure shows just the Euclidean norm of the signature vector. Notice that the c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. Two different images (d) are presented at various location in the visual field. The Euclidean distance between the signatures of a set of HW-modules at layer 2 with the same receptive field (the whole image) and a reference vector is shown in (e). The signature vector is invariant to global translation and discriminative (between the two faces). In this example the HW-module represents the top of a hierarchical, convolutional architecture. The images we used were 200×200 pixels*

of N. Pinto et al. all consist of hierarchically stacked modules of simple and complex cells. However, only the most recent of these – variants of HMAX that incorporate invariances to complex transformations learned from video – have been designed with this theory explicitly in mind.

In [12], we showed that our approach of pooling over stored views of template faces undergoing the transformation can be used to recognize novel faces robustly to rotations in depth from a single example view. More recently, we applied the same idea to unconstrained face recognition benchmarks: Labeled Faces in the Wild and PubFig83, and showed that they yield a system that performs comparably to the state of the art with considerably less engineering.

In versions of HMAX developed prior to this theory, and in some related models, rather than arbitrary invariances being learned from video, specific invariances to local translation (and sometimes scaling) are built in to the architecture. A convolutional architecture which *by design* computes responses to the same set of templates at every position (and scale) is equivalent to a model which *learned* to do this by seeing videos of each template object translating (and scaling) through every position.

The best-performing version of HMAX for generic object categorization is an improved version of [19] which scores 74% on the Caltech 101 dataset, competitive with the state-of-the-art for a single feature type. The original version achieved a near-perfect score on the UIUC car dataset. Another HMAX variant added a time dimension for action recognition [20], outperforming both human annotators and a state-of-the-art commercial system on a mouse behavioral phenotyping task. An HMAX model [21] was also shown to account for human performance in rapid scene categorization.

One of the observations that inspired our theory is that in convolutional architectures, random features perform nearly as well as features learned from objects [22, 23]. This includes models other than HMAX: [24] found that a convolutional network with randomized weights performed only 3% worse than the same network after training via backpropagation. [25] also found feature learning to be the least significant of several variables contributing to the performance of a hierarchical architecture.

A simple illustrative empirical demonstration of the HMAX properties of invariance, stability and uniqueness is in Fig. 6.

6.Unsupervised learning of the template orbit

While the templates need not be related to the test images (in the affine case), during development, the model still needs to observe the orbit of some templates. We conjectured that this could be done by unsupervised learning based on the temporal adjacency assumption [26, 27]. One might ask, do "errors of temporal association" happen all the time over the course of normal vision? Lights turn on and off, objects are occluded, you blink your eyes – all of these should cause errors. If temporal association is really the method by which all the images of the template orbits are associated with one another, why doesn't the fact that its assumptions are so often violated lead to huge errors in invariance?

The full orbit is needed, at least in theory. In practice we have found that significant scrambling is possible as long as the errors are not correlated. That is, normally an HW-module would pool all the $\langle I, g_i t^k \rangle$. We tested the effect of, for some i, replacing t^k with a different template $t^{k'}$. Even scrambling 50% of our model's connections in this manner only yielded very small effects on performance. These experiments were described in more detail in [28] for the case of translation. In that paper we modeled Li and DiCarlo's "invariance disruption" experiments in which they showed that a temporal association paradigm can induce individual IT neurons to change their stimulus preferences under specific transformation conditions [29, 30]. We also report similar results on another "non-uniform template orbit sampling" experiment with 3D rotation-in-depth of faces in [31].

1. H. Cramer and H. Wold. Some theorems on distribution functions. J. London Math. Soc., 4:290–294, 1936.
2. J. Cuesta-Albertos, R. Fraiman, and R. T. A sharp form of the cramer–wold theorem. Journal of Theoretical Probability, 20:201–209, 2007.
3. J. Cuesta-Albertos. How many random projections suffice to determine a probability distribution? IPMs sections, 2009.
4. A. Heppes. On the determination of probability distributions of more dimensions by their projections. Acta Mathematica Hungarica, 7(3):403–410, 1956.
5. D. L. Donoho, P. B. Stark Uncertainty principles and signal recovery SIAM J. Appl. Math.,49 ,3 ,906–931 , 1989
6. D. Gabor. Theory of communication. Part 1: The analysis of information. Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of,93 ,26 ,429-441 ,1946.
7. S. Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65(10):1331–1398, 2012.
8. T. Poggio. On optimal nonlinear associative recall. Biological Cybernetics, 19(4):201–209, 1975.
9. T. Plate, Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations, International Joint Conference on Artificial Intelligence, 30-35, 1991.
10. T. Poggio A theory of how the brain might work Cold Spring Harb Symp Quant Biol, 1990.
11. T. Poggio, T. Vetter, and M. I. O. T. C. A. I. LAB. Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries, 1992.
12. J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In Advances in Neural Information Processing Systems (NIPS), Granada, Spain, 2011.
13. L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio. The timing of invariant object recognition in the human visual system. Submitted, 2013.
14. F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio. Magic Materials: a theory of deep hierarchical architectures for learning sensory representations CBCL paper, Massachusetts Institute of Technology, Cambridge, MA, April 1, 2013.
15. T. Poggio and S. Smale. The mathematics of learning: Dealing with data. Notices of the American Mathematical Society (AMS), 50(5):537–544, 2003.
16. D. Arathorn. Computation in the higher visual cortices: Map-seeking circuit theory and application to machine vision. In Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop, AIPR '04, pages 73–78, Washington, DC, USA, 2004. IEEE Computer Society.
17. L. Sifre, S. Mallat, and P. France. Combined scattering for rotation invariant texture, 2012.
18. T. Poggio, S. Edelmann A network that learns to recognize three-dimensional objects. In Nature, 1990 Jan 18, 343(6255):263266.
19. J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. Computer Vision and Pattern Recognition 2006, 1:11–18, 2006.
20. H. Jhuang, E. Garrote, J. Mutch, X. Yu, V. Khilnani, T. Poggio, A. Steele and T. Serre, Automated home-cage behavioural phenotyping of mice Nature Communications, 1, 68, doi:10.1038/ncomms1064, 2010.
21. T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15):6424–6429, 2007.
22. J.Z. Leibo, J. Mutch, L. Rosasco, S. Ullman, T. Poggio, Learning Generic Invariances in Object Recognition: Translation and Scale MIT-CSAIL-TR-2010-061, CBCL-294, 2010
23. A. Saxe, P.W. Koh, Z. Chen, M. Bhand, B. Suresh, A. Ng, On Random Weights and Unsupervised Feature Learning, Proceedings of the 28th International Conference on Machine Learning (ICML-11), 1089-1096, 2011.
24. K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y . LeCun. What is the best multi-stage architecture for object recognition? IEEE International Conference on Computer Vision, 2146-2153, 2009.
25. N. Pinto, D. Doukhan, J.J. DiCarlo, D.D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Computational Biology, 5, 2009.
26. P. Földiák. Learning invariance from transformation sequences. Neural Computation, 3(2):194–200, 1991.
27. L. Wiskott, T.J. Sejnowski Slow feature analysis: Unsupervised learning of invariances Neural computation, 4, 14, 715-770, 2002.
28. L. Isik, J.Z. Leibo, T. Poggio Learning and disrupting invariance in visual recognition with a temporal association rule Frontiers in Computational Neuroscience,2,2012
29. N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321(5895):1502–7, Sept. 2008.
30. N. Li and J. J. DiCarlo. Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. Neuron, 67(6):1062–1075, 2010.
31. Q. Liao, J.Z. Leibo, T. Poggio. Learning invariant representations and applications to face verification Under revision.

A theory of deep learning architectures for sensory perception: the ventral stream.

28 April 2013

DRAFT[1]

Tomaso Poggio⋆,†, Jim Mutch⋆, Fabio Anselmi†, Lorenzo Rosasco†, Joel Z Leibo⋆, Andrea Tacchetti†
⋆ *CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*
† Istituto Italiano di Tecnologia, Genova, Italy

[1] **Online archived report: historical notes.** This is version 3.1 of a report first published online on July 20, 2011 (npre.2011.6117.1); it replaces version 3.0 which appeared on December 30th, 2012, as a CSAIL technical report.

Abstract

This paper explores the theoretical consequences of a simple assumption: the computational goal of the feedforward path in the ventral stream – from V1, V2, V4 to IT – is to discount image transformations, after learning them during development.

Part I assumes that a *basic neural operation* consists of dot products between input vectors and synaptic weights – which can be modified by learning. It proves that a multi-layer hierarchical architecture of dot-product modules can learn in an unsupervised way geometric transformations of images and then achieve the dual goals of invariance to global affine transformations and of stability. The basic module, which estimates a unique, invariant signature, has a surprisingly elegant implementation in terms of idealized simple and complex cells, which are predicted to perform respectively (random) projections and group averages, followed by a sigmoidal nonlinearity. These hierarchical architectures learn in an unsupervised way to be automatically invariant to transformations of a new object, achieving the goal of recognition with one or very few labeled examples. The theory of Part I should apply to a varying degree to a range of hierarchical architectures such as HMAX, convolutional networks and related feedforward models of the visual system and formally characterize some of their properties.

A *linking conjecture* in Part II assumes that storage of transformed templates during development – a stage implied by the theory of Part I – takes place via Hebbian-like developmental learning at the synapses in visual cortex. It follows that the cells' tuning will effectively converge during development to the top eigenvectors of the covariance of their inputs. The solution of the associated eigenvalue problem is surprisingly tolerant of details of the image spectrum. It predicts quantitative properties of the tuning of cells in the first layer – identified with simple cells in V1; in particular, they should converge during development to oriented Gabor-like wavelets with frequency inversely proportional to the size of an elliptic Gaussian envelope – in agreement with data from the cat, the macaque and the mouse. A similar analysis leads to predictions about receptive field tuning in higher visual areas – such as V2 and V4 – and in particular about the size of simple and complex receptive fields in each of the areas.

For non-affine transformations of the image – for instance induced by out-of-plane rotations of a 3D object or non-rigid deformations – it is possible to prove that the dot-product module of Part I can provide *approximate* invariance for certain classes of objects. Thus Part III considers modules that are class-specific – such as the face, the word and the body area – and predicts several properties of the macaque cortex face patches characterized by Freiwald and Tsao, including a patch (called AL) which contains mirror symmetric cells and is the input to the pose-invariant patch (AM).

Taken together, the results of the papers suggest a computational role for the ventral stream and derive detailed properties of the architecture and of the tuning of cells, including the role and quantitative properties of neurons in $V1$. A surprising implication of these theoretical results is that the computational goals and several of the tuning properties of cells in the ventral stream may follow from *symmetry properties* (in the sense of physics) of

the visual world through a process of unsupervised correlational learning, based on Hebbian synapses.

Contents

1 Summary **9**

2 Introduction **12**
2.1 Plan of the paper 12

3 Part I: Memory-based Learning of Invariance to Transformations **17**
3.1 Recognition is difficult because of image transformations 17
3.1.1 Suggestive empirical evidence 17
3.1.2 Intraclass and viewpoint complexity 20
3.1.3 Invariant representations and bounds on learning rates . 21
3.2 Templates and signatures 21
3.2.1 Preliminaries: resolution and size 22
3.2.2 Templatesets 24
3.2.3 Transformations and templatebooks 27
3.3 Invariance and discrimination 28
3.3.1 The invariance lemma 28
3.3.2 Orbits 30
3.4 Invariant and unique signatures 31
3.4.1 Orbits and probability distributions 31
3.4.2 Empirical measurements of probabilities of projections . 32
3.4.3 Computations by simple and complex cells 35
3.4.4 A theory of pooling 35
3.4.5 Stable signatures 36
3.4.6 Signatures for Partially Observable Groups (POG): Invariance, Uniqueness and Stability 38
3.4.7 Approximate Invariance of Signatures associated to POG 39
3.4.8 Uniqueness and Stability for POG signatures 40
3.5 Hierarchical architectures 40
3.5.1 The basic idea: wholes and parts 42
3.5.2 Hierarchical Architectures 44
3.5.3 Property 1 :covariance of the c_ℓ response 47
3.5.4 Property 2: partial and global invariance of c_ℓ response (whole and parts) 48
3.5.5 Property 3: stability of the c_ℓ response 50
3.5.6 A hierarchical architecture: locally compact groups . . . 52
3.6 Factorization of Invariances 53
3.7 Preliminaries 53
3.8 Factorization of transformations 54
3.9 Factorization of Invariance ranges 56
3.10 Approximate Factorization for Invariances in Object Classes . . 57
3.11 A mathematical summary (*incomplete*) 58

4 Part II: Learning Transformations and Spectral Properties 60
4.1 Apertures and Stratification . . . 60
4.1.1 Translation approximation for small apertures . . . 61
4.2 Linking conjecture: developmental memory is Hebbian . . . 64
4.2.1 Hebbian synapses and Oja flow . . . 64
4.3 Spectral properties of the templatebook covariance operator: cortical equation . . . 67
4.3.1 Eigenvectors of the covariance of the template book for the translation group . . . 70
4.4 Retina to V1: processing pipeline . . . 74
4.4.1 Spatial and temporal derivatives in the retina . . . 75
4.5 Cortical equation: predictions for simple cells in V1 . . . 76
4.6 Complex cells: wiring and invariance . . . 90
4.6.1 Complex cells invariance properties: mathematical description . . . 90
4.6.2 Hierarchical frequency remapping . . . 91
4.7 Beyond V1 . . . 92
4.7.1 Almost-diagonalization of non commuting operators . . . 92
4.7.2 Independent shifts and commutators . . . 93
4.7.3 Hierarchical wavelets: 4-cube wavelets . . . 93
4.7.4 Predictions for V2, V4, IT . . . 94

5 Part III: Class-specific Transformations and Modularity 99
5.1 Approximate invariance to non-generic transformations . . . 99
5.2 3D rotation is class-specific . . . 99
5.2.1 The 2D transformation . . . 101
5.2.2 An approximately invariant signature for 3D rotation . . . 102
5.3 Empirical results on class-specific transformations . . . 103
5.4 The macaque face-processing network . . . 107
5.4.1 Principal components and mirror-symmetric tuning curves 110
5.4.2 Models of the macaque face recognition hierarchy . . . 111
5.5 Other class-specific transformations: bodies and words . . . 111
5.6 Invariance to X and estimation of X . . . 118

6 Discussion 120
6.1 Some of the main ideas . . . 120
6.2 Extended model and previous model . . . 123
6.3 What is under the carpet . . . 124
6.4 Directions for future research . . . 125
6.4.1 Associative memories . . . 125
6.4.2 Invariance and Perception . . . 126
6.4.3 The dorsal stream . . . 127
6.4.4 Visual "concepts" . . . 127
6.4.5 Is the ventral stream a cortical mirror of the invariances of the physical world? . . . 127

7 Appendix: background from previous work **136**

8 Appendix: memory-based model and invariance **137**
8.1 Invariance lemma . . . 137
8.1.1 Old, original version of the Invariance Lemma . . . 138
8.1.2 Example: affine group and invariance . . . 139
8.1.3 More on Group Averages . . . 140
8.1.4 More on Templatebooks . . . 141
8.2 More on groups and orbits . . . 141

9 Appendix: stratification lemma **142**
9.1 Subgroups and invariances factorization . . . 142
9.2 Factorization of Invariances and Hierarchies . . . 143
9.3 Transformations: Stratification and Peeling Off . . . 144
9.3.1 Class-specific transformations . . . 145

10 Appendix: invariant discriminability **145**
10.1 Premise . . . 145
10.1.1 Basic Framework . . . 146
10.2 Similarity Among Orbits . . . 146
10.3 (Group) Invariance . . . 147
10.4 Discrimination . . . 148
10.4.1 (Non Linear) Measurements . . . 148
10.4.2 Pooling functions as moments . . . 149
10.4.3 Abstract measurements of probability distributions . . . 150
10.4.4 Comparing probability distributions . . . 152
10.4.5 Moments . . . 152
10.5 Random projections and invariants: an extension of J-L . . . 152
10.5.1 Frames and random projections . . . 153
10.6 Discriminability, invariance and robustness to diffeomorphisms 154
10.6.1 Templates and diffeomorphisms: from global to local . . 154
10.7 Complex cells invariance: $SIM(2)$ group . . . 157

11 Appendix: whole and parts **160**

12 Appendix: hierarchical frequency remapping **161**
12.1 Information in bandpass signals . . . 162
12.2 More on uniqueness of modulo square . . . 163
12.2.1 Information can be preserved . . . 163
12.2.2 Another approach: direct wavelet reconstruction from modulus square . . . 164
12.3 Predicting the size of the receptive field of simple and complex cells . . . 166

13 Appendix: hierarchical representation and computational advantages 168
13.1 Memory . . . 168
13.2 Higher order features . . . 168

14 Appendix: apertures and transformations **168**
14.1 Stratification . . . 168
14.1.1 Commutativity . . . 171
14.2 Local approximation of global diffeomorphisms . . . 173
14.2.1 Method I . . . 173
14.2.2 Method II . . . 174

15 Appendix: differential equation **175**
15.1 Derivation and solution . . . 175
15.1.1 Case: $1/\omega$ spectrum . . . 176
15.1.2 Aperture ratio . . . 177
15.1.3 Initial conditions . . . 177
15.1.4 Two dimensional problem . . . 177
15.1.5 Derivative in the motion direction . . . 178
15.1.6 Fisher information and templatebook eigenfunctions . . . 178
15.2 Special case: the covariance $t^{\circledast}(x)$ consists of two Fourier components . . . 178
15.3 Continuous spectrum: an alternative derivation of the differential equation . . . 179
15.3.1 Numerical and analytical study . . . 184
15.3.2 Perturbative methods for eq. (139) . . . 186

16 Appendix: spectral properties of the templatebook **186**
16.1 Spectral properties of the translation operator . . . 186
16.1.1 Spectral properties of the uniform scaling and rotation operators . . . 186
16.2 Single value decomposition of compact operators . . . 187
16.3 Wavelet transform and templatebook operator . . . 187
16.4 Fourier Transform on a compact group . . . 188
16.5 Diagonalizing the templatebook . . . 189
16.6 The choice of the square integrable function t . . . 190
16.7 Diagonalizing the templatebook with different templates . . . 190
16.8 Temporal and spatial filtering in the retina and LGN . . . 190
16.8.1 Phase distribution of Gabor tuning functions . . . 191
16.8.2 Retinal filtering . . . 191
16.9 Optimizing signatures: the antislowness principle . . . 191
16.9.1 Against a "naive slowness" principle . . . 191
16.9.2 Our selection rule . . . 191

17 Appendix: blue-sky ideas and remarks **198**
17.1 Visual abstractions . . . 198
17.2 Invariances and constraints . . . 199
17.3 Remarks and open problems . . . 199

18 Background Material: Groups **202**
18.1 What is a group? 202
18.2 Group representation 203
18.3 Few more definitions 203
18.4 Affine transformations in $\mathbb{R}^2$. 204
18.5 Similitude transformations in $\mathbb{R}^2$ 204
18.5.1 Discrete subgroups: any lattice is locally compact abelian 204
18.6 Lie algebra associated with the affine group 205
18.6.1 Affine group generators 205
18.6.2 Lie algebra generators commutation relations 206
18.6.3 Associated characters 206
18.6.4 Mixing non commuting transformations 206

19 Background Material: Frames, Wavelets **207**
19.1 Frames 207
19.2 Gabor and wavelet frames 207
19.3 Gabor frames 208
19.4 Gabor wavelets 208
19.5 Lattice conditions 208

20 Background Material: Hebbian Learning **209**
20.1 Oja's rule 209
20.1.1 Oja's flow and receptive field aperture 210
20.2 Foldiak trace rule: simple and complex cells 210

1 Summary

The starting assumption in the paper is that the sample complexity of (biological, feedforward) object recognition is mostly due to geometric image transformations. Thus our main conjecture is that the computational goal of the feedforward path in the ventral stream – from $V1$, $V2$, $V4$ and to IT – is to discount image transformations after learning them during development. A complementary assumption is about the basic biological computational operation: we assume that

- *dot products* between input vectors and stored templates (synaptic weights) are the basic operation (dot product in most of the paper refers to *normalized dot product*: the normalization is often hidden in the equations;
- *memory* is stored in the synaptic weights through a Hebbian-like rule.

Part I of the paper describes a class of biologically plausible memory-based modules that learn transformations from unsupervised visual experience. The idea is that neurons can store during development "neural frames", that is image patches' of an object transforming – for instance translating or looming. After development, the main operation consists of dot-products of the stored templates with a new image. The dot-products are followed by a transformations-average operation, which can be described as pooling. The main theorems show that this 1-layer module provides (from a single image of any new object) a *signature* which is automatically invariant to global affine transformations and approximately invariant to other transformations. These results are derived in the case of random templates, using the Johnson-Lindenstrauss lemma in a special way; they are also valid in the case of sets of basis functions which are a frame. Surprisingly the theory predicts that the group average by complex cells should be followed by sigmoids and then by linear combination of a bank of complex cells. This one-layer architecture, though invariant, and optimal for clutter, is however not robust against local perturbations (unless a prohibitively large set of templates is stored). A multi-layer hierarchical architecture is needed to achieve the dual goal of local and global invariance. A key result of Part I is that a hierarchical architecture of the modules introduced earlier with "receptive fields" of increasing size, provides global invariance and stability to local perturbations (and in particular tolerance to local deformations). Interestingly, the *whole-parts theorem* implicitly defines "object parts" as small patches of the image which are locally invariant and occur often in images. The theory predicts a stratification of ranges of invariance in the ventral stream: size and position invariance should develop in a sequential order meaning that smaller transformations are invariant before larger ones, in earlier layers of the hierarchy. Translations would be the transformation associated with small apertures.

Part II studies spectral properties associated with the hierarchical architectures introduced in Part I. The motivation is given by a *Linking Conjecture*: instead of storing a sequence of frames during development, it is biologically

plausible to assume that there is Hebbian-like learning at the synapses in visual cortex. We will show that, as a consequence, the cells will effectively computes online the eigenvectors of the covariance of their inputs during development and store them in their synaptic weights. Thus the tuning of each cell is predicted to converge to one of the eigenvectors. We assume that the development of tuning in the cortical cells takes place in stages – one area – often called "layer" – at the time. We also assume that the development of tuning starts in V1 with Gaussian apertures for the simple cells. Translations are effectively selected as the only learnable transformations during development by small apertures – e.g. small receptive fields – in the first layer. The solution of the associated eigenvalue problem predicts that the tuning of cells in the first layer – identified with simple cells in V1 – can be approximately described as oriented Gabor-like functions. This follows in a parameter-free way from properties of shifts, e.g. the translation group. Further, rather weak, assumptions about the spectrum of natural images imply that the eigenfunctions should in fact be Gabor-like with a finite wavelength which is proportional to to the variance of the Gaussian in the direction of the modulation. The theory also predicts an elliptic Gaussian envelope. Complex cells result from a local group average of simple cells. The hypothesis of a second stage of hebbian learning at the level above the complex cells leads to wavelets-of-wavelets at higher layers representing local shifts in the $4-$cube of x,y, scale, orientation learned at the first layer. We derive simple properties of the number of eigenvectors and of the decay of eigenvalues as a function of the size of the receptive fields, to predict that the top learned eigenvectors – and therefore the tuning of cells – become increasingly complex and closer to each other in eigenvalue. Simulations show tuning similar to physiology data in V2 and $V4$.

Part III considers modules that are class-specific. For non-affine transformations of the image – for instance induced by out-of-plane rotations of a 3D object or non-rigid deformations – it is possible to prove that the dot-product technique of Part I can provide *approximate* invariance for certain classes of objects. A natural consequence of the theory is thus that non-affine transformations, such as rotation in depth of a face or change in pose of a body, can be approximated well by the same hierarchical architecture for classes of objects that have enough similarity in 3D properties, such as faces, bodies, perspective. Thus class-specific cortical areas make sense for invariant signatures. In particular, the theory predicts several properties of the macaque cortex face patches characterized by Freiwald and Tsao ([100, 101]), including a patch (called AL) which contains mirror symmetric cells and is the input to the pose-invariant patch (AM, [21]) – again because of spectral symmetry properties of the face templates.

A surprising implication of these theoretical results is that the computational goals and several of the tuning properties of cells in the ventral stream may follow from *symmetry properties* (in the sense of physics) of the visual world[2] through a process of unsupervised correlational learning, based on

[2]A symmetry – like bilateral symmetry – is defined as invariance under a transformation.

Hebbian synapses. In particular, simple and complex cells do not directly care about oriented bars: their tuning is a side effect of their role in translation invariance. Across the whole ventral stream the preferred features reported for neurons in different areas are only a symptom of the invariances computed and represented.

The results of each of the three parts stand on their own independently of each other. Together this theory-in-fieri makes several broad predictions, some of which are:

- invariance to small translations is the main operation of V1;
- invariance to larger translations and local changes in scale and scalings and rotations takes place in areas such as V2 and V4;
- class-specific transformations are learned and represented at the top of the ventral stream hierarchy; thus class-specific modules – such as faces, places and possibly body areas – should exist in IT;
- tuning properties of the cells are shaped by visual experience of image transformations during developmental (and adult) plasticity and can be altered by manipulating them;
- while features must be both discriminative and invariant, invariance to specific transformations is the primary determinant of the tuning of cortical neurons.
- homeostatic control of synaptic weights during development is required for hebbian synapses that perform online PCA learning.
- motion is key in development and evolution;
- invariance to small transformations in early visual areas may underly stability of visual perception (suggested by Stu Geman);
- the signatures (computed at different levels of the hierarchy) are used to retrieve information from an associative memory which includes labels of objects and verification routines to disambiguate recognition candidates. Back-projections execute the visual routines and control attentional focus to counter clutter.

The theory is broadly consistent with the current version of the HMAX model. It provides theoretical reasons for it while extending it by providing an algorithm for the unsupervised learning stage, considering a broader class of transformation invariances and higher level modules. We suspect that the performance of HMAX can be improved by an implementation taking into account the theory of this paper (at least in the case of class-specific transformations of faces and bodies [57]) but we still do not know.

The theory may also provide a theoretical justification for several forms of convolutional networks and for their good performance in tasks of visual

recognition as well as in speech recognition tasks (e.g. [50, 52, 47, 73, 6, 49]); it may provide even better performance by learning appropriate invariances from unsupervised experience instead of hard-wiring them.

The goal of this paper is to sketch a comprehensive theory with little regard for mathematical niceties: the proofs of several theorems are only sketched. If the theory turns out to be useful there will be scope for interesting mathematics, ranging from group representation tools to wavelet theory to dynamics of learning.

2 Introduction

The ventral stream is widely believed to have a key role in the task of object recognition. A significant body of data is available about the anatomy and the physiology of neurons in the different visual areas. Feedforward hierarchical models (see [85, 90, 92, 91] and references therein, see also section 7—in the appendix), are faithful to the anatomy, summarize several of the physiological properties, are consistent with biophysics of cortical neurons and achieve good performance in some object recognition tasks. However, despite these empirical and the modeling advances the ventral stream is still a puzzle: Until now we have not had a broad theoretical understanding of the main aspects of its function and of how the function informs the architecture. The theory sketched here is an attempt to solve the puzzle. It can be viewed as an extension and a theoretical justification of the hierarchical models we have been working on. It has the potential to lead to more powerful models of the hierarchical type. It also gives fundamental reasons for the hierarchy and how properties of the visual world determine properties of cells at each level of the ventral stream. Simulations and experiments will soon say whether the theory has some promise or whether it is nonsense.

As background to this paper, we assume that the content of past work of our group on models of the ventral stream is known from old papers [85, 90, 92, 91] to more recent technical reports [58, 59, 55, 56]. See also the section *Background* in Supp. Mat. [79]. After writing previous versions of this report, TP found a few interesting and old references about transformations, invariances and receptive fields, see [75, 34, 44]. It is important to stress that a key assumption of this paper is that in this initial theory and modeling it is possible to neglect subcortical structures such as the pulvinar, as well as cortical backprojections (discussed later).

2.1 Plan of the paper

Part I begins with the conjecture that the sample complexity of object recognition is mostly due to geometric image transformations, e.g. different viewpoints, and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. Part I deals with theoretical results that are independent of specific models. They are motivated by a one-layer

architecture "looking" at images (or at "neural images") through a number of small "apertures" corresponding to receptive fields, on a 2D lattice or layer. We have in mind a *memory-based architecture* in which learning consists of "storing" patches of neural activation. The argument of Part I is developed for this "batch" version; a biologically plausible "online" version is the subject of Part II. The first two results are

1. recording transformed templates - together called *the templatebook* – provides a simple and biologically plausible way to obtain a 2D-affine invariant *signature* for any new object, even if seen only once. The signature – a vector – is meant to be used for recognition. This is the *invariance lemma* in section 3.3.1.

2. several *aggregation* (eg pooling) functions including the energy function and the the max can be used to compute an invariant signature in this one-layer architecture (see 3.3.1).

Section **??** discusses limitations of the architecture, with respect to robustness to local perturbations. The conclusion is that multilayer, hierarchical architectures are needed to provide local and global invariance at increasing scales. In part II we will shows that global transformations can be approximated by local affine transformations. The key result of Part I is a characterization of the hierarchical architecture in terms of its *covariance and invariance* properties.

Part II studies spectral properties associated with the hierarchical architectures introduced in Part I. The motivation is given by a *Linking Conjecture*: instead of storing frames during development, learning is performed online by Hebbian synapses. Thus the conjecture implies that the tuning of cells in each area should converge to one of the eigenvectors of the covariance of the inputs. The size of the receptive fields in the hierarchy affects which transformations dominate and thus the spectral properties. In particular, the range of the transformations seen and "learned" at a layer depends on the aperture size: we call this phenomenon *stratification*. In fact translations are effectively selected as the only learnable transformations during development by the small apertures, e.g. small receptive fields, in the first layer. The solution of the associated eigenvalue problem – the *cortical equation* –predicts that the tuning of cells in the first layer, identified with simple cells in V1, should be oriented Gabor wavelets (in quadrature pair) with frequency inversely proportional to the size of an elliptic Gaussian envelope. These predictions follow in a parameter-free way from properties of the translation group. A similar analysis lead to wavelets-of-wavelets at higher layers representing local shifts in the 4-cube of x,y, scale, orientation learned at the first layer. Simulations show tuning similar to physiology data in $V2$ and $V4$. Simple results on the number of eigenvectors and the decay of eigenvalues as a function of the size of the receptive fields predict that the top learned eigenvectors, and therefore the tuning of cells, become increasingly complex and closer to each other in eigenvalue. The latter property implies that a larger variety of top eigenfunctions are likely to emerge

during developmental online learning in the presence of lateral inhibition (see section 4.2.1).

Together with the arguments of the previous sections this theory provides the following speculative framework. From the fact that there is a hierarchy of areas with receptive fields of increasing size, it follows that the size of the receptive fields determines the range of transformations learned during development and then factored out during normal processing; and that the transformation represented in an area influences – via the spectral properties of the covariance of the signals – the tuning of the neurons in the area.

Part III considers modules that are class-specific. A natural consequence of the theory of Part I is that for non-affine transformations such as rotation in depth of a face or change in pose of a body the signatures cannot be exactly invariant but can be approximately invariant. The approximate invariance can be obtained for classes of objects that have enough similarity in 3D properties, such as faces, bodies, perspective scenes. Thus class-specific cortical areas make sense for approximately invariant signatures. In particular, the theory predicts several properties of the face patches characterized by Freiwald and Tsao [100, 101], including a patch containing mirror symmetric cells before the pose-invariant patch [21] – again because of spectral properties of the face templates.

Remarks

- **Recurrent architecture, bottom-up and top-down** The ventral stream is a recurrent architecture, bottom-up *and* top-down, at the level of connectivity between cortical areas (in addition to the recurrent circuits within each area). Anatomy and psychophycs point clearly in this direction, though physiology is less clear. The theory here is very much based on the assumption of a recurrent circuitry (in the same spirit, digital computers are recurrent architectures). Our hypothetical architecture involves memory access from different levels of the hierarchy as well as top-down attentional effects, possibly driven by partial retrieval from an associative memory. The neural implementation of the architecture requires local feedback loops within areas (for instance for normalization operations). Thus the architecture we propose is hierarchical; its most basic skeleton is feedforward. The theory is most developed for the feedforward skeleton which is probably responsible for the first 100 msec of perception/recognition. This is the part of the theory described in this paper.

- **Memory access** A full image signature is a vector describing the "full image" seen by a set of neurons sharing a "full visual field" at the top layer, say, of the hierarchy. Intermediate signatures for image patches – some of them corresponding to object parts – are computed at intermediate layers. All the signatures from all level are used to access memory for recognition. The model of figure 1 shows an associative memory module that can be also regarded as a classifier.

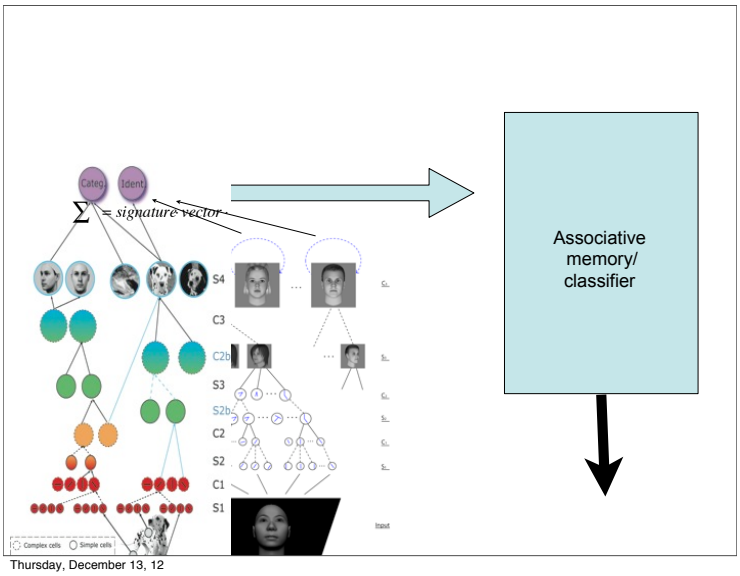


Figure 1: *Signatures from every level access associative memory modules.*

- **Identity-specific, pose-invariant vs identity-invariant, pose-specific representation** Part I develops a theory that says that invariance to a transformation can be achieved by pooling over transformed templates memorized during development. Part II says that an equivalent, more biological way to achieve invariance to a transformation is to store eigenvectors of a sequence of transformations of a template for several templates and then to pool the moduli of the eigenvectors.

  In this way different cortical patches can be invariant to identity and specific for pose and vice-versa. Notice that affine transformations are likely to be so important that cortex achieves more and more affine invariance through several areas in a sequence (≈ 3 areas).

- **Generic and class-specific transformations** We distinguish (as we did in past papers, see [82, 85]) between generic image-based transformations that apply to every object, such as scale, 2D rotation, 2D translation, and class specific transformations, such as rotation in depth for a specific class of objects such as faces. Affine transformations in $\mathbb{R}^2$ are generic. Class-specific transformations can be learned by associating templates from the images of an object of the class undergoing the transformation. They can be applied only to images of objects of the same class – provided the class is "nice" enough. This predicts modularity of the architecture for recognition because of the need to route – or reroute – information to transformation modules which are class specific [56, 57].

- **Memory-based architectures, correlation and associative learning** The

architectures discussed in this paper implement memory-based learning of transformations by storing templates (or principal components of a set of templates) which can be thought of as frames of a patch of an object/image at different times of a transformation. This is a very *simple, general and powerful way to learn rather unconstrained transformations*. Unsupervised (Hebbian) learning is the main mechanism at the level of simple cells. For those "complex" cells which may pool over several simple cells, the key is an unsupervised Foldiak-type rule: *cells that fire together are wired together*. At the level of complex cells this rule determines *classes of equivalence* among simple cells – reflecting observed *time correlations in the real world, that is transformations* of the image. The main function of each (simple + complex) layer of the hierarchy is thus to learn invariances via association of templates memorized during transformations in time. There is a general and powerful principle of time continuity here, induced by the Markovian (eg low-order differential equations) physics of the world, that allows associative labeling of stimuli based on their temporal contiguity[3].

- **Subcortical structures and recognition** We neglect the role of cortical backprojections and of subcortical structures such as the pulvinar. It is a significant assumption of the theory that this can be dealt with later, without jeopardizing the skeleton of the theory. The default hypothesis at this point is that inter-areas backprojections subserve attentional and gaze-directed vision, including the use of visual routines, all of which is critically important to deal with recognition in clutter. In this view, backprojections would be especially important in hyperfoveal regions (less than 20 minutes of visual angle in humans). Of course, inter-areas backprojections are likely to play a role in control signals for learning, general high-level modulations, hand-shakes of various types. Intra-areas feedback are needed even in a purely feed-forward model for several basic operations such as for instance normalization.

[3]There are many alternative formulations of temporal contiguity based learning rules in the literature. These include: [18, 107, 97, 39, 64, 19]. There is also psychophysics and physiology evidence for these [10, 106, 61, 60]

3 Part I: Memory-based Learning of Invariance to Transformations

Summary of Part I. *Part I assumes that an important computational primitive in cortex consists of normalized dot products between input vectors and synaptic weights. It shows that the following sequence of operation allows learning invariance to transformations for an image. During development a number of objects (templates) are observed during affine transformations; for each template a sequence of transformed images is stored. At run-time when a new image is observed its dot-products with the transformed templates (for each template) are computed; then the moduli of each term are pooled to provide a component of the signature vector of the image. The signature is an invariant of the image. Later in Part I we show that a multi-layer hierarchical architecture of dot-product modules can learn in an unsupervised way geometric transformations of images and then achieve the dual goal of invariance to global affine transformations and of robustness to image perturbations. These architectures learn in an unsupervised way to be automatically invariant to transformations of a new object, achieving the goal of recognition with one or very few labeled examples. The theory of Part I should apply to a varying degree to hierarchical architectures such as HMAX, convolutional networks and related feedforward models of the visual system and formally characterize some of their properties.*

3.1 Recognition is difficult because of image transformations

Summary. This section motivates the main assumption of the theory: a main difficulty of recognition is dealing with image transformations and this is the problem solved by the ventral stream. We show suggestive empirical observation and pose an open problem for learning theory: is it possible to show that invariances improve the sample complexity of a learning problem?

The motivation of this paper is the conjecture that the "main" difficulty, in the sense of *sample complexity*, of (clutter-less) object categorization (say dogs vs horses) is due to all the transformations that the image of an object is usually subject to: translation, scale (distance), illumination, rotations in depth (pose). The conjecture implies that recognition – i.e. both identification (say of a specific face relative to other faces) as well as categorization (say distinguishing between cats and dogs and generalizing from specific cats to other cats) – is easy (eg a small number of training example is needed for a given level of performance), if the images of objects are rectified with respect to all transformations.

3.1.1 Suggestive empirical evidence

To give a feeling for the arguments consider the empirical evidence – so far just suggestive and at the anecdotal level – of the "horse vs dogs" challenge (see Figures 3 and 2). The figure shows that if we factor out all transformations in images of many different dogs and many different horses – obtaining "normal-

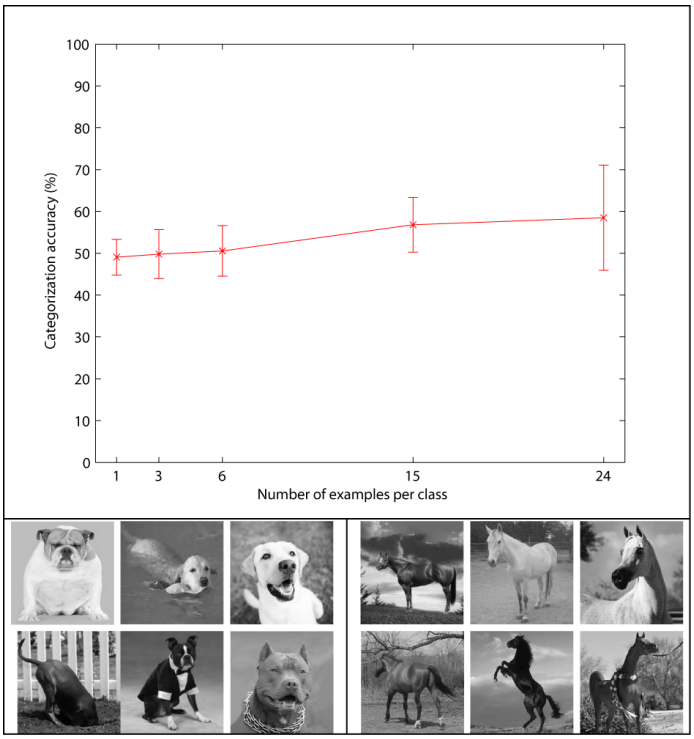


Figure 2: *Images of dogs and horses, in the wild, with arbitrary viewpoints (and clutter, eg background).The performance of a regularized least squares classifier (linear kernel, as in the next figure) is around chance. There are 60 images in total (30 per class) from Google. The x axis gives the number of training examples per class. Both clutter and viewpoint are likely to make the problem difficult. This demonstration leaves unclear the relative role of the two.*

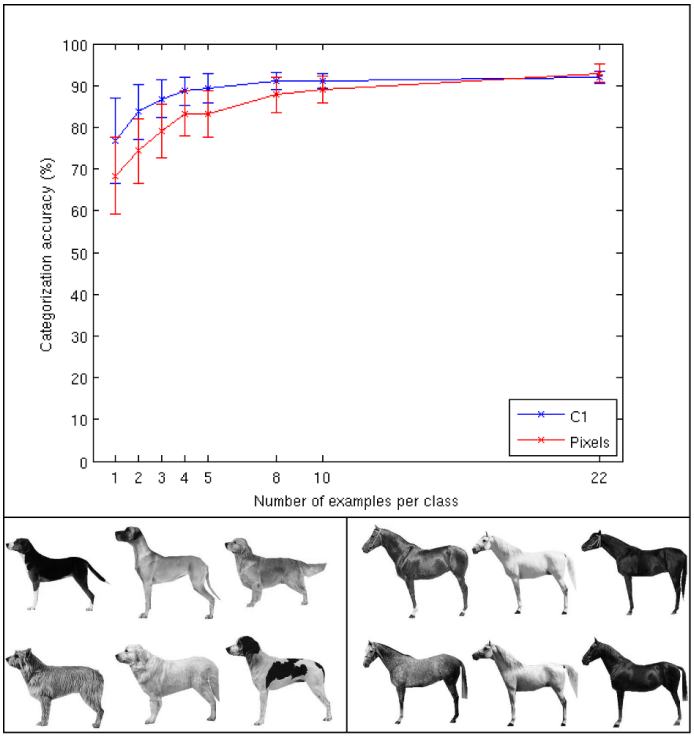


Figure 3: *Images of dogs and horses, 'normalized' with respect to image transformations. A regularized least squares classifier (linear kernel) tested on more than 150 dogs and 150 horses does well with little training. Error bars represent +/- 1 standard deviation computed over 100 train/test splits. This presegmented image dataset was provided by Krista Ehinger and Aude Oliva.*

ized" images with respect to viewpoint, illumination, position and scale – the problem of categorizing horses vs dogs is very easy: it can be done accurately with few training examples – ideally from a single training image of a dog and a single training image of a horse – by a simple classifier. In other words, the sample complexity of this problem is – empirically – very low. The task in the figure is to correctly categorize dogs vs horses with a very small number of training examples (eg small sample complexity). All the 300 dogs and horses are images obtained by setting roughly the same viewing parameters – distance, pose, position. With these "rectified" images, there is no significant difference between running the classifier directly on the pixel representation versus using a more powerful set of features (the C1 layer of the HMAX model).

3.1.2 Intraclass and viewpoint complexity

Additional motivation is provided by the following back-of-the-envelope estimates. Let us try to estimate whether the cardinality of the universe of possible images generated by an object originates more from intraclass variability – eg different types of dogs – or more from the range of possible viewpoints – including scale, position and rotation in 3D. Assuming a granularity of a few minutes of arc in terms of resolution and a visual field of say 10 degrees, one would get $10^3 - 10^5$ different images of the same object from x, y translations, another factor of $10^3 - 10^5$ from rotations in depth, a factor of $10 - 10^2$ from rotations in the image plane and another factor of $10 - 10^2$ from scaling. This gives on the order of $10^8 - 10^{14}$ distinguishable images for a single object. On the other hand, how many different distinguishable (for humans) types of dogs exist within the "dog" category? It is unlikely that there are more than, say, $10^2 - 10^3$. From this point of view, it is a much greater win to be able to factor out the geometric transformations than the intracategory differences.

Thus we conjecture that the key problem that determined the evolution of the ventral stream was recognizing objects – that is identifying and categorizing – from a single training image, *invariant* to geometric transformations. In computer vision, it has been known for a long time that this problem can be solved if the correspondence of enough points between stored models and a new image can be computed. As one of the simplest results, it turns out that under the assumption of correspondence, two training images are enough for orthographic projection (see [103]). Recent techniques for normalizing for affine transformations are now well developed (see [109] for a review or [**?**, **?**] for a novel method in transformations estimations). Various attempts at learning transformations have been reported over the years (see for example [83, 47] and for additional references the paper by Hinton [33]).

Our goal here is instead to explore approaches to the problem that do not rely on explicit correspondence operations and provide a plausible biological theory for the ventral stream. Our conjecture is that *the main computational goal of the ventral stream is to learn to factor out image transformations*. We show here several interesting consequences follow from this conjecture such as the hier-

archical architecture of the ventral stream. Notice that discrimination *without any invariance* can be done very well by a classifier which reads the pattern of activity in simple cells in V1 – or, for that matter, the pattern of activity of the retinal cones.

3.1.3 Invariant representations and bounds on learning rates

Open Problem *It seems obvious that learning/using an input representation which is invariant to natural transformations (eg contained in the distribution) should reduce the sample complexity of supervised learning. It is less obvious what is the best formalization and proof of the conjecture in the framework of learning theory.*

Theorem

The key observation is that we can estimate the compression coefficient associated with an invariant representation and that there are results connection compression coefficient of the training set and probability of test error (see Vapnik around p226)

more here!!!!

3.2 Templates and signatures

Summary. In this section we argue for another assumption in the theory: a primitive computation performed by neurons is a (normalized) dot product. This operation can be used by cortex to compute a signature for any image as a set of dot products of the image with a number of templates stored in memory. It can be regarded as a vector of similarities to a fixed set of templates. Signatures are stored in memory: recognition requires matching a signature with an item in memory.

The theory we develop in Part I is informed by the assumption that a *basic neural operation* carried by a neuron can be described by the dot product between an input vectors and a vector of synaptic weights on a dendritic tree. Part II will depend from the additional assumption that the vector of synaptic weights can be stored and modified by an online process of Hebb-like learning. These two hypothesis are broadly accepted.

In this paper we have in mind layered architectures of the general type shown in Figure 5. The computational architecture is memory-based in the sense that it stores during development sensory inputs and does very little in terms of additional computations: it computes normalized dot products and *pooling* (also called *aggregation*) functions. The results of this section are independent of the specifics of the hierarchical architecture and of explicit references to the visual cortex. They deal with the computational problem of invariant recognition from one training image in a layered, memory-based architecture.

The basic idea is the following. Consider a single aperture. Assume a mechanism that stores "frames", seen through the aperture, as an initial pattern "out in the world" transforms from $t = 1$ to $t = N$ under the action of a specific transformation (such as rotation). For simplicity assume that the set of transformations is a group. This is the "developmental" phase of learning the

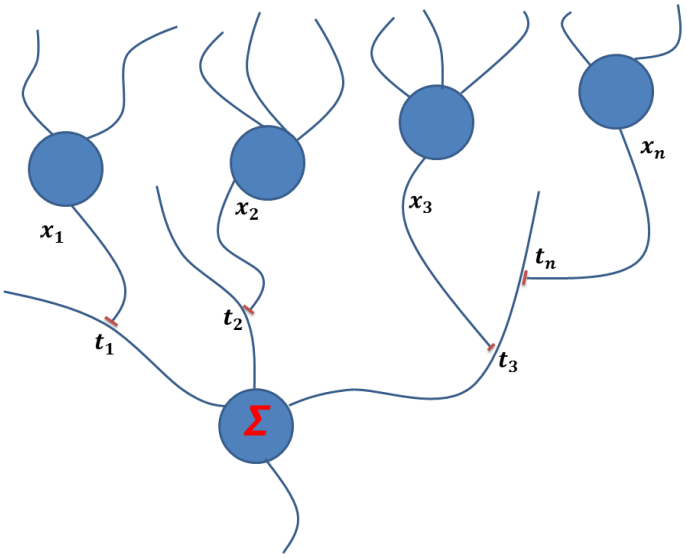


Figure 4: *A neuron receives on its dendritic tree in the order of* $10^3 - 10^4$ *synaptic inputs from other neurons. To a first approximation each synapse contributes a current which depends on the product of the input signal and the synapse. Since the soma of the neuron can be regarded as summating all these contributions, the neuron computes* xt *which is then encoded in spike trains.*

templates. At run time an image patch is seen through the aperture, and a set of normalized dot products with each of the stored templates (eg all transformations of each template) is computed. A vector called "signature" is then produced by an aggregation function – typically a group average over non-linear functions of the dot product with each template. Suppose now that at some later time (after development is concluded) the same image is shown, transformed in some way. The claim is that if the templates are closed under the same group of transformations then the signature remains the same. Several aggregation functions, such as the average or even the max (on the group), acting on the signature, will then be invariant to the learned transformation.

3.2.1 Preliminaries: resolution and size

The images we consider here are functions of two spatial variables x, y and time t. The images that the optics forms at the level of the retina are well-behaved functions, in fact entire analytic functions in $\mathbb{R}^2$, since they are bandlimited by the optics of the eye to about $60\ cycles/degree$ (in humans). The photoreceptors sample the image in the fovea according to Shannon's sampling theorem on a hexagonal lattice with a distance between samples equal to the diameter of the cones (which are tightly packed in the fovea) which is 27 seconds of arc. The

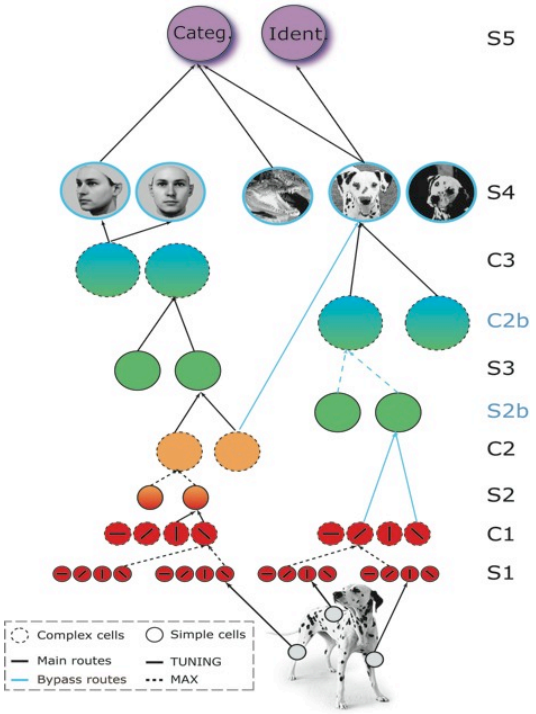


Figure 5: *Hierarchical feedforward model of the ventral stream – a modern interpretation of the Hubel and Wiesel proposal (see [84]). The theoretical framework proposed in this paper provides foundations for this model and how the synaptic weights may be learned during development (and with adult plasticity). It also suggests extensions of the model such as class specific modules at the top.*

sampled image is then processed by retinal neurons; the result is transmitted to the LGN and then to primary visual cortex through the optic nerve, consisting of axons of the retinal ganglion cells. At the LGN level there are probably two neural "images" in the fovea: they may be roughly described as the result of DOG (Difference-of-Gaussian or the similar Laplacian-of-Gaussian) spatial filtering (and sampling) of the original image at two different scales corresponding to the magno and the parvo system. The parvo or midget system is spatially bandpass (but with a DC component). There is also high-pass filtering in time at the level of the retina which can be approximated by a time derivative component or more accurately as a filter providing, in the Fourier domain, $\beta F(\omega_x, \omega_y, \omega_t) + i\omega_t F(\omega_x, \omega_y, \omega_t)$ where F is the fourier transform of the image. Thus the neural image seen by the cortex is bandpass in space and time. The finest grain of it is set by the highest spatial frequency (notice that if λ_u corresponds to the highest spatial frequency then sampling at the Shannon rate, eg on a lattice with edges of length $\frac{\lambda_u}{2}$ preserves all the information.)

3.2.2 Templatesets

Since the goal of visual recognition in the brain is not reconstruction but identification or categorization, a representation possibly used by the ventral stream and suggested by models such as Figure 5, is in terms of an overcomplete set of measurements on the image, a vector that we will call here a *measurement*.

It is interesting to notice that the *nature of the measurements may not be terribly important* as long as they are reasonable and there are enough of them. A historical motivation and example for this argument is provided by OCR algorithms based on counting intersections of characters with a random, fixed set of lines (see 6. A more mathematical *motivation* is provided by a theorem due to Johnson and Lindenstrauss. Their classic result says informally that any set of n points in d-dimensional Euclidean space can be embedded into k-dimensional Euclidean space where k is logarithmic in n and independent of d via random projections so that all pairwise distances are maintained within an arbitrarily small factor. The theorem will be discussed later together with more classical approximate embeddings as provided by *finite frames*. We mention it here as a suggestion that since there are no special conditions on the projections (though the assumption of randomness is actually strong) most measurements will work to some degree, as long as there are enough independent measurements (but still with $k << n$ in most cases of interest). Notice for future use that the *discriminative power* of the measurements depends on k.

In summary we assume

- The ventral stream computes a representation of images that supports the task of recognition (identification and categorization). It does not need to support image reconstruction.

Figure 6: *Number of intersection per line (out of an arbitrary, random but fixed set) provides an effective set of measurements for OCR.*

- The ventral stream provides a *signature* which is invariant to geometric transformations of the image and to deformations that are locally approximated by affine transformations

- Images (of objects) can be represented by a set of functionals of the image, eg measurements. Neuroscience suggests that a natural way for a neuron to compute a simple image measurements is a (possibly normalized) dot product between the image and a vector of synaptic weights corresponding to the tuning of the neuron.

Before showing how to built and invariant signature let us give a few definitions:

Definition 1. *Space of images:* $\mathcal{X} \subseteq L^2(\mathbb{R}^2)$ *(or* $\mathbb{R}^d$*) where*

$$L^2(\mathbb{R}^2) = \{I : \mathbb{R}^2 \to \mathbb{R},\ s.t.\ \int \mid I(x,y) \mid^2 dxdy < \infty\}$$

$$\langle I, t \rangle = \int I(x,y)t(x,y)dxdy, \quad I, t \in \mathcal{X}$$

the space of square integrable functions equipped with dot product.

Definition 2. *Template set:* $\mathcal{T} \subseteq \mathcal{X},\ (or\ \mathbb{R}^d)$*: a set of images (or, more generally, image patches)*

Given a finite template set ($|\mathcal{T}| = K < \infty$) we define a set of linear functionals of the image I:

$$\langle I, t^k \rangle, \quad k = 1, ..., K.$$

which we call measurements.

Definition 3. *The image I can be represented in terms of its measurement vector defined with respect to the templateset $\mathcal{T}$:*

$$\Delta_I = (\langle I, t^1 \rangle, \langle I, t^2 \rangle, ..., \langle I, t^K \rangle)^T$$

We consider here two examples for choosing a set of templates. Both examples are relevant for the rest of the paper. Consider as an example the set of images in $\mathcal{X} \in \mathbb{R}^d$. The obvious choice for the set of templates is to be an orthonormal basis in the space of "images patches", eg in $\mathbb{R}^d$. Our first example is a variation of this case: the templateset $\mathcal{T}$ is assumed to be a *frame* (see Appendix 19.1) for the n-dimensional space $\mathcal{X}$ spanned by n chosen images in R^d, that is the following holds

$$A||I||^2 \leq \sum_{k=1}^{T} | < I, t^k > |^2 \leq B||I||^2 \tag{1}$$

where $I \in \mathbb{R}^d$ and $A \leq B$. We can later assume that $A = 1 - \epsilon$ and $B = 1 + \epsilon$ where ϵ can be controlled by the cardinality T of the templateset $\mathcal{T}$. In this example consider for instance $n \leq T < d$.

This means that we can represent n images by projecting them from $I \in \mathbb{R}^d$ to $\mathbb{R}^T$ by using templates. This map $F : \mathbb{R}^d \to \mathbb{R}^T$ is such that for all $I, I' \in \mathcal{X}$ (where $\mathcal{X}$ is a n-dimensional subspace of R^d)

$$A \parallel I - I' \parallel \leq \parallel FI - FI' \parallel \leq B \parallel I - I' \parallel .$$

If $A = 1 - \epsilon$ and $B \leq 1 + \epsilon$ where $\epsilon = \epsilon(T)$ the projections of I and I' in $\mathbb{R}^T$ maintains the distance within a factor ϵ: the map is a quasi-isometry and can be used for tasks such as classification. The second example is based on the choice of *random templates* and a result due to Johnson and Lindenstrauss (J-L).

Proposition 1. *For any set V of n points in $\mathbb{R}^d$, there exists a map $P : \mathbb{R}^d \to \mathbb{R}^T$ such that for all $I, I' \in V$*

$$(1 - \epsilon) \parallel I - I' \parallel \leq \parallel PI - PI' \parallel \leq (1 + \epsilon) \parallel I - I' \parallel$$

where the map P is a random projection on $\mathbb{R}^T$ and

$$kC(\epsilon) \geq ln(n), \quad C(\epsilon) = \frac{1}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right).$$

The JL theorem suggests that good representations for classification and discrimination of n images may be provided by K dot products with *random* templates since they provide a quasi-isometric embedding of images.

Remarks

- The dimensionality of the measurement vector given by JL depends on n but not on d;
- The dimensions of the measurement vector are logarithmic in n;
- The fact that random templates are sufficient suggests that the precise choice of the templates is not important, *contrary* to the present folk wisdom of the computer vision community.
- The Johnson-Lindenstrauss result implies that if I and I' are very close (that is $||I - I'|| \leq \epsilon$), their projections $P(I)$ and $P(I')$ are very close in every norm, in particular component-wise (that is $max_k |P(I)_k - P(I')_k| \leq \delta$).

3.2.3 Transformations and templatebooks

The question now is how to compute a measurement vector that is capable not only of discriminating different images but is also *invariant* to certain transformations of the images. We consider geometric transformations of images due to changes in viewpoints.
We define as *geometric transformations* of the image I the action of the operator $U(T) : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ acting as:

$$[U(T)I](x, y) = I(T^{-1}(x, y)) = I(x', y'), \quad I \in L^2(\mathbb{R}^2)$$

where $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a coordinate change.
In general $U(T) : \mathbb{R}^2 \to \mathbb{R}^2$ isn't a unitary operator. However it can be made unitary defining

$$[U(T)I](x, y) = |J_T|^{-\frac{1}{2}} I(T^{-1}(x, y))$$

where $|J_T|$ is the determinant of the Jacobian of the transformation. Unitarity of the operator will be useful in the next paragraph.
A key example of T is the affine case, eg

$$\mathbf{x}' = A\mathbf{x} + \mathbf{t}_\mathbf{x}$$

where $A \in GL(2, \mathbb{R})$ the linear group in dimension two and $\mathbf{t}_\mathbf{x} \in \mathbb{R}^2$.

In fact, in most of this paper we will consider transformations that correspond to the affine group $Aff(2, \mathbb{R})$ which is an extension of $GL(2, \mathbb{R})$ (the general linear group in $\mathbb{R}^2$) by the group of translations in $\mathbb{R}^2$. Consider a finite group G whose elements we indicate with g_i, $i = 1, ..., |G|$ and a finite set of templates, t^k, $k = 1, ..., K$. Let us now define a key object of the paper:

Definition 4. *Suppose the set of templates are closed under the action of a group of transformations, i.e.*

$$g_i t^k = t^l, \quad \exists\, l \,\forall\, i = 1, ..., |G|, \;\; k = 1, ..., K.$$

We assume that the basic element of our architecture, the memory based module, stores (during development) sequences of transformed templates for each template in the templateset. We define the Templatebook as

$$\mathbb{T}_{t_1,...,t_T} = \begin{pmatrix} g_0 t_1, \; g_0 t_2, & ..., g_0 t_T \\ \vdots & \\ g_{|G|} t_1, \; g_{|G|} t_2, & ..., g_{|G|} t_T \\ \cdot & \end{pmatrix}$$

the collection of all transformed templates. Each row corresponds to the orbit of the template under the transformations of G.

3.3 Invariance and discrimination

Summary. If a signature is a dot product between the image and a template, then the average of any function of the dot product between all the transformations of the image and the template is an invariant. Under some assumptions this is equivalent to the average of any function of the dot product of the image and all the transformations of the template. Thus an invariant can be obtained from a single image. However, invariance is not enough: discrimination is also important. Thus we start considering the full orbit induced by the action of the group on the image. For compact groups if two orbits have a point in common then they are the same orbit. A distribution P_I induced by the group acting on the image I is associated to each orbit: the distribution is invariant and discriminant. Thus the discriminability question is answered if the distribution can be characterized uniquely by a set of empirical averages. This section shows that group averaging of projections followed by sigmoidal nonlinearities can be used for an empirical estimation of one-dimensional projections of P_I. In turns, such estimators characterize uniquely the distribution, according to J-L type results for distributions. Thus a set of estimators provides a discriminant signature vector with invariant components. The biologically elegant prediction is that linear combinations of complex-like cells with a sigmoidal nonlinearity can provide discriminative and invariant signatures.

We start with a rather idealized situation (group is compact, the image does not contain clutter) for simplicity. We will make our framework more realistic in section **??**. For a more mathematically formal description of the problem and an alternative formulation see appendix 9 and10.

3.3.1 The invariance lemma

Consider the dot products of all transformation of an image with one component of the templateset t

$$\Delta_{G,I} = (\langle g_0 I, t\rangle, \langle g_1 I, t\rangle ..., \langle g_{|G|} I, t\rangle)^T$$

Clearly, if the transformation is unitary (i.e. $g^\dagger = g^{-1}$)

$$\Delta_{G,I} = (\langle g_0 I, t\rangle, \langle g_1 I, t\rangle, ..., \langle g_{|G|} I, t\rangle)^T = (\langle I, g_0^{-1} t\rangle, \langle I, g_1^{-1} t\rangle, ..., \langle I, g_{|G|}^{-1} t\rangle)^T$$

where g^{-1} is the inverse transformation of g and $\Delta_{G,I}$ is the measurement vector of the image w.r.t the transformations of one template, that is the orbit obtained by the action of the group on the dot product. Note that the following is mathematically trivial but important from the point of view of object recognition. To get measurements of an image *and all its transformations* it is not necessary to "see" all the transformed images: a single image is sufficient provided a templatebook is available. In our case we need for any image, just one row of a templatebook, that is all the transformations of one template:

$$\mathbb{T}_t = (g_0 t, g_1 t, ..., g_{|G|} t)^T.$$

Note that the orbits $\Delta_{I,G}$ and $\Delta_{gI,G}$ are the same set of measurements apart from ordering). The following *invariance lemma* follows.

Proposition 2. *(***Invariance lemma***) Given $\Delta_{I,G}$ for each component of the templateset an invariant signature Σ can be computed as the group average of a nonlinear function, $\eta : \mathbb{R} \to \mathbb{R}$, of the measurements which are the dot products of the image with all transformations of one of the templates, for each template:*

$$\mu^k(I) = \frac{1}{|G|} \sum_{g \in G} \eta(\langle I, g t^k\rangle), \quad k = 1, ..., K. \tag{2}$$

A classical example of invariant is $\eta(\cdot) \equiv |\cdot|^2$, the energy

$$\mu^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} |\langle I, g_i t^k\rangle|^2$$

Other examples of invariant group functionals are

- Max: $\mu^k(I) = max_i \langle I, g_i t^k\rangle$
- Average: $\mu^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \langle I, g_i t^k\rangle$

We call the functions μ in eq. 2 *pooling or aggregation functions*. The original HMAX model uses a max of $I \circ g_i t^k$ over i or the average of $I \circ g_i t^k$ over i or the average of $(I \circ g_i t^k)^2$ over i. In convolutional networks the pooling function is often the average of η terms, where η is a sigmoidal function describing the threshold operation of a neuron. Such aggregation operations can also be approximated by the generalized polynomial

$$y = \frac{\sum_{i=1}^{n} w_i \, {x_i}^p}{k + \left(\sum_{i=1}^{n} {x_i}^q \right)^r} \tag{3}$$

for appropriate values of the parameters (see [46]). Notice that by defining the p-norm of x with $||x||_p = (\sum |x_i|^p)^{\frac{1}{p}}$, it follows that $max(x) = ||x||_\infty$ and $energy - operation(x) = ||x||_2$. in any case, the invariant *signature*,

$$\Sigma(I) = (\mu^1(I), \mu^2(I), ..., \mu^K(I))^T$$

is a vector which is invariant under the transformations g_i.

Remarks

- *Signature* Notice that *not all* individual components of the signature (a vector) have to be discriminative wrt a given image – whereas *all* have to be invariant. In particular, a number of poorly responding templates could be together quite discriminative.
- *Group averages* Image blur corresponds to local average of pixel values. It is thus a (local) group average providing the first image moment.

3.3.2 Orbits

An invariant signature based on the arithmetic average is invariant but most of the times not enough discriminative. As we will see later, signatures consisting of ceratin nonlinear functions of the same dot products (instead of a single η) function are discriminant.

In the case of invariant recognition wrt a group of transformations the basic object to be discriminated is the set of images generated by the group from a single image – called an orbit (see 3.3.1.). Two "objects" are different independently of their transformation if their orbits are different. This makes sense because if two orbits intersect in one point they are identical everywhere. Thus equivalence of two orbits implies that at least one point, eg one image, is in common – but then all are. In the same spirit, recall that *iff* a group is compact then the quotient group is a metric space. This implies that a distance between orbits can be defined (see Proposition 7).

How can two orbits be characterized and compared in a biologically plausible way? Intuitively two empirical orbits are the same *irrespectively of the ordering of their points*. Thus a natural approach in the finite case is to rank all the points of the $\mathbf{I}$ set and do the same for the $\mathbf{I}'$ set. Then a comparison should be easy (computationally). The next two mathematical sections describe the more general approach of comparing the probability distribution associated with the $\mathbf{I}$ set with the distribution associated with the $\mathbf{I}'$ set. We will discuss the following axiom that we take as a definition of equivalence between the orbits generated by G acting on the points I and I',

Definition 5.

$$P_I = P_{I'} \iff I \sim I'$$

where P is the probability distribution induced by the group.

After the mathematical interlude of section 3.4.1 we will address in section 3.4.2 the key question of how to obtain empirical estimates of P_I in a biological plausible way.

3.4 Invariant and unique signatures

Let $\mathcal{X}$ be a Hilbert space with norm and inner product denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. We can think of $\mathcal{X}$ as the space of images. We typically consider $\mathcal{X} = \mathbb{R}^d,\ L^2(\mathbb{R}),\ L^2(\mathbb{R}^2)$. If G is a group with an abuse of notation, we denote by g both a group element in G and its action/representation on $\mathcal{X}$.

Given an image $I \in \mathcal{X}$ and a group representation g, the orbit $O_I = \{I' \in \mathcal{X}\ s.t.\ I' = gI, g \in G\}$ is uniquely associated to an image and all its transformations. The orbit provides an invariant representation of I, i.e. $O_I = O_{gI}$ for all $g \in G$. Indeed, we can view an orbit as the realization of a random variable with distribution P_I induced by the group action. From this observation, a (vector) signature can be derived for compact groups, by using results characterizing probability distributions via their one dimensional projections.

In this section we discuss and study the signature given by

$$\Sigma(I) = (\mu_1^1(I), \dots, \mu_N^K(I)) = (\mu^1(I), \dots, \mu^k(I)),$$

where each vector $\mu^k(I) \in \mathbb{R}^N$ is a histogram corresponding to a one dimensional projection defined by a template $t^k \in \mathcal{X}$. For the rest of this section we let $\mathcal{X} = \mathbb{R}^d$.

3.4.1 Orbits and probability distributions

If G is a compact group, the associated Haar measure dg can be normalized to be a probability measure, so that, for any $I \in \mathbb{R}^d$, we can define the random variable,

$$Z_I : G \to \mathbb{R}^d, \qquad Z_I(g) = gI.$$

The corresponding distribution P_I is defined as $P_I(A) = dg(Z_I^{-1}(A))$ for any Borel set $A \subset \mathbb{R}^d$ (with some abuse of notation we let dg be the normalized Haar measure).
Recall that we define two images, $I, I' \in \mathcal{X}$ equivalent (and we indicate it with $I \sim I'$) if there exists $g \in G$ s.t. $I = gI'$; we have the following:

Theorem 1. *The distribution P_I is invariant and unique i.e. $I \sim I' \ \Leftrightarrow\ P_I = P_{I'}$.*

Proof. We first prove that $I \sim I' \Rightarrow P_I = P_{I'}$. By definition $P_I = P_{I'}$ iff $\int_A dP_I(s) = \int_A dP_{I'}(s), \forall\, A \subseteq \mathcal{X}$, that is $\int_{Z_I^{-1}(A)} dg = \int_{Z_{I'}^{-1}(A)} dg$, where,

$$Z_I^{-1}(A) = \{g \in G \; s.t. \; gI \subseteq A\}$$
$$Z_{I'}^{-1}(A) = \{g \in G \; s.t. \; gI' \in A\} = \{g \in G \; s.t. \; g\bar{g}I \subseteq A\},$$

$\forall\, A \subseteq \mathcal{X}$. Note that $\forall\, A \subseteq \mathcal{X}$ if $gI \in A \Rightarrow g\bar{g}^{-1}\bar{g}I = g\bar{g}^{-1}I' \in A$, so that $g \in Z_I^{-1}(A) \Rightarrow g\bar{g}^{-1} \in Z_{I'}^{-1}(A)$, i.e. $Z_I^{-1}(A) \subseteq Z_{I'}^{-1}(A)$. Conversely $g \in Z_{I'}^{-1}(A) \Rightarrow g\bar{g} \in Z_I^{-1}(A)$, so that $Z_I^{-1}(A) = Z_{I'}^{-1}(A)\bar{g},\ \forall A$. Using this observation we have,

$$\int_{Z_I^{-1}(A)} dg = \int_{(Z_{I'}^{-1}(A))\bar{g}} dg = \int_{Z_{I'}^{-1}(A)} d\hat{g}$$

where in the last integral we used the change of variable $\hat{g} = g\bar{g}^{-1}$ and the invariance property of the Haar measure: this proves the implication.
To prove that $P_I = P_{I'} \Rightarrow I \sim I'$, note that $P_I(A) - P_{I'}(A) = 0, \forall\, A \subseteq \mathcal{X}$, is equivalent to

$$\int_{Z_{I'}^{-1}(A)} dg - \int_{Z_I^{-1}(A)} dg = \int_{Z_I^{-1}(A) \triangle Z_{I'}^{-1}(A)} dg = 0, \; \forall A \in \mathcal{X}$$

where $\triangle$ denotes the symmetric difference. This implies $Z_I^{-1}(A) \triangle Z_{I'}^{-1}(A) = \emptyset$ or equivalently

$$Z_I^{-1}(A) = Z_{I'}^{-1}(A), \; \forall\, A \in \mathcal{X}$$

In other words of any element in A there exist $g', g'' \in G$ such that $g'I = g''I'$. This implies $I = {g'}^{-1}g''I' = \bar{g}I', \; \bar{g} = {g'}^{-1}g''$, i.e. $I \sim I'$. □

3.4.2 Empirical measurements of probabilities of projections

Thus the distribution P_I is invariant and then discriminative. How can neurons obtain an empirical estimate – a signature that uniquely characterize P_I? Of course we are not interested in computing the probability distributions and or compare them (see Appendix 10.4.4). We are interested in a set of measurements that characterize uniquely the distribution. An approach which seems relevant, though indirectly, for neuroscience is related to the characterization of distributions in terms of (multivariate) moments. Notice that univariate moments can be estimated using pooling functions such as in Equation (2) with appropriate choices of the function η (but it is not clear how to biologically estimate multivariate moments). In addition, a sufficient (possibly infinite) number of moments uniquely characterizes a probability distribution (see Appendix 10.4.2). There is however a simpler and more elegant way (because of biological plausibility) to obtain selective and invariant quantities in terms of group averages such as Equation (2). We first show that empirical estimates of lower dimensional projections of P_I can be estimated using a surprisingly simple and biologically plausible operation. We then show that such lower dimensional estimates uniquely characterize P_I.

Probability distributions via one dimensional projections. Given the above discussion a *signature* could be associated to I by constructing a histogram approximation of P_I, but this would require dealing with high dimensional histograms. The following classic theorem gives a way around this problem.
For a *template* $t \in \mathbb{S}(\mathbb{R}^d)$, where $\mathbb{S}(\mathbb{R}^d)$ is unit sphere in $\mathbb{R}^d$, let $I \mapsto \langle I, t\rangle$ be the associated projection. Moreover, let $P_{\langle I,t\rangle}$ be the distribution associated to the random variable $g \mapsto \langle gI, t\rangle$ (or equivalently $g \mapsto \langle I, g^{-1}t\rangle$, if g is unitary). Let $\mathcal{E} = [t \in \mathbb{S}(\mathbb{R}^d),\ s.t.\ P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}]$.

Theorem 2. *(Cramer-Wold, [11]) For any pair P, Q of probability distributions on $\mathbb{R}^d$, we have that $P = Q$ if and only if $\mathcal{E} = \mathbb{S}(\mathbb{R}^d)$.*

In words two probability distributions are equal if and only if their projections on any of the unit sphere directions is equal. The result can be equivalently stated as follows.

Theorem 3. *(consequence of Theorem 3.4 in [13]) Let P and Q two probability distributions on $\mathbb{R}^d$. Let λ be the normalized uniform measure on $\mathbb{S}(\mathbb{R}^d)$. We have that $\lambda(\mathcal{E}) > 0 \Leftrightarrow P = Q$.*

The latter formulation implies in particular that the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 1 if and only if $P = Q$ and the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 0 if and only if $P \neq Q$. Moreover it suggests to define a metric on distributions (orbits) as follows,

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle}) d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_0 is any metric on one dimensional probability distributions. Indeed, it is easy to check that d is a metric. In particular note that, in view of Cramer Wold Theorem, $d(P, Q) = 0$ if and only if $P = Q$. As discussed in the main text, each one dimensional distribution $P_{\langle I,t\rangle}$ can be approximated by a suitable histogram $\mu^t(I) = (\mu^t_n(I))_{n=1,\dots,N} \in R^N$, so that, in the limit in which the histogram approximation is accurate

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I')) d\lambda(t), \quad \forall I, I' \in \mathcal{X}, \tag{4}$$

where d_μ is the metric on histograms induced by d_0.

A natural question is whether there are situations in which a finite number of projections suffice to discriminate any two probability distributions, that is $P_I \neq P'_I \Leftrightarrow d(P_I, P_{I'}) \neq 0$. Indeed, empirically it seems that this could be the case with a small number of templates, as shown for instance in [12]. The problem of characterizing mathematically situations in which a finite number of (one-dimensional) projections is sufficient is challenging. The next result provides a partial answer to this question.

We start observing that the metric (4) can be approximated by uniformly sampling K templates and considering

$$\hat{d}_K(P_I, P_{I'}) = \frac{1}{K}\sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')), \tag{5}$$

where $\mu^k = \mu^{t^k}$. The following result shows that a finite number K of templates is sufficient to obtain an approximation within a given precision ϵ. Towards this end let

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\|\mu^k(I) - \mu^k(I)\right\|_{\mathbb{R}^N}. \tag{6}$$

The following result holds.

Theorem 4. *Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{c}{\epsilon^2}\log\frac{n}{\delta}$, where c is a universal constant. Then*

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon, \tag{7}$$

with probability $1-\delta^2$, for all $I, I' \in \mathcal{X}_n$.

Proof. The proof follows from an application of Höeffding inequality and a union bound.
Fix $I, I' \in \mathcal{X}_n$. Define the real random variable $Z : \mathbb{S}(\mathbb{R}^d) \to \mathbb{R}$, $Z(t^k) = \left\|\mu^k(I) - \mu^k(I)\right\|_{\mathbb{R}^N}$, with $k = 1, \ldots, K$. From the definitions it follows that $\|Z\| \leq c$ and $\mathbb{E}(Z) = d(P_I, P_{I'})$. Then Höeffding inequality implies

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| = |\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(Z) - Z(t^k)| \geq \epsilon,$$

with probability at most $e^{-c\epsilon^2 k}$. A union bound implies a result holding uniformly on $\mathcal{X}_n$; the probability becomes at most $n^2 e^{-c\epsilon^2 K}$. The desired result is obtained noting that this probability is less than δ^2 as soon as $n^2 e^{-c\epsilon^2 K} < \delta^2$ that is $K \geq \frac{c}{\epsilon^2}\log\frac{n}{\delta}$. □

Thus the discriminability question can be answered in terms of empirical estimates of the one-dimensional distributions of projections of the image and transformations induced by the group on a number of templates $t^k,\ k = 1, ..., K$.

Remark 1. *The above result can be compared to a version of Cramer Wold Theorem for discrete probability distributions. Indeed, the following theorem holds [32].*

Theorem 5. *Let P be a discrete probability distribution on $\mathbb{R}^d$ with a support made with exactly k distinct atoms. Assume that $V_1, ..., V_{k+1}$ are subspaces of $\mathbb{R}^d$ of respective dimensions $d_1, ..., d_{k+1}$ such that no couple of them is contained in a hyperplane (i.e. no straight line is perpendicular to more than one of them). Then, for any probability distribution Q in $\mathbb{R}^d$, we have $P = Q$ if and only if $\pi_{V_i} \in \mathcal{E}_{d_i}(P, Q)$, for every $1 \leq i \leq k+1$, where π_{V_i} is the projector onto the subspace V_i and $\mathcal{E}_{d_i}(P, Q)$ is the set of subspaces of dimensionality d_i with equal P and Q projections.*

In particular, for a probability distribution consisting of k atoms in $\mathbb{R}^d$, we see that at most $k+1$ directions ($d_1 = d_2 = ... = d_{k+1} = 1$) are enough to characterize the distribution, thus a finite – albeit large – number of one-dimensional projections.

3.4.3 Computations by simple and complex cells

The approach described above map the computation of an invariant signature onto well-known capabilities of cortical neurons. We start from an old observation. A key difference between the basic elements of our digital computers and neurons is the number of connections: 3 vs. $10^3 - 10^4$ synapses per cortical neuron. Taking into account basic properties of synapses, it follows that a single neuron can compute high-dimensional ($10^3 - 10^4$) inner products between input vectors and the stored vector of synaptic weights. A natural scenario is then the following (see also Fig. **??**). Consider a module of "simple" and "complex" cells [36] looking at the image through a window defined by their common receptive fields. During development—and more generally, during visual experience—a set of $|G|$ simple cells store in their synapses an image patch t^k and its transformations $g_1 t^k, ..., g_{|G|} t^k$—one transformation step per simple cell. This is done for several image patches t^k (templates), $k = 1, \cdots, K$. Later, when an image is presented, the simple cells compute $\langle I, g_i t^k \rangle$ for $i = 1, ..., |G|$. The next step is to estimate the one-dimensional probability distribution of such a projection—that is, the distribution of the outputs of the simple cells. One idea is to compute a histogram. It is generally assumed that complex cells pool the outputs of simple cells. Thus a complex cell could compute $\mu_n^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + n\Delta)$ where σ is a smooth version of the step function ($\sigma(x) = 0$ for $x \leq 0$, $\sigma(x) = 1$ for $x > 0$) and $n = 1, ..., N$. Each of these N complex cells computes one bin of an approximated CDF (cumulative distribution function) for $P_{\langle I, t^k \rangle}$. The CDF suits our purpose equally well, with the measurements $\mu_n^k(I)$ containing sufficient information for an empirical estimate of the distribution at resolution Δ. A visual system does not need to recover the actual probabilities from the empirical estimate in order to compute a unique signature. The set of $\mu_n^k(I)$ values is sufficient, since it uniquely identifies the associated orbit (see box 1).

3.4.4 A theory of pooling

The arguments above make a few predictions. They require an effective normalization of the elements of the inner product (e.g. $\langle I, g_i t^k \rangle \mapsto \frac{\langle I, g_i t^k \rangle}{\|I\| \|g_i t^k\|}$) for the property $\langle gI, t^k \rangle = \langle I, g^{-1} t^k \rangle$ to be valid. They also imply that pooling in general requires a nonlinearity of the complex cells. These predictions, as well as the overall scheme, are supported by physiological data. Notice that invariant signatures can be computed in several ways from one-dimensional probability distributions. Instead of the $\mu_n^k(I)$ components representing directly the empirical distribution, complex cells could compute the moments $m_n^k(I) = 1/|G| \sum_{i=1}^{|G|} (\langle I, g_i t^k \rangle)^n$ of the same distribution [45]. Under some

rather weak conditions, the set of *all* moments uniquely characterizes the one-dimensional distribution $P_{\langle I,t^k\rangle}$ (and thus P_I). $n = 1$ corresponds to pooling via sum/average (and is the only pooling not requiring a nonlinearity). $n = 2$ corresponds to an energy model of complex cells [2]; very large n corresponds to a max operation. In our simulations, using just one of these moments seems to provide sufficient selectivity to a hierarchical architecture. Other nonlinearities are also possible; see [80]. Conventional wisdom interprets available evidence to suggest that simple/complex cells in V1 and "simple" cells in the macaque anterior lateral face patch may be described in terms of energy models, but our alternative suggestion of empirical histogramming by sigmoidal nonlinearities with different offsets may fit the diversity of data even better. The arguments of this section may begin to provide a comprehensive theory of "pooling" and offer a possible explanation for the persistence of many different pooling schemes—max vs sum vs sum-of-squares—in current deep convolutional networks. Under our interpretation, these different pooling functions all play the same role.

In summary learning invariance and computing an invariant representation for a new image patch requires the following steps (see also box): 1. store, once for all, one randomly chosen image t^1 (called template) and its $|G| - 1$ transformations for a total of $|G|$ stored images $t^1, t^1_1, \cdots, t^1_{|G|}$; 2. to obtain an invariant signature for a new image a) compute $|G|$ dot products of the new image with each of the $|G|$ stored templates b) compute a statistics of these $|G|$ numbers such as the histogram or one or more of its moments such as the second moment. Repeat for other templates $t^2, \cdots, t^K$. The resulting signature is a vector which is invariant and arbitrarily selective for the image patch. Step 1 corresponds to learning the tuning of the "simple" cells; step 2a corresponds to the response of each of the $|G|$ simple cell to a new image; step 2b corresponds to the output of one of the K complex cells.

3.4.5 Stable signatures

If $\Sigma(I) \in \mathbb{R}^p$ denotes the signature of an image, and $\|\Sigma(I) - \Sigma(I')\|_s$, $I, I' \in \mathcal{X}$, is a metric, we say that a *signature* Σ *is stable if it is Lipschitz continuous* (see [63]), that is

$$\|\Sigma(I) - \Sigma(I')\|_s \leq L \|I - I'\|, \quad \forall I, I' \in \mathcal{X}. \tag{8}$$

In this section we study the stability of the empirical signature $\Sigma(I) = (\mu^1_1(I), \ldots, \mu^K_N(I)) \in \mathbb{R}^p, p = NK$ provided with the metric (5) (together with (6)). In particular, we consider the case in which $\mu^k_n(I)$ is computed as in (**??**),(**??**). For the sake of simplicity here we consider the group G to be finite. By definition,

$$\begin{aligned} & \left\|\mu^k(I) - \mu^k(I')\right\|_{\mathbb{R}^N} \\ = & \sqrt{\sum_{n=1}^{N} \Big(\frac{1}{|G|} (\sum_{g\in G} \eta_n(\langle gI, t^k\rangle) - \sum_{g\in G} \eta_n(\langle gI', t^k\rangle)) \Big)^2} \end{aligned}$$

where $\mu^k(I) = (\mu_1^k(I), ..., \mu_N^k(I))$.
If the non linearities η_n are Lipschitz continuous, for all $n = 1, \ldots, N$, with Lipschitz constant L_{η_n}, it follows that

$$\begin{aligned}
& \left\|\mu^k(I) - \mu^k(I')\right\|_{\mathbb{R}^N} \\
\leq & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N}\Big(\sum_{g\in G}\eta_n(\langle gI, t^k\rangle) - \eta_n(\langle gI', t^k\rangle)\Big)^2} \\
\leq & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N}\Big(\sum_{g\in G}L_{\eta_n}|\,\langle gI, t^k\rangle - \langle gI', t^k\rangle\,|\Big)^2} \\
\leq & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N}\Big(\sum_{g\in G}L_{\eta_n}|\,\langle gI - I', t^k\rangle\,|\Big)^2} \\
\leq & \frac{1}{|G|}\sqrt{\sum_{n=1}^{N}L_{\eta_n}^2\sum_{g\in G}(|\,\langle g(I - I'), t^k\rangle\,|)^2},
\end{aligned}$$

where we used the linearity of the inner product and Jensen's inequality. Applying Schwartz inequality we obtain

$$\left\|\mu^k(I) - \mu^k(I')\right\|_{\mathbb{R}^N} \leq \frac{L_\eta}{|G|}\sqrt{\sum_{n=1}^{N}\sum_{g\in G}\|I - I'\|^2\left\|g^{-1}t^k\right\|^2}$$

where $L_\eta = \max_n(L_{\eta_n})$. If we assume the templates and their transformations to be normalize in the sense that $\left\|g^{-1}t^k\right\|/N,\ \forall g \in G$ then we finally have,

$$\left\|\mu^k(I) - \mu^k(I')\right\|_{\mathbb{R}^N} \leq L_\eta \left\|I - I'\right\|.$$

which is (8). In particular if $L_\eta \leq 1$ the map is non expansive. The same result holds if we further sum over K templates and divide by $1/K$ and if a compact group, rather than a finite group, is considered.

The above reasoning proves the following result.

Theorem 6. *Let $\|\Sigma(I) - \Sigma(I')\|_s = d_\mu(\mu^k(I), \mu^k(I'))$ and assume that $\mu_n^k(I) = \int dg\eta_n(\langle gI, t^k\rangle)$ for $n = 1, \ldots, N$ and $k = 1, \ldots, K$. Assume the templates to be normalized so that $\left\|g^{-1}t^k\right\|/N,\ \forall g \in G$ and $L_\eta = \max_n(L_{\eta_n}) \leq 1$. Then*

$$\|\Sigma(I) - \Sigma(I')\|_s \leq \|I - I'\|,$$

for all $I, I' \in \mathcal{X}$.

We note that above proof shows that it is important that the non-linearities are Lipschitz continuous with constants such that $L_\eta \leq 1$. Sigmoidal nonlinearities which are sufficiently smooth ensures stability, whereas their discontinuous limit of a binary threshold function does not.

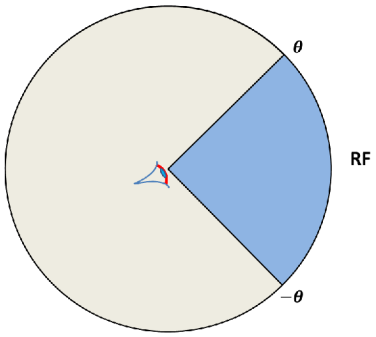


Figure 7: *A partially observable compact shift: an object is seen through a window in different positions. The object is fully contained in the window and is isolated (blank background). A compact group of periodic translations acts on it; only a part of the orbit is observable.*

3.4.6 Signatures for Partially Observable Groups (POG): Invariance, Uniqueness and Stability

This section outlines invariance, uniqueness and stability properties of the signature obtained in the case in which transformations of a (locally compact) group are observable only within a *window* (a *receptive field*) "over" the orbit. The prototypical case is an object moving around the observer along a circle (see Figure 7): then its image on the eye is subject to translation in the x, y coordinates of the image plane. Let us assume that the image of the object is fully contained within the window and is observable only within that window $[-\theta, +\theta]$. The object is subject to transformations which can be regarded as shifts in x on a torus: thus the group is compact but only partially observable. Every x, y translation in the visual field can be modeled in this way.

Following the above discussion we let G be a finite group and $G_0 \subseteq G$ a subset. The subset of transformations G_0 can be seen as the set of transformations that can be *observed*. A *local* signature associated to the partial observation of G can be defined considering

$$\mu_n^k(I) = \frac{1}{|G_0|} \sum_{g \in G_0} \eta_n(\langle gI, t^k \rangle), \tag{9}$$

and $\Sigma_{G_0}(I) = (\mu_n^k(I))_{n,k}$. This definition can be generalized to any locally compact group considering,

$$\mu_n^k(I) = \frac{1}{V_0} \int_{G_0} \eta_n(\langle gI, t^k \rangle) dg, \quad V_0 = \int_{G_0} dg. \tag{10}$$

Note that the constant V_0 normalizes the Haar measure, restricted to G_0, to define a probability distribution. The latter is the distribution of the images subject to the group transformations which are observable. The above definitions

can be compared to definitions (**??**) and (**??**) in the fully observable groups case. In the next paragraph we discuss the properties of the above signature. While stability and uniqueness follow essentially from the analysis of the previous section, invariance requires developing a new analysis.

3.4.7 Approximate Invariance of Signatures associated to POG

Since the group is only partially observable we cannot expect to have complete invariance, but only a form of *approximate* invariance. The following lemma makes the above statement precise.

Lemma 1. *Let $g' \in G$ and $G_0 \subset G$. If*

$$\eta_n(\langle gI, t^k \rangle) = 0, \forall g \in G_0 \Delta {g'}^{-1} G_0 \tag{11}$$

then,

$$\mu_n^k(I) = \mu_n^k(g'I)$$

Proof. Calculating the difference between the signature of the translated and not translated image we have

$$\begin{aligned} & \mu_n^k(I) - \mu_n^k(g'I) \\ = & \frac{1}{V_0} \int_{G_0} \eta_n(\langle gI, t^k \rangle) dg - \int_{G_0} \eta_n(\langle gg'I, t^k \rangle) dg \\ = & \frac{1}{V_0} \int_{G_0} \eta_n(\langle gI, t^k \rangle) - \int_{{g'}^{-1} G_0} \eta_n(\langle gI, t^k \rangle) dg \\ = & \frac{1}{V_0} \int_{G_0 \Delta {g'}^{-1} G_0} \eta_n(\langle gI, t^k \rangle) dg. \end{aligned}$$

□

The above result and the meaning of condition (11) is better explained considering a specific example.

Invariance: a simple example with translations and images in $1D$. Consider $1D$ signals and the group of translations so that $\mathcal{X} \subset L^2(\mathbb{R})$, $G = \mathbb{R}$. Let $T_\xi : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, $(T_\xi I)(\tau) = I(\tau - \xi)$, $I \in L^2(\mathbb{R}), \xi \in \mathbb{R}$ Assume that $G_0 = [-b, b]$, $b > 0$, then Eq. (10) becomes

$$\mu_n^k(I) = \frac{1}{2b} \int_{-b}^{b} d\xi \eta_n(\langle T_\xi I, t^k \rangle).$$

For the sake of simplicity assume $I, t^k \in \mathcal{X}$ to be compactly supported functions, with $\text{supp}(I) = [-a, a]$, $a < b$ and assume the support of t to be much smaller than that of I, so that $\mu_n^k(I) \approx \frac{1}{2b} \int_{-b}^{b} d\xi \eta_n(I(\xi))$. In this setting we are interested into characterizing for which $c \in \mathbb{R}$, $\mu_n^k(T_c I) = \mu_n^k(I)$. Following Lemma 1, we have to ensure that $\eta_n(\langle T_\xi I, t^k \rangle) \approx \eta_n(I(\xi)) = 0$ for

$\xi \in G_0 \Delta {g'}^{-1} G_0 = [-b, b] \cup [-b-c, b-c]$. If we assume $\text{supp}(\eta_n \circ I) \approx \text{supp}(I)$, then it is easy to see the condition is ensured for $c < b - a$. In words, the $\mu_n^k(I)$ are invariant as long as the object is isolated, and fully viewed through the observable window while undergoing small translations.

3.4.8 Uniqueness and Stability for POG signatures

A direct consequence of Theorem 1 is that *any two orbits with a common point are identical*. This follows from the fact that if $gI \in O_I$, $g'I \in O_{I'}$, $g, g' \in G$ is a common point of the orbits, then

$$g'I' = gI \;\Rightarrow\; I' = (g')^{-1} gI.$$

Thus the two images are one the transformed of the other: therefore $O_I = O_{I'}$. Suppose now that only a fragment of the orbits – the part within the window – is observable; the reasoning above is still valid since if the orbits are different or equal so must be any of their "corresponding" parts. In particular, it sufficient to observe through the window one complete image of the object in one orbit: if it is the same as a single image in another orbit (or part of another orbit) then the two orbits are the same. The viceversa (when part of the orbit is different from part of another orbit) implies that the orbits are different but only if the parts correspond (as it is our case of a "fixed" window).
Finally, we discuss stability of POG signatures, noting that the reasoning in 3.4.5 can be repeated without any significative change in the case that a subset of the transformations is considered. In fact, only the normalization over the transformations is modified accordingly.

3.5 Hierarchical architectures

We focused so far on the basic "simple and complex cells" modules. Architectures consisting of such modules can be single-layers as well as multi-layers and hierarchical (see Fig. 3.5.5). One-layer networks can achieve invariance to *global* transformations of the whole image (exact invariance if the transformations are a subgroup of the affine group in $\mathbb{R}^2$) while providing a unique global signature which is stable with respect to small perturbations of the image. One could imagine local and global one-layer architectures used in the same visual system. We conjecture that: *Hierarchical architectures, as depicted in Fig. 3.5.5, are the only networks built from simple-complex modules satisfying simultaneously the following set of properties*:

1. *Optimization of local connections* and optimal reuse of computational elements. Despite the high number of synapses on each neuron it would be impossible for a complex cell to pool information across all the simple cells needed to cover an entire image.

2. *Compositionality*. A hierarchical architecture provides signatures of larger and larger patches of the image in terms of lower level signatures. Because of this, it can access memory in a way that matches naturally with

the linguistic ability to describe a scene as a whole and as a hierarchy of parts.

3. *Approximate factorization*. In architectures such as the network sketched in Fig. 3.5.5, approximate invariance to transformations specific for an object class can be learned and computed in different stages. This property may provide an advantage in terms of the sample complexity of multi-stage learning [81]. For instance, approximate class-specific invariance to pose (e.g. for faces) can be computed on top of a translation-and-scale-invariant representation [57, 54]. Thus the implementation of invariance can, in some cases, be"factorized" into different steps corresponding to different transformations. (see also [4, 93] for related ideas).

Probably all three properties together are the reason evolution developed hierarchies. It is interesting that with respect to the range of invariance the following property holds (SI4 and [80]), for multilayer architectures such as in Fig. 3.5.5, in which the output of a layer is defined as covariant if it transforms in the same way as the input. *For a given transformation of an image or part of it, the signature from complex cells at a certain layer is either invariant or covariant with respect to the group of transformations; if it is covariant there will be a higher layer in the network at which it is invariant*. This property predicts a *stratification* of ranges of invariance in the ventral stream: invariances should appear in a sequential order meaning that smaller transformations will be invariant before larger ones, in earlier layers of the hierarchy (see [40]). In addition to the issues of sample complexity and connectivity, one-stage architectures are unable to capture the hierarchical organization of the visual world where scenes are composed of objects which are themselves composed of parts. Objects (i.e. parts) can move in a scene relative to each other without changing their identity and often changing only in a minor way the scene (i.e., the object). Thus global and local signatures from all levels of the hierarchy must be able to access memory in order to enable the categorization and identification of whole scenes as well as of patches of the image corresponding to objects and their parts. Fig. 9 show examples of invariance and stability for wholes and parts. Fig. 3.5.5 sketches a $1D$ hierarchical architecture of "simple-complex cells": a $\bigwedge$ module provides uniqueness, invariance and stability at different levels over increasing ranges from bottom to top. HMAX [85, 67, 91], Convolutional Neural Networks [22, 77, 51, 50, 52, 7] and Deep Feedforward Neural Networks [1, 47, 48] are examples of this class (see [94, 98]) of architectures—as is, possibly, the feedforward organization of the ventral steam. Notice that the architecture based on the simple-complex module cannot deal with non-affine transformations (in $\mathbb{R}^2$) such as rotations in depth or changes of expression of a face. Approximate invariance, however, can be obtained with the same basic modules for certain "nice" classes of objects that share a similar 3D structure [57, 54].

The property of compositionality discussed above is related to the efficacy of hierarchical architectures vs one-layer architectures in dealing with the problem of partial occlusion and the more difficult problem of clutter and context in

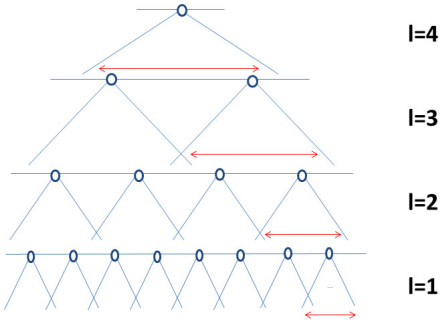


Figure 8: *A hierarchical architecture built from simple-complex cell modules. Circles represent complex cells and red double arrows pooling ranges at different layers,* $\ell = 1, 2, 3, 4$.

object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks [90]. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g. recurrent) architectures are needed for human-level recognition in clutter (see for instance [8, 27, 26]) and for other aspects of human vision. It is likely that most of the circuitry of the visual cortex is required by these recurrent computations, not considered in this paper.

3.5.1 The basic idea: wholes and parts

Consider a hierarchical, 1D,"discrete" architecture such as in Figure 10. We assume that each of the nodes $\bigwedge$ is invariant for shifts of a pattern within its receptive field; we also assume that the output layer at each level is covariant (see later). Assume that the receptive field of each node overlaps by $\frac{1}{2}$ the neighboring ones on each side at each level of the architecture. Start from the highest level. Assume that deformations are local translation of a patch. Consider now the following examples. First assume that there is a minimal distance between patches (A and B in the figure) of 3 pixels. It is easy to see that each of A and B has a distinct signature at the first level in 2 different nodes. Each of A or B can shift by arbitrary amounts without changing their signature. So each one is an "object" at the first level in terms of their signatures, invariant to shifts. They compose a new object (AB) at the second level if their distance is between $3 \leq d \leq 4$ and so on for higher levels. This is a situation in which A and B are each a part – like an eye and a mouth in a face, each part is invariant to shifts, the object AB is also invariant and is tolerant to "small" deformations (distance between A and B). There other cases. For instance, assume that the

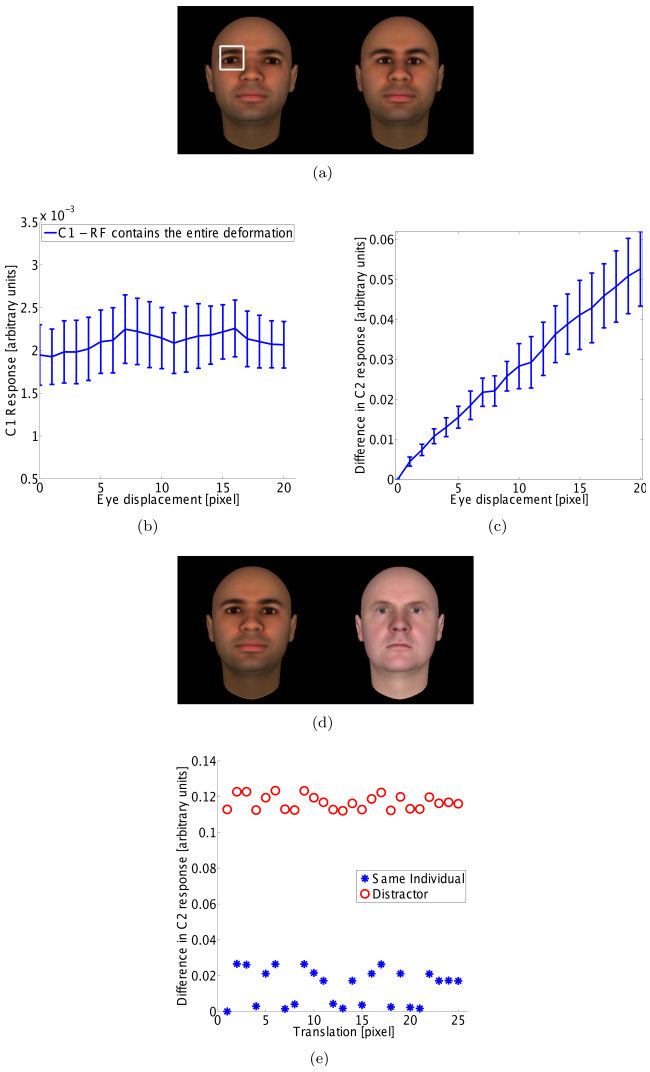


Figure 9: *(a) shows the reference image on the left and a local deformation of it (the eyes are closer to each other) on the right; (b) shows that a complex cell at layer 1 (c_1) signature from complex cells whose receptive fields covers the left eye is invariant to the deformation; in (c) Complex cells at layer 2 (c_2) whose receptive fields contain the whole face are (Lipschitz) stable with respect to the deformation. In all cases just the Euclidean norm of the response is shown on the y axis; the c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. Two stimuli (d) are presented at various location in the visual field. We consider the response of a set of c_2 cells with the same receptive field (the whole image) but different tuning. The Euclidean distance between the c_2 response vector corresponding to the two images and a reference response vector is reported (e). The response is invariant to global translation and discriminative (between the two faces). In this example the c_2 units represent the top of a hierarchical, convolutional architecture. The images we used were 200×200 pixels*

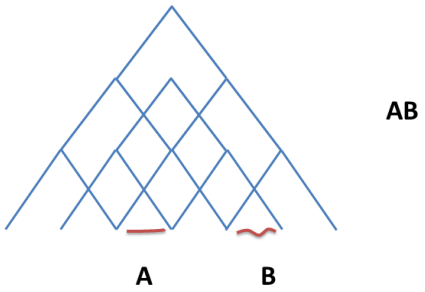


Figure 10: *Each of the nodes $\bigwedge$ is invariant for shifts of a pattern within its receptive field; we also assume that the output layer at each level is covariant.*

distance between A and B is $1 \leq d \leq 3$. Then for each shift there is always a $\bigwedge$ which "sees" A, another one which "sees" B and a third one which "sees" AB. In this case AB are parts of an object AB, all represented in an invariant way at the first level. However, the object AB is not tolerant to deformations of the distance between A and B (this happens only if objects are represented at higher levels than parts in the hierarchy). Finally, if the distance between A and B is less than 1 then AB is always an object at all levels. It is intriguing to *conjecture* that this kind of properties may be related to the minimal distances involved in crowding.

3.5.2 Hierarchical Architectures

So far have studied the invariance, uniqueness and stability properties of signatures both in the case when a whole group of transformations is observable (see (**??**) and (**??**)) as well as in the case in which it is only partially observable (see (9) and (10)). We now discuss how the above ideas can be iterated to define a multilayer architecture. Consider first the case when G is finite. Given a subset $G_0 \subset G$, we can associate a *window* gG_0 to each $g \in G$. Then, we can use definition (9) to define for each window a signature $\Sigma(I)(g)$ given by the measurements,

$$\mu_n^k(I)(g) = \frac{1}{|G_0|} \sum_{\bar{g} \in gG_0} \eta_n(\langle \bar{g}I, t^k \rangle).$$

For fixed n, k, a set of measurements corresponding to different windows can be seen as a $|G|$ dimensional vector. A signature $\Sigma(I)$ for the whole image is obtained as *a signature of signatures*, that is a collection of signatures $(\Sigma(I)(g_1), \ldots, \Sigma(I)(g_{|G|})$ associated to each window.
The above discussion can be easily extended to continuous (locally compact) groups considering,

$$\mu_n^k(I)(g) = \frac{1}{V_0} \int_{gG_0} d\bar{g} \eta_n(\langle \bar{g} I, t^k \rangle), \quad V_0 = \int_{G_0} d\bar{g},$$

where, for fixed n, k, $\mu_n^k(I) : G \to \mathbb{R}$ can now be seen as a function on the group. In particular, if we denote by $K_0 : G \to \mathbb{R}$ the indicator function on G_0, then we can write

$$\mu_n^k(I)(g) = \frac{1}{V_0} \int d\bar{g} K_0(\bar{g}^{-1} g) \eta_n(\langle \bar{g} I, t^k \rangle).$$

The signature for an image can again be seen as a collection of signatures corresponding to different windows, but in this case it is a function $\Sigma(I) : G \to \mathbb{R}^{NK}$, where $\Sigma(I)(g) \in \mathbb{R}^{NK}$, is a signature corresponding to the window G_0 "centered" at $g \in G$.

The above construction can be iterated to define a hierarchy of signatures. Consider a sequence $G_1 \subset G_2, \ldots, \subset G_L = G$. For $h : G \to \mathbb{R}^p$, $p \in \mathbb{N}$ with an abuse of notion we let $gh(\bar{g}) = h(g^{-1}\bar{g})$. Then we can consider the following construction.

We call *complex cell operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $c_\ell(I) : G \to \mathbb{R}^{NK}$ where

$$c_\ell^{n,k}(I)(g) = \frac{1}{|G_\ell|} \sum_{\bar{g} \in g G_\ell} \eta_n \left(s_\ell^k(I)(\bar{g}) \right), \tag{12}$$

and *simple cell operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $s_\ell(I) : G \to \mathbb{R}^K$

$$s_\ell^k(I)(g) = \langle g c_{\ell-1}(I), t_\ell^k \rangle \tag{13}$$

with t_ℓ^k the k^{th} template at layer ℓ and $c_0(I) = I$. Several comments are in order:

- beside the first layer, the inner product defining the simple cell operator is that in $L^2(G) = \{h : G \to \mathbb{R}^{NK} \mid \int dg |h(g)|^2 < \infty\}$;
- The index ℓ corresponds to different layers, corresponding to different subsets G_ℓ.
- At each layer a (finite) set of templates $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ ($\mathcal{T}_0 \subset \mathcal{X}$) is assumed to be available. For simplicity, in the above discussion we have assumed that $|\mathcal{T}_\ell| = K$, for all $\ell = 1, \ldots, L$. The templates at

layer ℓ can be thought of as *compactly supported functions*, with support much smaller than the corresponding set G_ℓ. Typically templates can be seen as image patches in the space the complex operator response, that is $t_\ell = c_{\ell-1}(\bar{t})$ for some $\bar{t} \in \mathcal{X}$.

- Similarly we have assumed that the number of non linearities η_n, considered at every layer, is the same.

Following the above discussion, the extension to continuous (locally compact) groups is straightforward. We collect it in the following definition.

Definition 6 (Simple and complex response)**.** *For $\ell = 1, \dots, L$, let $\mathcal{T}_\ell = (t_\ell^1, \dots, t_\ell^K) \subset L^2(G)$ (and $\mathcal{T}_0 \subset \mathcal{X}$) be a sequence of template sets. The complex cell operator at layer ℓ maps an image $I \in \mathcal{X}$ to a function $c_\ell(I) : G \to \mathbb{R}^{NK}$; in components*

$$c_\ell^{n,k}(I)(g) = \frac{1}{V_\ell} \int d\bar{g} K_\ell(\bar{g}^{-1} g) \eta_n \left(s_\ell^k(I)(\bar{g}) \right), \; g \in G \tag{14}$$

where K_ℓ is the indicator function on G_ℓ, $V_\ell = \int_{G_\ell} d\bar{g}$ and where

$$s_\ell^k(I)(g) = \left\langle g c_{\ell-1}(I), t_\ell^k \right\rangle, \quad g \in G \tag{15}$$

($c_0(I) = I$) is the simple cell operator at layer ℓ that maps an image $I \in \mathcal{X}$ to a function $s_\ell(I) : G \to \mathbb{R}^K$.

Remark 2. *Note that eq.* (14) *can be written as:*

$$c_\ell^{n,k}(I) = K_\ell * \eta_n(s_\ell^k(I))$$

where $*$ is the group convolution.

Hierarchical architecture: $1D$ **translation case.** We specialize the above definitions to the case of images viewed as one dimensional signals $\mathcal{X} = L^2(\mathbb{R})$ and consider the transformations to be $1D$ translations, so that gI is given by $T_\tau I(x) = I(x - \tau)$. Note that in this case we can identify $G = \mathbb{R}$ and $L^2(G) = L^2(\mathbb{R}) = \mathcal{X}$. The sets $G_1 \subset G_2 \subset, \dots, \subset G_L = G$ can now be seen as (centered) intervals $P_\ell \subseteq \mathbb{R}$, where $P_{\ell-1} \subseteq P_\ell$, and $V_\ell = |P_\ell|$, $\forall \ell = 1, ..., L$, $P_L = \mathbb{R}$. In particular, we consider, for simplicity, the case where there is just one template, t, and one non linear function, η, per layer. In this case, we can rewrite definition (6) as follows. The complex response operator $c_\ell : \mathcal{X} \to \mathcal{X}$, is iteratively defined as,

$$c_\ell(I)(\xi) = \frac{1}{|P_\ell|} \int_{\mathbb{R}} d\tau \; K_\ell(\xi - \tau) \eta(s_\ell(I)(\tau)) \tag{16}$$

and the simple response operator $s_\ell : \mathcal{X} \to \mathcal{X}$ as

$$s_\ell(I)(\xi) = \langle T_\xi c_{\ell-1}(I), t \rangle, \; I, \; t \in \mathcal{X}, \xi \in \mathbb{R} \tag{17}$$

where $c_0(I) = I$.

For the sake of simplicity, in the reminder of this section we focus on the above special case, however most results hold in the general setting.

3.5.3 Property 1 :covariance of the c_ℓ response

We recall that a map is **covariant** iff

$$\Sigma(gI) = g^{-1}\Sigma(I), \quad \forall g \in G, I \in \mathcal{X}.$$

In the one dimensional setting, discussed at the end of the previous section, we have $\Sigma(T_\xi(I)) = T_{-\xi}(\Sigma(I)),\ \forall\ I \in \mathcal{X},\ \xi \in \mathbb{R}$, with $\Sigma : \mathcal{X} \to \mathcal{X}$.
In this paragraph we prove a form of covariance for the complex response, i.e.

Proposition 3. *The operator c_ℓ is covariant with respect to translations, that is*

$$\begin{aligned} c_\ell(T_{\bar\tau} I) &= T_{-\bar\tau}(c_\ell(I)),\ \ell\ odd\ \forall \bar\tau \in \mathbb{R} \\ c_\ell(T_{\bar\tau} I) &= T_{\bar\tau}(c_\ell(I)),\ \ell\ even\ \forall \bar\tau \in \mathbb{R} \end{aligned}$$

The proof proceeds as following. For $\ell = 1$ the covariance of the $s_1(I)$ function follows from:

$$\begin{aligned} s_1(T_{\bar\tau} I)(\tau) &= \langle T_\tau T_{\bar\tau} I, t\rangle = \langle T_{\tau+\bar\tau} I, t\rangle = s_1(I)(\tau+\bar\tau) \\ &= (T_{-\bar\tau} s_1(I))(\tau) \end{aligned}$$

The covariance of the $c_1(I)$ follows from:

$$\begin{aligned} & c_1(T_{\bar\tau} I)(\tau) = \frac{1}{|P_1|}\int K_1(\tau-\tilde\tau)\eta(s_1(T_{\bar\tau} I))(\tilde\tau)d\tilde\tau \\ = & \frac{1}{|P_1|}\int K_1(\tau-\tilde\tau)\eta(s_1(I))(\tilde\tau+\bar\tau)d\tilde\tau \\ = & \frac{1}{|P_1|}\int K_1(\tau+\bar\tau-\hat\tau)\eta(s_1(I)(\hat\tau))d\hat\tau \\ = & c_1(I)(\tau+\bar\tau) = (T_{-\bar\tau} c_1(I))(\tau) \end{aligned}$$

where on the second line we used the covariance property of $s_1(I)$ and on the third we used the change of variable $\hat\tau = \tilde\tau + \bar\tau$.
Let us now calculate the complex and simple response at the second layer for a translated image. For $s_2(I)$ we have

$$\begin{aligned} s_2(T_{\bar\tau} I)(\tau) &= \langle T_\tau c_1(T_{\bar\tau} I), t\rangle = \langle T_\tau T_{-\bar\tau} c_1(I), t\rangle \\ &= T_{\bar\tau} s_2(I)(\tau) = s_2(I)(\tau-\bar\tau) = T_{\bar\tau} s_2(I)(\tau) \end{aligned}$$

where we used the property $c_1(T_{\bar\tau})(I) = T_{-\bar\tau} c_1(I)$.
For $c_2(I)$

$$\begin{aligned} c_2(T_{\bar\tau} I)(\tau) &= \frac{1}{|P_2|}\int K_2(\tau-\tilde\tau)\eta(s_2(T_{\bar\tau} I))(\tilde\tau)d\tilde\tau \\ &= \frac{1}{|P_2|}\int K_2(\tau-\tilde\tau)\eta(s_1(I))(\tilde\tau-\bar\tau)d\tilde\tau \\ &= \frac{1}{|P_2|}\int K_2(\tau-\bar\tau-\hat\tau)\eta(s_2(I)(\hat\tau))d\hat\tau \\ &= c_2(I)(\tau-\bar\tau) = (T_{\bar\tau} c_1(I))(\tau) \end{aligned}$$

Note that the covariance property of the complex response is different depending on the layer. The above calculation suggests to prove the theorem by induction. We omit the proof since it is a straightforward repetition of the case $\ell = 1, 2$ above.

3.5.4 Property 2: partial and global invariance of c_ℓ response (whole and parts)

We now prove that the functions c_ℓ are locally invariant to translations, i.e. invariant within the restricted range of the pooling if the non linearity η satisfies certain conditions. We further prove that the range of invariance increases from layer to layer in the hierarchical architecture. The fact that for an image, in general, no more global invariance is guaranteed allows, as we will see , a novel definition of "parts" of an image.
Let us first reformulate the result in Lemma 1 in the context of a hierarchical architecture.

Proposition 4. *Let c_ℓ be the complex response at layer ℓ and $I \in \mathcal{X}$ then:*

$$c_\ell(T_{\bar\tau} I)(\tau) = c_\ell(I)(\tau), \quad I \in \mathcal{X}, \tag{18}$$

if

$$\eta(s_l(I)(\tau)) = \eta(\langle T_\tau c_{\ell-1}(I), t\rangle) = 0 \quad \tau \in P_\ell \Delta T_{\bar\tau} P_\ell.$$

Proof. The proof follows the reasoning done in par. 3.4.7 with I substituted by $s_l(I)$. We prove the statement for ℓ odd and use the covariance property found in proposition 3. A similar proof holds for ℓ even.
Let the pooling at layer l be achieved by a characteristic function on the interval P_ℓ; we have

$$\begin{aligned}
&c_\ell(T_{\bar\tau} I)(\tau) - c_\ell(I)(\tau) = c_\ell(I)(\tau + \bar\tau) - c_\ell(I)(\tau) \\
&= \frac{1}{|P_\ell|} \int_{\mathbb{R}} \Big(K_\ell(\tau + \bar\tau - \tilde\tau) - K_\ell(\tau - \tilde\tau) \Big) \eta(s_\ell(I)(\tilde\tau)) d\tilde\tau \\
&= \frac{1}{|P_\ell|} \Big(\int_{T_{\bar\tau} P_\ell} \eta(s_\ell(I)(\tilde\tau)) d\tilde\tau - \int_{P_\ell} \eta(s_l(I)(\tilde\tau)) d\tilde\tau \Big) \\
&= \frac{1}{|P_\ell|} \int_{P_\ell \Delta T_{\bar\tau} P_\ell} \eta(s_\ell(I)(\tilde\tau)) d\tilde\tau \\
&= \frac{1}{|P_\ell|} \int_{P_\ell \Delta T_{\bar\tau} P_\ell} \eta(\langle T_{\bar\tau} c_{\ell-1}(I), t\rangle) d\tilde\tau
\end{aligned}$$

where on the first line we used the covariance properties of the function $c_\ell(I)$. The above expression is zero if $\eta(\langle T_\tau c_{\ell-1}(I), t\rangle) = 0 \; \forall \tau \in P_\ell \Delta T_{\bar\tau} P_\ell$. □

We can give now a formal definition of *object part* as the subset of the signal I whose complex response, at layer ℓ, is invariant under transformations in the range of the pooling at that layer.
This definition is consistent since the invariance is increasing from layer to

layer (as formally proved below) therefore allowing bigger and bigger parts. Consequently for each transformation there will exists a layer $\bar{\ell}$ such that any signal subset will be a part at that layer. Before formulating this result in the form of a theorem let us introduce a new definition: an operator $f : \mathcal{X} \to \mathcal{X}$ is called $r-$invariant for $I \in \mathcal{X}$ with respect to translations if:

$$\Sigma(T_\xi(I)) = \Sigma(I), \quad \forall \xi \in [0, r], \quad r \in \mathbb{R}.$$

We can now state the following:

Theorem 7. *Whole and parts. Let $I \in \mathcal{X}$ (an image or a subset of it) and $c_\ell : \mathcal{X} \to \mathcal{X}$ the complex response at layer ℓ. Let $P_0 \subset \cdots \subset P_\ell \subset \cdots \subset P_L = \mathbb{R}$ a set of nested intervals. Suppose $\eta(0) = 0$ and that the template t and the complex response at each layer has finite support. Then $\forall \bar{\tau} \in \mathbb{R}$ c_ℓ is $\bar{\tau}-$invariant for some $\ell = \bar{\ell}$, i.e.*

$$c_m(T_\tau I) = c_m(I), \quad \exists\, \bar{\ell} \;\; s.t. \;\; \forall m \geq \bar{\ell}, \; \forall \tau \in [0, \bar{\tau}].$$

Proof. We prove the statement for ℓ odd and use the covariance property found in proposition 3. A similar proof holds for ℓ even. As proved in Lemma 1 the complex response $c_\ell(I)(0)$ (we choose $\tau = 0$ without loss of generality) is invariant for a translation of the image in the range $[0, \bar{\tau}]$ iff

$$\eta(\langle T_{\tau'} c_{\ell-1}(I), t\rangle) = \eta(\langle c_{\ell-1}(I), T_{-\tau'} t\rangle) = 0, \quad \forall \tau' \in P_\ell \Delta T_{\bar{\tau}} P_\ell. \tag{19}$$

Thus the theorem is proved if for any choice of $\bar{\tau}$ there exists an odd $\bar{\ell}$ such that the condition above is satisfied. Since $P_0 \subset \cdots \subset P_\ell \subset \cdots \subset P_L = \mathbb{R}$ there will exists (see, e.g., Fig. 11), being the supports of t and $c_{\ell-1}$ finite, for any fixed $\bar{\tau} \in \mathbb{R}^+$, an $\bar{\ell}$ such that

$$supp(T_{-\tau'} t) \cap supp(c_{\bar{\ell}-1}(I)) = \emptyset \Rightarrow \left\langle T_{\tau'} c_{\bar{\ell}-1}(I), t \right\rangle = 0.$$

$\forall \tau' \in P_{\bar{\ell}} \Delta T_{\bar{\tau}} P_{\bar{\ell}}$. Being $\eta(0) = 0$ we have $\eta(\langle T_{\tau'} c_{\bar{\ell}-1}(I), t\rangle) = 0, \forall \tau' \in P_{\bar{\ell}} \Delta T_{\bar{\tau}} P_{\bar{\ell}}$, i.e. the response is invariant at layer $\bar{\ell}$. □

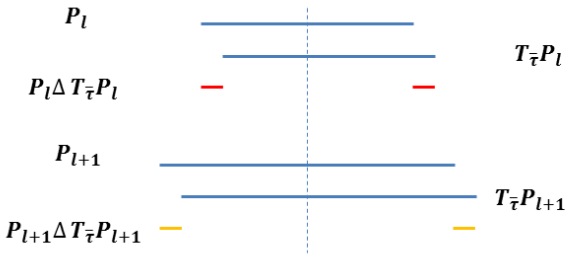


Figure 11: *Behavior of the symmetric difference for growing layer number .*

The same proof can be repeated in the case of a locally compact group G.

3.5.5 Property 3: stability of the c_ℓ response

Using the definition of stability given in 3.4.5, in the simple case where we have one template, t, and one non linearity, η we can formulate the following theorem characterizing stability for the complex response:

Theorem 8. *Stability. Let $I, I' \in \mathcal{X}$ and $c_\ell : \mathcal{X} \to \mathcal{X}$ the complex response at layer l. Let the nonlinearity η a Lipschitz function with Lipschitz constant $L_\eta \leq 1$. Then*

$$\|c_\ell(I) - c_\ell(I')\| \leq \|I - I'\|, \ \forall\ \ell, \ \forall\ I, I' \in \mathcal{X}. \tag{20}$$

Proof. Using the definition of complex response in eq. (53) and remark 2 we can write the l.h.s. of eq. (20) as:

$$\|c_\ell(I) - c_\ell(I')\| = \frac{1}{|P_\ell|} \|K_\ell * [\eta(s_\ell(I)) - \eta(s_\ell(I'))]\|.$$

Using the inequality $\|f * g\| \leq \|f\|_1 \|g\|, \ f, g \in \mathcal{X}$,

$$\|c_\ell(I) - c_\ell(I')\| \leq \frac{1}{|P_\ell|} \|K_\ell\|_1 \|\eta(s_\ell(I)) - \eta(s_\ell(I'))\|.$$

Since K_ℓ is, by construction, the characteristic function of the interval P_ℓ we have $\|K_\ell\|_1 / |P_\ell| < 1$ and being η a Lipschitz function

$$\|c_\ell(I) - c_\ell(I')\| \leq \|\eta(s_\ell(I)) - \eta(s_\ell(I'))\| \leq L_\eta \|s_\ell(I) - s_\ell(I')\|.$$

Applying the definition of the simple response, $s_\ell(I)(\tau) = \langle T_\tau c_{\ell-1}(I), t\rangle = \langle c_{\ell-1}(I), T_{-\tau} t\rangle$ and the Schwartz inequality we have

$$\begin{aligned} \|c_\ell(I) - c_\ell(I')\| &\leq L_\eta \|T_{-\tau} t\| \|c_{\ell-1}(I) - c_{\ell-1}(I')\| \\ &= L_\eta \|c_{\ell-1}(I) - c_{\ell-1}(I')\|. \end{aligned}$$

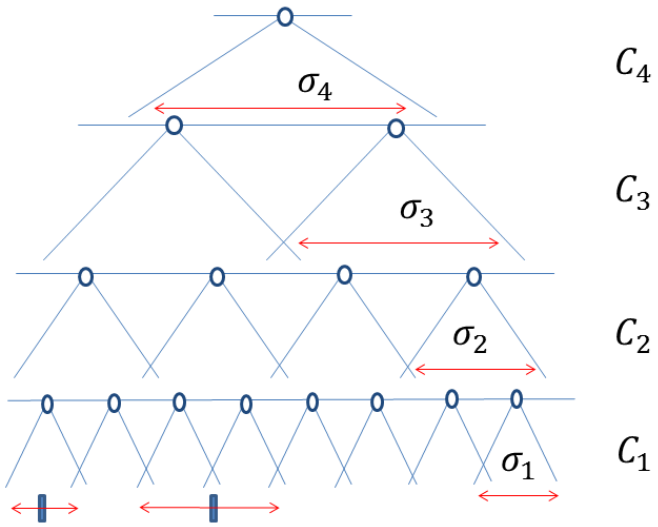
where we used the invariance to translations of the norm and supposed normalized templates, $\|t\| = 1$.
Repeating the reasoning l times we finally obtain:

$$\|c_\ell(I) - c_\ell(I')\| \leq L_\eta^l \|I - I'\|.$$

If $L_\eta \leq 1$ the statement of the theorem follows. □

Remark 3. *The above definitions and the demonstrations in the next paragraphs are done for one dimensional square integral signals undergoing translation transformations in one dimension. They can be generalized to square integral signals on locally compact groups $I \in L^2(G, dg)$ undergoing group transformations.*

Pictorially, indicating each function c_n with a $\bigwedge$ we can consider a network composed by different receptive fields $\bigwedge$:



c_n is the complex cell response at layer n and σ_n may be equal or larger than σ_{n-1}.
Notice that the patch of the image seen at layer n is at least as large as the patch seen at level $n-1$, that is $\sigma_{eff}^n \geq \sigma_{eff}^{n-1}$. In general σ_{eff}^n increases (rapidly) with n where with σ_{eff} we mean the image part effectively seen by a complex response at layer n.

Remark 4. *Notice that many non-linear functionals are so-called space- or time-invariant, e.g. NL-L systems, Volterra series, etc.. In this paper, we assume that cortical layers in visual cortex can be modeled by linear convolutions, which are trivially covariant, followed by memoryless non-linearities, which maintain covariance.*

Remark 5. *In principle, the arguments of these sections apply also to scale and rotation under the assumption that the network is X-invariant (instead of simply shift-invariant). If the network treats scale or rotation in the same way as x, y (with convolution and local pooling) the same arguments should apply. In practice, as we will show, in the hierarchical architecture after the first layer all transformations can be approximated as shifts within a 4-cube of wavelets (see later).*

Remark 6. *If in 4 instead of choosing the characteristic function of the interval P_n we*

use a Gaussian function $exp(-x^2/2\sigma_n)$ a similar result is obtained:

$$c_n(T_{\bar{\tau}}I)(\tau) = c_n(I)(\tau) + O\Big(\frac{\bar{\tau}^2}{2\sigma_n^2}\Big), \;\; \forall \bar{\tau} \in [0, \sigma_n], \; \tau \in \mathbb{R}.$$

Remark 7. *There are multiple scales (at each layer). We can think of them as different resolution units corresponding to different sizes of complex cells – like multiple lattices of photoreceptors of different sizes and different separations. The size of the complex cells also increases from layer to layer and defines how large a part is at each layer – from small parts in the first layer to parts of parts in intermediate layers, to whole objects and scenes at the top. Notice that the term parts here really refers to patches of the image. Notice that our theory may provide a* **novel definition of Part** *as a patch which has an invariant signature – at some level of the architecture – under local affine transformations. An interesting conjecture of invariance of signatures here is that the signature of a part at some level of the architecture may remain the same at another level when the part is seen under different conditions (think of an eye as part of a face and the same eye in isolation, occupying the whole image). The question is about which properties of the architecture are required for this.*

3.5.6 A hierarchical architecture: locally compact groups

In the following we are going to extend the reasoning done in the previous paragraphs to a general transformation of a locally compact group G implemented by the operator

$$T_g : \mathcal{X}/\mathcal{Y} \to \mathcal{Y}, \;\; (T_g I)(\tau) = I(g\tau), \; I \in \mathcal{X}/\mathcal{Y}, g \in G.$$

where $\mathcal{X}$ and $\mathcal{Y}$ are defined below among other basic objects:

- $\mathcal{X} = L^2(\mathbb{R}^2)$.
- $\mathcal{Y} = L^2(G, dg)$, where dg is the group invariant Haar measure.
- $\mathcal{T} \subset \mathcal{X}/\mathcal{Y}$, $|\mathcal{T}| < \infty$, the template set.
- $\eta : \mathcal{Y} \to \mathcal{Y}$ a non-linear function.
- $K_n : \mathcal{Y} \to \mathcal{Y}$ the characteristic function of the intervals $P_1 \subseteq ... \subseteq P_n$, $P_i \subset \mathcal{Y}$ or Gaussians with σ_n width.

The definitions of simple and complex response are similar to those given for the one dimensional translation group. However there is a major difference, although irrelevant for the covariance, invariance properties of the construction: the first simple response is an operator that maps the image space $\mathcal{X}$ into $\mathcal{Y}$; higher order responses instead are operators defined from $\mathcal{Y}$ into itself.

Definition 7. Simple and complex response
The complex response operator $c_n : \mathcal{Y} \to \mathcal{Y}$, is iteratively defined as:

$$c_n(I)(\xi) = (K_n * \eta(s_n(I)))(\xi) = \langle K_n, T_g \eta(s_n(I)) \rangle \tag{21}$$

in terms of the simple response operator $s_n : \mathcal{X}/\mathcal{Y} \to \mathcal{Y}$:

$$s_n(I)(\xi) = (c_{n-1}(I) * t)(\xi) = \langle c_{n-1}(I), T_\xi t \rangle \, , \; t \in \mathcal{T}, \; I \in \mathcal{X}, g \in G \tag{22}$$

where $c_0(I) \equiv I$.

Same kind of results obtained before for covariance, invariance and robustness to local perturbations can be obtained. For details on invariance properties of complex responses for the similitude group see 10.7.

3.6 Factorization of Invariances

The first version of [78] we conjectured that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g. the subgroup of translations of the affine group) and then adding the invariance w.r.t. another subgroup (e.g. rotations). It also assumed that factorization of the range of invariance was possible. The second version of the memo scrapped the first conjecture while endorsing the second. We show here that with the specific architecture we assume it is indeed mostly impossible to factorize effectively (to reduce number of templates to be learned) subgroups but that it is possible to factorize the range of invariance. Questions that remains open are (a) under which conditions approximate factorization is possible and (b) whether factorization helps reducing the sample complexity of unsupervised learning.

3.7 Preliminaries

We begin by considering an example where $\mathcal{X} \subset L^2(\mathbb{R}^2)$. We are interested in invariance w.r.t. $2D$ translations. We show how invariance can be achieved by first computing an invariant signature with respect to the $x-$translation and to the $y-$translation and discuss how a similar strategy can be used with more general groups (and subgroups). We then show that this factorization cannot be used to learn separately x and y invariance through independent templates.

We consider an architecture composed by two layers. Each layer has an associated set of templates, $t^1, \ldots, t^K \in \mathcal{X}$ and $\bar{t}^1, \ldots, \bar{t}^K \in \mathcal{X}$ (we assume for simplicity the number of templates is the same). We also assume to have just one non-linearity per layer and that the set of translations have been suitably discretized. We denote by T_x, T_y, $x, y \in \mathbb{R}$, the action of the translations on any $I \in \mathcal{X}$.
The first layer defines a signature $\mu : \mathcal{X} \to \mathbb{R}^K$ as

$$\mu^k(I) = \sum_{x' \in \mathbb{R}} \eta(\langle T_{x'} I, t^k \rangle), \quad k = 1, ..., K. \tag{23}$$

where clearly $\langle T_{x'} I, t^k \rangle = \langle I, T_{x'}^{-1} t^k \rangle$, since the representation of the translation is unitary. The signature at second layer is $\nu : \mathcal{X} \to \mathbb{R}^L$

$$\nu^l(I) = \sum_{y' \in \mathbb{R}} \eta(\langle \mu(T_{y'} I), s^l \rangle), \quad l = 1, ..., K \tag{24}$$

where the set of templates $s^1, \ldots, s^K$ can be thought of as a the signatures of a set of templates with respect to the first layer i.e. $s^i = \mu(\bar{t}^i)$, for $\bar{t}^i \in \mathcal{X}$, $i = 1, \ldots, K$, with $s^i \in \mathbb{R}^K$.

Indeed, we can show that ν is invariant to $2D$ translations $\nu(I) = \nu(T_x T_y I)$. Recalling equation (**??**), we note that, $\mu(T_x T_y I) = \mu(T_y I)$ since μ is defined by a group integration with respect to the $x-$translation. Then

$$\nu^l(T_x T_y I) = \sum_{y' \in \mathbb{R}} \eta(\langle \mu(T_{y'} T_y I), s^l \rangle) = \nu^l(T_y I), \quad l = 1, ..., K \tag{25}$$

and finally $\nu^l(T_x T_y I) = \nu^l(T_y I) = \nu(I)$ since ν is defined by a group integration with respect to the $y-$translation.

An inspection of the above reasoning shows that a similar factorization holds for many transformations beyond translations. Indeed, we have the following result.

Lemma 2 (Factorization of Invariances)**.** *Let G, R be two locally compact groups with Haar measures dg, dr respectively. Let $\mu : \mathcal{X} \to \mathbb{R}^K$ and $\nu : \mathcal{X} \to \mathbb{R}^L$ be defined by,*

$$\mu^k(I) = \int dg \eta(\langle gI, t^k \rangle), \qquad \nu^l(I) = \int dr \eta(\langle \mu(rI), s^l \rangle), \tag{26}$$

then $\nu(grI) = \nu(I)$.

We add one final remark.

Remark 8. *We briefly comment on the order in which transformations need to be applied to an image to still have invariance. Clearly, factorization of invariances happens in the architecture with an order given by the group integration at each layer, so in general it might not be true that $\nu(rgI) = \nu(I)$. However, invariance clearly still holds if the groups actions commute, $rg = gr$. Moreover it also holds under the weaker requirement that for all $(g, r) \in G \times R$ there exist $(g', r') \in G \times R$ such that $rg = g'r'$. The latter property holds for example if we take G, R to be 2-D translations and rotations respectively. This is because the semidirect product of the abelian group $\mathbb{R}^2$ (which describes translations) and the group $SO(2)$ of orthogonal 22 matrices (which describes rotations and reflections that keep the origin fixed) is isomorphic to the Euclidean group. The group of translations is a normal subgroup of the Euclidean group and the Euclidean group is a semidirect product of the translation group and $SO(2)$.*

3.8 Factorization of transformations

The results above would be much more interesting if ν of equation 24 were "covariant" in the following sense: $\langle \mu(g_{y'} I), s^k \rangle) = \left\langle \mu(I), \mu(g_{y'}^{-1} \bar{t}^i) \right\rangle$ which we

rewrite for simplicity as

$$\langle \mu(rI), \mu(r\bar{t}) \rangle = \langle \mu(I), \mu(\bar{t}) \rangle \tag{27}$$

where r is the group element corresponding to $g_{y'}$. If this were true then one invariance could be learned via the templates t^k and another could be learned in the next layer via the templates s^i. This covariance property however cannot be expected for general η.

We first sketch the "linear" situation: η is the identity function. We also assume that the set of t^k is an orthonormal basis. In this case

$$\begin{aligned} \langle \mu(rI), \mu(r\bar{t}) \rangle &= \sum_k \int dg \left\langle grI, t^k \right\rangle \int dg' \left\langle g'r\bar{t}, t^k \right\rangle \\ &= \int dg \int dg' \sum_k \left\langle grI, t^k \right\rangle \left\langle g'r\bar{t}, t^k \right\rangle \end{aligned}$$

and, if the transformations g and r do commute we have

$$\begin{aligned} \langle \mu(rI), \mu(r\bar{t}) \rangle &= \int dg \int dg' \left\langle grI, g'r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'r^{-1}r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'\bar{t} \right\rangle \\ &= \langle \mu(I), \mu(\bar{t}) \rangle \end{aligned} \tag{28}$$

Note that if the transformations g and r do not commute the result in (28) does not hold. Even a weaker result, i.e., $\langle \mu(rI), \mu(r\bar{t}) \rangle = \langle \mu(I), \mu(r'\bar{t}) \rangle$ for some r', does not hold. In fact, using Remark 8 we have that for each fixed g, g', r there exists $\tilde{g}, \hat{g}, r'(g, g')$ such that

$$\langle grI, g'r\bar{t} \rangle = \langle \tilde{g}I, \hat{g}r'(g, g')\bar{t} \rangle .$$

However r' is different for any choice of g, g'. Therefore we can neither obtain the same result as in in the commutative case nor have a weaker form of it, i.e. $\langle \mu(rI), \mu(r\bar{t}) \rangle = \langle \mu(I), \mu(r'\bar{t}) \rangle$ for some r'. We now sketch another case of practical interest in which $\eta(x) = x^2$. We make the same simplifying assumptions. Then

$$\begin{aligned} \langle \mu(rI), \mu(r\bar{t}) \rangle &= \sum_k (\int dg \left\langle grI, t^k \right\rangle)^2 (\int dg' \left\langle g'r\bar{t}, t^k \right\rangle)^2 \\ &= \int dg \int dg' \sum_k (\left\langle grI, t^k \right\rangle)^2 (\left\langle g'r\bar{t}, t^k \right\rangle)^2. \end{aligned}$$

This leads to consider the related expression

$$\sum_k (P_k Rx)^p (P_k R^* y)^p$$

where we assume that R and R^* are in principle different matrices. Typically R represents one of the elements of the $SO(2)$ group (rotations). P_k is the projection vector s.t. $(P_k)^T x = (x)_k$. Notice that $\sum_k (P_k Rx)^T (P_k Ry) = \sum_k x^T R^T (P_k)^T (P_k) Ry = x^T R^T [\sum_k (P_k)^T (P_k)] Ry = x^T R^T Ry = x^T y$ since $\sum_k (P_k)^T (P_k) = I$. We consider the important case of $p = 2$. Then

$$\begin{aligned} \sum_k (P_k Rx)^2 (P_k R^* y)^2 &= x^T [\sum_k (R^T \bar{P}_k R) x y^T (R^{*T} \bar{P}_k R^*)] y \\ &= x^T D y = \sum_k x_k^2 y_k^2 r_k^2 r_k^{*2} \end{aligned}$$

since D is a $K \times K$ diagonal matrix with elements $D_{i,i} = x_k y_k {r_k}^2 {r_k^*}^2$ where $r_k = R_{k,k}$ and ${R^*}_{k,k} = r_k^*$. Thus *for the specific network* here, factorization holds only for functions η which are linear (or equivalently only for the first moment of the distribution) and only for commutative groups.

Thus *for the specific network* here, factorization holds only for functions η which are linear (or equivalently only for the first moment of the distribution) and only for commutative groups.

3.9 Factorization of Invariance ranges

Here we quickly review Theorem 7 in the setting considered in this section.

A more interesting architecture would compute invariance to translations at the level of the first layer over small ranges on a grid of μ units (with different x, y positions). At the next layer invariance to rotations and scales (and possibly translations) over larger ranges is computed by another μ layer consisting of a sparser grid of μ units. At the third level a still sparser grid of μ units could compute invariance to larger translations.

Here we show, as an example, how to built a sequence ("factorize") of μ units (signature) invariant to larger and larger ranges of $1D$ translations. For simplicity we assume we have only one template for each layer.

Theorem 9. *Factorization of invariant ranges. Let $I \in \mathcal{X}$ and $\mu_\ell : \mathcal{X} \to \mathcal{X}$ the μ_ℓ unit at layer l. Let $supp(t_\ell) \subset supp(\mu_\ell(I)) \subset P_\ell$, where respectively, P_ℓ is the pooling range and t_ℓ the template at layer ℓ. Then, for any ℓ*

$$\mu_\ell(T_\tau I) = \mu_\ell(I), \quad \forall I \in \mathcal{X},\ \tau \in [0, \bar{\tau}], \quad \exists\, \bar{\tau} \in \mathbb{R}$$

i.e. at each layer there is a range of transformations for which the μ unit is invariant and

$$\mu_{\ell+1}(T_\tau I) = \mu_{\ell+1}(I), \quad \forall I \in \mathcal{X},\ \tau \in [0, \bar{\tau}'], \quad \exists\, \bar{\tau}' > \bar{\tau}.$$

i.e. the invariance range is larger and larger with growing layer number.

Proof. We prove the statement for the fist two layers but the proof follows the same steps for any two subsequent layers. Following the definitions of par. 3.7

we call $\mu_1 \equiv \mu$, $\mu_2 \equiv \nu$ and $t_1 \equiv t$, $t_2 \equiv \bar{t}$.
For $\ell = 1$ being

$$\mu(I)(\tau) = \int_{\tau P_1} d\tilde{\tau}\ \eta(\langle T_{\tilde{\tau}} I, t\rangle) \tag{29}$$

it is clear that if $supp(t) \subset supp(I)$ there will be a $\bar{\tau} > 0$ such that $\langle T_\tau I, t\rangle = \langle I, T_{-\tau} t\rangle = \langle I, t\rangle$, $\forall\ \tau \in [0, \bar{\tau}]$ and consequently $\mu(T_\tau I) = \mu(I)$.
We prove now that the response at the second layer has a bigger range of invariance. By definition, for $\ell = 2$

$$\nu(I)(\tau) = \int_{\tau P_2} d\tilde{\tau}\ \eta(\langle T_{\tilde{\tau}} \mu(I), \bar{t}\rangle). \tag{30}$$

Remember that by invariance of the $\mu(I)$ unit we have $\nu(T_\tau I) = \nu(I),\ \forall \tau \in [0, \bar{\tau}]$. Note now that for any $\bar{\tau}' > \bar{\tau}$ we can write

$$\mu(T_{\bar{\tau}'} I) = \mu(T_{\bar{\tau}+\tau^*} I) = T_{-\tau^*} \mu(T_{\bar{\tau}} I) = T_{-\tau^*} \mu(I) \tag{31}$$

where $\tau^* = \bar{\tau}' - \bar{\tau}$. Proceeding as before, being $supp(\bar{t}) \subset supp(\mu(I))$, we have that there exists $\check{\tau}$ s.t. $\forall \tau \in [0, \check{\tau}]$, $\langle T_\tau \mu(I), \bar{t}\rangle = \langle \mu(I), T_{-\tau} \bar{t}\rangle = \langle \mu(I), \bar{t}\rangle$. We can then choose for an opportune $\bar{\tau}'$ such that $\tau^* = \check{\tau}$. This implies that the invariance range at the second layer is larger than at the first layer. The same argument provides a factorization of invariance to the whole range in layers of growing invariant ranges. □

3.10 Approximate Factorization for Invariances in Object Classes

The first version of [78] we conjectured that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g. the subgroup of translations of the affine group) and then adding the invariance w.r.t. another subgroup (e.g. rotations). It can be shown that in general this operation is impossible (see Theorem 4 and **[?]**). However, there are particular, interesting situations in which approximate factorization is possible. In the following we show that this is the case if the templates are in the same *nice object class*.
To fix ideas let us consider the example of a face rotating in depth. Assume that at some level of the hierarchy the output is a single vector $\mu(I)$ corresponding to an input image $I \in \mathcal{X}$; for example assume that $\mu(I)$ is invariant w.r.t. translation of I in $, x, y$. Now consider transformations (e.g. rotations in depth) of one template t (e.g. another face), that we indicate as $T^j t = t^j$, with $j = 1, \cdots, N$ denoting different rotations. If $\mu(I)$ and the set of templates t^j belongs to the same object class (e.g. of faces) and there are enough templates, then it is reasonable to assume that for any specific transformation R of the image, there exists j s.t. $\left\|\mu(RI) - \mu(t^j)\right\|_k < \delta$. This then implies that the set of neurons each tuned to each template t^k can represent an approximation

of $P_t(\langle \mu(I), t\rangle)$ and that the probability is approximatively invariant to transformations R of I, that is $P_t(\langle \mu(I), t\rangle) \approx P_t(\langle \mu(RI), t\rangle)$ (where the notation P_t indicates the distribution over the set of templates).

3.11 A mathematical summary (*incomplete*)

The theory just described has a simple mathematical structure, aside from the biological details. We summarize it in this appendix.

1. Setting

Let $(\mathcal{X}, \langle \cdot, \cdot \rangle)$ be a real separable Hilbert space, e.g. $\mathcal{X} = L^2(\mathbb{R}^2)$. Let $\mathcal{L}(\mathcal{X})$ be the space of linear operators to and from $\mathcal{X}$.

A measurement is defined as a functional $m : \mathcal{X} \to \mathbb{R}$. A signature is a map $\phi : \mathcal{X} \to \ell^2$ and can be viewed as a collection of measurements.

2. Linear measurements: bases, frames and Johnson Lindenstrauss lemma

Claim: Linear measurements give rise to isometric or quasi-isometric signatures.

Let $\mathcal{T} \subset \mathcal{X}$ be countable and

$$s : \mathcal{X} \to \ell^2, \quad s_t(I) = \langle I, t\rangle \quad t \in \mathcal{T}.$$

If $\mathcal{T}$ is an orthonormal basis $I = \sum_{t\in\mathcal{T}} s_t(I)t$ and $\|s(I)\|_2 = \|I\|$ where $\|s(I)\|_2^2 = \sum_t s_t(I)^2$.
If $\mathcal{T}$ is a frame, by definition, $A\|I\| \le \|s(I)\|_2 \le B\|I\|$, with $0 < A \le B < \infty$.
Finally, if $\mathcal{X}$ is a set of n points in $\mathbb{R}^N$ and $\mathcal{T}$ a suitable finite set of p, possibly random, vectors. by the Jonson and Lindenstrauss lemma $(1-\epsilon)\|I\| \le \|s(I)\|_2 \le (1+\epsilon)\|I\|$, as soon as $p \ge 8\log n/\epsilon^2$.

3. Invariant measurements via group integration

Claim: Invariant measurements can be obtained via local group integration.

Let $\mathcal{G}$ be a locally compact Abelian group and dg the associated Haar measure. Let $T : \mathcal{G} \to \mathcal{B}(\mathcal{X})$, $T_g = T(g)$ be a representation of $\mathcal{G}$ on $\mathcal{X}$. Let $m, h : \mathcal{X} \to \mathbb{R}$ with $m(I) = \int h(T_g I)dg$. Then m is *invariant*:

$$m(T_{g'}I) = \int h(T_{g'}T_g I)dg = \int h(T_{g'g}I)dg = m(I),$$

for all $I \in \mathcal{X}, g' \in \mathcal{G}$.

Example 1. *Let $\mathcal{X} = L^2(\mathbb{R}^2)$, and $(T_g I)(I) = I(\sigma_g^{-1}(I))$, where $\sigma_g : \mathbb{R}^2 \to \mathbb{R}^2$, with $g \in \mathcal{G}$, is a representation of a group $\mathcal{G}$. In particular we can consider $\mathcal{G}$ to be the affine group so that $\sigma_g r = Ar + b$ and $\sigma_g^{-1} r = A^{-1}r - b$, where $b \in \mathbb{R}^2$ and $A : \mathbb{R}^2 \to \mathbb{R}^2$ is a unitary matrix. It is easy to see that in this case T_g is linear and $T_g^* x(r) = x(\sigma_g r)$ for all $g \in \mathcal{G}$ and $r \in \mathbb{R}^2$. Moreover,redefining the representation dividing by the transformation Jacobian we have $T_g^* T_g = I$ so that $g \mapsto T_g$ is a unitary representation of $\mathcal{G}$ on $\mathcal{X}$.*

Observation If $t, I \in \mathcal{X}$, and T is a unitary representation, then

$$\langle T_g I, t\rangle = \left\langle I, T_g^* t\right\rangle = \left\langle I, T_g^{-1} t\right\rangle = \left\langle I, T_{g^{-1}} t\right\rangle.$$

For $\sigma : \mathbb{R} \to \mathbb{R}$ measurable, let

$$c : \mathcal{X} \to \mathbb{R}, \qquad c(I) = \int \sigma(\langle I, T_g t\rangle) dg,$$

then c is invariant.

4. Approximately invariant measurements via local group integration

Claim: Approximately invariant measurements can be obtained via local group integration.

Lemma 3. *If $\mathcal{G}_{0,I} = \{g' \in \mathcal{G} \mid h(T_g I) = 0, \forall g \in \mathcal{G}_0 \Delta g'^{-1} \mathcal{G}_0\}$, then,*

$$m_0(I) = m_0(T_{g'} I), \quad \forall g' \in \mathcal{G}_{0,I}.$$

Proof. Let $\mathcal{G}_0 \subset \mathcal{G}$ and $m_0, h : \mathcal{X} \to \mathbb{R}$ with $m_0(I) = \int_{\mathcal{G}_0} h(T_g I) dg$.
Clearly, in this case m_0 is not invariant, but

$$\begin{aligned} m_0(I) - m_0(T_{g'} I) &= \int_{\mathcal{G}_0} h(T_g I) dg - \int_{\mathcal{G}_0} h(T_{g' g} I) \\ &= \int_{\mathcal{G}_0} h(T_g I) dg - \int_{g'^{-1} \mathcal{G}_0} h(T_g I) dg = \int_{\mathcal{G}_0 \Delta g'^{-1} \mathcal{G}_0} h(T_g I) dg. \end{aligned}$$

The statement is then clear. □

Example 2. *The interpretation of $\mathcal{G}_{0,I}$ can be made clear considering $\mathcal{X} = L^2(\mathbb{R})$ and $h(I) = |f(x)|^2$, $I \in \mathcal{X}$. Let $(T_\tau I)(x) = I(x + \tau)$, $I \in \mathcal{X}$, $\tau \in \mathbb{R}$ and $\mathcal{G}_0 = [-\pi, \pi]$. In this case, $g'^{-1} \mathcal{G}_0 = [-\pi - \tau, \pi - \tau]$.*

5. Signature of approximately invariant measurements

Claim: A signature consisting of a collection of measurements obtained via partial integration is covariant.

6. Discrimination properties of invariant and approximately invariant signatures

Claim: If the considered group is compact, then it is possible to built (possibly countably many) nonlinear measurements that can *discriminate* signals which do not belong to the same orbit.

7. Hierarchies approximately invariant measurements

Claim: An appropriate cascade of linear measurements and approximately invariant measurements (obtained via partial integration) give rise to signatures which are covariant and eventually invariant.

8. Whole vs parts and memory based retrieval Biological Conjecture: Signatures obtained from complex cells at each level access an (associative) memory which also is involved in top-down control.

4 Part II: Learning Transformations and Spectral Properties

Summary of Part II. *Part I proves that pooling over sequences of transformed images stored during development allows the computation at run-time of invariant signatures for any image. Part II assumes that storage of sequences of images is performed online via Hebbian synapses. Because of this assumption, it is possible then to connect invariances to tuning of cortical cells. We start by relating the size of the receptive field – called aperture – and transformations "seen through the aperture". During development, translations are effectively the only learnable transformations by small apertures – eg small receptive fields – in the first layer (see also appendix 14). We then introduce a Linking Conjecture: instead of explicitly storing a sequence of frames during development as assumed in the abstract framework of Part I, it is biologically more plausible to assume that there is Hebbian-like learning at the synapses in visual cortex. We will show that, as a consequence, the cells will effectively store and compress input "frames" by computing online the eigenvectors of their covariance during development and storing them in their synaptic weights. Thus the tuning of each cell is predicted to converge to one of the eigenvectors. Furthermore, invariance is now obtained by pooling nonlinear functions, such as the modulo, of the dot products between the eigenfunctions (computed over the transformation of interest) and the new image. The section further shows that numerical simulations predict well quantitative properties of the tuning of simple cells in V1 across different species. Further, at higher cortical levels, similar developmental learning on the V1 representation generates 4-dimensional wavelets. The prediction seems qualitatively consistent with physiology data.*

4.1 Apertures and Stratification

Summary. In this short section we argue that size and position invariance develop in a sequential order, meaning that smaller transformations are invariant before larger ones; size and position invariance are computed in stages by a hierarchical system that builds invariance in a feedforward manner. The transformations of interest include all object transformations which are part of our visual experience. They include perspective projections of (rigid) objects moving in 3D (thus transforming under the action of the euclidean group). They also include *nonrigid* transformations (think of changes of expression of a face or pose of a body): the memory-based architecture described in part I can deal – exactly or approximately – with all these transformations.

Remember that the hierarchical architecture has layers with receptive fields of increasing size. The intuition is that transformations represented at each level of the hierarchy will begin with "small" affine transformations – that is over a small range of translation, scale and rotation. The "size" of the transformations represented in the set of transformed templates will increase with the level of the hierarchy and the size of the apertures. In addition it seems intuitive that only translations will be "seen" by small apertures with scale and

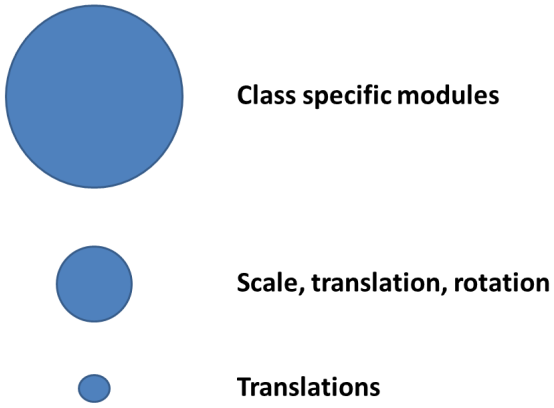


Figure 12: *The conjecture is that receptive field sizes affects not only the size but also the type of transformations that is learned and represented by the templates. In particular, small apertures (such as in V1) only "see" (small) translations.*

orientation changes been relevant later in the hierarchy.

Let us be more specific. Suppose that the first layer consists of an array of "small apertures" – in fact corresponding to the receptive fields of V1 cells – and focus on one of the apertures. We will show that the only transformations that can be "seen" by a small aperture are small translations, even if the transformation of the image is more complex.

4.1.1 Translation approximation for small apertures

The purpose of this section is to show that a twice differentiable flow, when perceived for a sufficiently short time through a sufficiently small aperture, is well approximated by a translation. Other two derivations of the same result can be found in the Appendix 14.2.

Let $I \subseteq \mathbb{R}$ be a bounded interval and $\Omega \subseteq \mathbb{R}^N$ an open set and let $\Phi = (\Phi_1, ..., \Phi_N) : I \times \Omega \to \mathbb{R}^N$ be $\mathcal{C}_2$, where $\Phi(0, .)$ is the identity map. Here $\mathbb{R}^N$ is assumed to model the image plane, intuitively we should take $N = 2$, but general values of N allow our result to apply in subsequent, more complex processing stages, for example continuous wavelet expansions, where the image is also parameterized in scale and orientation, in which case we should take $N = 4$. We write (t, x) for points in $I \times \Omega$, and interpret $\Phi(t, x)$ as the position in the image at time t of an observed surface feature which is mapped to $x = \Phi(0, x)$ at time zero. The map Φ results from the (not necessarily rigid) motions of the observed object, the motions of the observer and the properties of the imaging apparatus. The implicit assumption here is that no surface features which are visible in Ω at time zero are lost within the time interval I. The

assumption that Φ is twice differentiable reflects assumed smoothness properties of the surface manifold, the fact that object and observer are assumed massive, and corresponding smoothness properties of the imaging apparatus, including eventual processing.

Now consider a closed ball $B \subset \Omega$ of radius $\delta > 0$ which models the aperture of observation. We may assume B to be centered at zero, and we may equally take the time of observation to be $t_0 = 0 \in I$. Let

$$K_t = \sup_{(t,x)\in I\times B} \left\| \frac{\partial^2}{\partial t^2}\Phi\left(t,x\right)\right\|_{\mathbb{R}^N}, \quad K_x = \sup_{x\in B} \left\| \frac{\partial^2}{\partial x \partial t}\Phi\left(0,x\right)\right\|_{\mathbb{R}^{N\times N}}.$$

Here $(\partial/\partial x)$ is the spatial gradient in $\mathbb{R}^M$, so that the last expression is spelled out as

$$K_x = \sup_{x\in B}\left(\sum_{l=1}^N \sum_{i=1}^N \left(\frac{\partial^2}{\partial x_i \partial t}\Phi_l\left(0,x\right)\right)^2\right)^{1/2}.$$

Of course, by compactness of $I \times B$ and the $\mathcal{C}_2$-assumption, both K_t and K_x are finite.

Theorem 10. *(Poggio-Maurer) There exists $V \in \mathbb{R}^N$ such that for all $(t,x) \in I \times B$*

$$\left\|\Phi\left(t,x\right) - \left[x + tV\right]\right\|_{\mathbb{R}^N} \leq K_x \delta \left|t\right| + K_t \frac{t^2}{2}.$$

As one might suspect, the proof reveals this to be just a special case of Taylor's theorem.

Proof. Denote $V\left(t,x\right) = \left(V_1, ..., V_l\right)\left(t,x\right) = \left(\partial/\partial t\right)\Phi\left(t,x\right)$, $\dot{V}\left(t,x\right) = \left(\dot{V}_1, ..., \dot{V}_l\right)\left(t,x\right) = \left(\partial^2/\partial t^2\right)\Phi\left(t,x\right)$, and set $V := V\left(0,0\right)$. For $s \in [0,1]$ we have with Cauchy-Schwartz

$$\left\|\frac{d}{ds}V\left(0,sx\right)\right\|_{\mathbb{R}^N}^2 = \sum_{l=1}^N\sum_{i=1}^N\left(\left(\frac{\partial^2}{\partial x_i \partial t}\Phi_l\left(0,sx\right)\right)x_i\right)^2 \leq K_x^2\left\|x\right\|^2 \leq K_x^2\delta^2,$$

whence

$$\begin{aligned}
& \left\|\Phi\left(t,x\right) - \left[x+tV\right]\right\| \\
= & \left\|\int_0^t V\left(s,x\right)ds - tV\left(0,0\right)\right\| \\
= & \left\|\int_0^t\left[\int_0^s \dot{V}\left(r,x\right)dr + V\left(0,x\right)\right]ds - tV\left(0,0\right)\right\| \\
= & \left\|\int_0^t\int_0^s \frac{\partial^2}{\partial t^2}\Phi\left(r,x\right)drds + t\int_0^1 \frac{d}{ds}V\left(0,sx\right)ds\right\| \\
\leq & \int_0^t\int_0^s\left\|\frac{\partial^2}{\partial t^2}\Phi\left(r,x\right)\right\|drds + \left|t\right|\int_0^1\left\|\frac{d}{ds}V\left(0,sx\right)\right\|ds \\
\leq & K_t\frac{t^2}{2} + K_x\left|t\right|\delta.
\end{aligned}$$

□

Of course we are more interested in the visible features themselves, than in the underlying point transformation. If $f : \mathbb{R}^N \to \mathbb{R}$ represents these features, for example as a spatial distribution of gray values observed at time $t = 0$, then we would like to estimate the evolved image $f(\Phi(t, x))$ by a translate $f(x + tV)$ of the original f. It is clear that this is possible only under some regularity assumption on f. The simplest one is that f is globally Lipschitz. We immediately obtain the following

Corollary 1. *Under the above assumptions suppose that $f : \mathbb{R}^N \to \mathbb{R}$ satisfies*

$$|f(x) - f(y)| \leq c \|x - y\|$$

for some $c > 0$ and all $x, y \in \mathbb{R}^N$. Then there exists $V \in \mathbb{R}^N$ such that for all $(t, x) \in I \times B$

$$|f(\Phi(t, x)) - f(x + tV)| \leq c \left(K_x |t| \delta + K_t \frac{t^2}{2} \right).$$

An example As a simple example we take rigid rotation with angular velocity ω about a point v in the image plane, observed in a neighborhood of radius δ about the origin. Then

$$\Phi(t, x_1, x_2) = \begin{pmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{pmatrix} \begin{pmatrix} x_1 - v_1 \\ x_2 - v_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

and with some calculation we obtain the bounds $K_t \leq \omega^2 (\|v\| + \delta)$ and $K_x \leq \sqrt{2} |\omega|$. The error bound in the theorem then becomes

$$(\|v\| + \delta) \omega^2 t^2 / 2 + \sqrt{2} |\omega| t\delta.$$

If we take $v = 0$, so that the center of rotation is observed , we see that we considerably overestimate the true error for large t, but for $t \to 0$ we also see that we have the right order in δ and that the constant is correct up to $\sqrt{2}$.
A one-layer system comprising the full image (a large aperture) would require a memory-based module to store all the transformations induced by all elements g of the full group of transformations at all ranges. Because this should include all possible local transformations as well (for instance for an object which is a small part of an image), this quickly becomes computationally infeasible as a general solution. A hierarchical architecture dealing with small, local transformations first – which can be assumed to be affine (because of Lemma 12) – can solve this problem and may have been evolution's solution for the vertebrate visual system. It is natural that layers with apertures of increasing size learn and discount transformations – in a sequence, from local transformations to more global ones. The learning of transformations during development in a sequence of layers with increasing range of invariance corresponds to the term *stratification*.

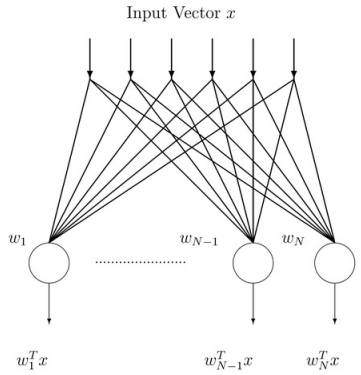
4.2 Linking conjecture: developmental memory is Hebbian

Summary. Here we introduce the hypothesis that memory of transformations during development is Hebbian. Thus instead of storing a sequence of frame of a template transforming, synapses store online updates due to the same sequences.

We introduce here a biologically motivated *Linking Conjecture*: instead of explicitly storing a sequence of frames during development as assumed in Part I, we assume that there is Hebbian-like learning at the synapses in visual cortex. The conjecture consists of the following points:

Linking Conjecture

- The memory in a layer of cells (such as simple cells in V1) is stored in the weights of the connections between the neurons and the inputs (from the previous layers).
- Instead of storing a sequence of discrete frames as assumed in Part I, online learning is more likely, with synaptic weights being incrementally modified.
- Hebbian-like synapses exist in visual cortex.
- Hebbian-like learning is equivalent to an online algorithm computing PCAs.
- As a consequence, the tuning of simple cortical cells is dictated by the top PCAs of the templatebook, since Hebbian-like learning such as the Oja flow converges to the top PCA.



4.2.1 Hebbian synapses and Oja flow

The algorithm outlined in part I in which transformations are "learned" by memorizing sequences of a patch undergoing a transformation is an algorithm similar to the existing HMAX (in which S2 tunings are learned by sampling and

memorizing random patches of images and invariance is hardwired). A biologically more plausible online learning rule is somewhat different: synapses would change as an effect of the inputs, effectively compressing information contained in the templates and possibly making signatures more robust to noise. Plausible online learning rules for this goal are associative Hebbian-like rules. As we will see later, Hebbian-like synaptic rules are expected to lead to tuning of the simple cells according to the eigenvectors of the templatebooks.

We discuss here the specific case of the Oja's flow. Oja's rule [69, 42] defines the change in presynaptic weights $\mathbf{w}$ given the output response $\mathbf{y}$ of a neuron to its inputs to be

$$\Delta \mathbf{w} \;=\; \mathbf{w}_{n+1} - \mathbf{w}_n \;=\; \eta\, y_n(\mathbf{x}_n - y_n \mathbf{w}_n) \tag{32}$$

where is the "learning rate" and $y = \mathbf{w}^{\mathbf{T}}\mathbf{x}$. The equation follows from expanding to the first order Hebb rule normalized to avoid divergence of the weights. Its continuous equivalent is

$$\dot{\mathbf{w}} \;=\; \gamma\, y(\mathbf{x} - y\mathbf{w}) \tag{33}$$

Hebb's original rule, which states in conceptual terms that "neurons that fire together, wire together", is written as $\Delta \mathbf{w} \;=\; \eta\, y(\mathbf{x}_n)\mathbf{x}_n$, yielding synaptic weights that approach infinity with a positive learning rate. In order for this algorithm to actually work, the weights have to be normalized so that each weight's magnitude is restricted between 0, corresponding to no weight, and 1, corresponding to being the only input neuron with any weight. Mathematically, this requires a modified Hebbian rule:

$$w_i(n+1) \;=\; \frac{w_i + \eta\, y(\mathbf{x}) x_i}{\left(\sum_{j=1}^{m} [w_j + \eta\, y(\mathbf{x}) x_j]^p\right)^{1/p}} \tag{34}$$

of which Oja's rule is an approximation.

Several theoretical papers on Hebbian learning rules show that selective changes in synaptic weights are difficult to achieve without building in some homeostatic or normalizing mechanism to regulate total synaptic strength or excitability. In the meantime, homeostatic control of synaptic plasticity – corresponding to the normalizing term in Oja equation – ([102]) is in fact experimentally well established.

The above learning rules converge to the PCA with the largest eigenvalue (see Appendix 20). It is a key conjecture of Part II of this paper that Oja's flow or some variation of it (with appropriate circuitry), may link the spectral properties of the templatebook to receptive field tuning in visual areas. The conjecture is based on Oja's and other results, summarized by:

Proposition 5. *The Oja flow (Equation 32) generates synaptic weights that converge to the top real eigenvector of the input patterns covariance matrix, that is the covariance matrix of the templatebook (in the noiseless case).*

In principle, local invariance to translation can be achieved by averaging a function over a number of Principal Components for each aperture (ideally all, in practice a small number) corresponding to the "movie" of one transformation sequence. The PCA do in fact span the variability due to the transformation (translation in the case of simple cells): thus this average is equivalent to averaging over frames of the templatebook, as described in Part I. An empirical observation is that most of the PCA for the translation case appear as quadrature pairs (this is also true for the other subgroups of the affine group since the characters are always Fourier components). It follows that the *energy* aggregation function is *locally* invariant (because $|e^{i\omega nx+\theta}| = 1$) to the transformation (see Figure 21).

In the hypothesis of Oja-type online learning, one possible scenario is that that different simple cells which "look" at the same aperture converge to a single top principal component. Several Oja-like learning rules converge to principal components [88, 70]. In the presence of lateral inhibition, different cells with the same aperture may converge to different eigenvectors with the same eigenvalue (such as the odd and even component of a quadrature pair (see Figure 21). A complex cell then aggregates the square or the modulo of two or more simple cells corresponding to different PCAs. Though diversity in the PCAs to fit the observed RF of simple cells may come from online learning in the presence of various types of noise, it is much more likely that there is lateral inhibition between nearby simple cells to avoid that they converge to eigenvectors of the same order (nearby neurons may also be driven by local interaction to converge to Gabor-like functions with similar orientation). In addition, Foldiak-type learning mechanisms (see Appendix 20.2) maybe responsible for wiring simple cells with the "same" orientation to the same complex cell.

It has been customary (for instance see [48] to state a single "slowness" maximization principle, formulated in such a way to imply both Oja's-like learning at the level of simple cells and wiring of the complex cells according to a Foldiak-like rule. Though such a principle does not seem to reflect any obvious biological plasticity property, it cannot be excluded that a single biological mechanisms – as opposed to a single abstract optimization principle – determines both the tuning of the simple cells and their pooling into complex cells. In a similar spirit, simple cells may be a group of inputs on a dendritic branch of a complex cell.

Notice that a relatively small change in the Oja equation gives an online algorithm for computing ICAs instead of PCAs [38]. Which kind of plasticity is closer to the biology remains an open question. We expect ICAs to be similar to PCAs described here but not identical. Our spectral analysis would not carry over to ICAs – at least not exactly – and instead direct simulations of the dynamic online equations will have to be done.

Let us summarize the main implications of this section in terms of templates, signatures and simple and complex cells. Notice that the templatebook $\mathbb{T}$ is a tensor with $\tau_{i,j}$ being an array. There are D PCA components for each $\mathbb{T}$: for instance retaining the first two PCA components shown in Figure 21

corresponds to replacing $\mathbb{T}$ with $\hat{\mathbb{T}}$ with 2 rows. From this point of view, what do we expect it will happen during developmental learning using a Hebb-like rule? Repeated exposure to stimuli sequences corresponding to the rows of the $\mathbb{T}$ should induce, through the learning rule, simple cell tunings corresponding for instance to the two PCA in quadrature pair of Figure 21. Simple cells tuned to these Principal Components would be pooled by the same complex cell.

4.3 Spectral properties of the templatebook covariance operator: cortical equation

Summary. This section focuses on characterizing the spectral properties associated with the covariance of the templatebook. It proposes a "cortical equation" whose solution provides the eigenfunctions of the covariance (the exact solution in some particular cases can be found in 15 and more material on the templatebook spectral properties can be found in 16). Hebbian synaptic rules imply that during development the tuning of simple cells when exposed to inputs from the retina will converge to the top eigenfunction(s). We start with the $1D$ analysis; the 2D problem is somewhat more interesting because of the "symmetry breaking" induced by motion.

We consider a layer of $2D$ "apertures" and the covariance of the templatebook associated with each aperture resulting from transformations of images "seen" through one of these apertures. This will lead later to an explicit solution for the first layer in the case of translations.

For any fixed t we want to solve the spectral problem associated to the templatebook:

$$\mathbb{T}_t = (g_0 t, g_1 t, ..., g_{|G|} t, ...)^T$$

i.e. we want to find the eigenvalues λ_i and eigenfunctions ψ_i such that

$$\mathbb{T}_t^* \mathbb{T}_t \psi_i = \lambda_i \psi_i, \ i = 1, ..., N \tag{35}$$

To state the problem precisely we need some definitions. We start first with the $1D$ problem for simplicity

We show how to derive an analytical expression of the visual cortex cells tuning based on the following hypothesis:

1. **Observables**: images, transforming by a locally compact group, looked through an "aperture" better specified later.
2. **Hebbian learning**: hebbian like synapses exists in visual cortex.

We fix few objects:

- $\mathcal{X}$ space of signals: $L^2(\mathbb{C}, dx)$.
- $\mathcal{T} \subseteq \mathcal{X}$ the template set.

We will solve the eigenproblem associated to the continuous version of (35): in this case the basic observable given by the operator $T : \mathcal{X} \to \mathcal{X}$

$$(TI)(y) \equiv [t * M_a I](y) = \int dx t(y - x) a(x) I(x), \ \ t \in \mathcal{T}, \ a, I \in \mathcal{X} \tag{36}$$

where

$$(M_a I)(x) = a(x)I(x),\ a \in \mathcal{X}$$

The equation (36) is the mathematical expression of the observable T i.e. a translating template t looked through the function a which will be called the aperture.

Remark 9. *T is linear (from the properties of the convolution operator) and bounded (from $\|T\| = \|\mathfrak{F}(T)\| = \|t\|\,\|a\|$).*

Remark 10. *M_a is a selfadjoint operator.*

The adjoint operator $T^* : \mathcal{X} \to \mathcal{X}$ is given by

$$\begin{aligned}\langle TI, I' \rangle &= \int dy\ \bar{I}'(y) \int dx\ t(y-x)a(x)I(x) \\ &= \int dx\ I(x)a(x) \int dy\ t(y-x)\bar{I}'(y) = \langle I, T^* I' \rangle\end{aligned}$$

which implies $T^* I = M_a(t^- * I),\ t^-(x) = t(-x)$. Note that $\|t\| = \|t^-\| \Rightarrow \|T\| = \|T^*\|$, i.e. $\|T^*\|$ is bounded.
Assuming Hebbian learning we have that the tuning of the cortical cells is given by the solution of the spectral problem of the covariance operator associated to T, $T^*T : \mathcal{X} \to \mathcal{X}$

$$\begin{aligned}[T^*TI](y) &= M_a[t^- * (t * (M_a I))](y) = M_a[(t^- * t) * (M_a I)](y) \\ &= M_a(t^{\circledast} * (M_a I)) = a(y) \int dx\ a(x)t^{\circledast}(y-x)I(x)\end{aligned}$$

The above expression can be written as

$$[T^*TI](y) = \int dx\ K(x,y)I(x), \quad K(x,y) = a(x)a(y)t^{\circledast}(y-x).$$

Being the kernel K Hilbert-Schmidt, i.e.

$$Tr(K) = \int dx\ K(x,x) = \int dx\ a^2(x)t^{\circledast}(0) < \infty$$

we have:

- the eigenfunctions corresponding to distinct eigenvalues are orthogonal.
- the eigenvalues are real and positive.
- there is at least one eigenvalues and one eigenfunctions (when K is almost everywhere nonzero) and in general a countable set of eigenfunctions.

In the following paragraphs we aim to find $\psi_n \in \mathcal{X}$ and $\lambda_n \in \mathbb{R}$ such that

$$a(y) \int dx\ a(x) t^{\circledast}(y-x)\psi_n(x) = \lambda_n \psi_n(y) \tag{37}$$

and study their properties. In particular in the next paragraphs we are going to find approximate solutions and show that they are a Gabor-like wavelets. An exact solution in some particular cases can be found in the appendix 15.

Remarks

- *Square aperture, circulants and Fourier*
  We start from the simplest discrete "toy" case in which we assume periodic boundary conditions on each aperture (one in a layer of receptive fields) resulting on a circulant structure of the templatebook.

  Define as templatebook T the circulant matrix where each column represents a template t shifted relative to the previous column. This corresponds to assuming that the visual world translates and is "seen through a square aperture" with periodic boundary conditions. Let us assume in this example that the image is one dimensional. Thus the image seen through an aperture

  $$a(x) \text{ s.t. } a(x) = 1 \text{ for } 0 \leq x \leq A \text{ and } a(x) = 0 \text{ otherwise}$$

  is $t(x-y)a(x)$ when the image is shifted by y. We are led to the following problem: find the eigenvectors of the symmetric matrix $T^T T$ where T is a circulant matrix[4]. If we consider the continuous version of the problem, that is the eigenvalue problem

  $$\int_0^A dx \psi_n(x) t^{\circledast}(y-x)dx = \lambda_n \psi_n(y)$$

  with $t^{\circledast}(x)$ being the autocorrelation function associated with t. The solution is $\psi_n(x) = e^{-i2\pi \frac{n}{A} x}$ which is the Fourier basis between 0 and A.

- *Translation invariance of the correlation function of natural images*

  In the toy example above the two point correlation function $t(x, y)$ has the form $t(x, y) = t(x - y)$ because of shifting the vector t. In the case of natural images, the expected two-point correlation function is always translation invariant even if the images are sampled randomly [87] (instead of being successive frames of a movie). In 1-D there is therefore no difference between the continuous motion case of one image translating and random sampling of different natural images (apart signal to noise issues). As we will see later, sampling from smooth translation is however needed for symmetry breaking of the 2D eigenvalue problem – and thus convergence of the eigenfunctions to directions orthogonal to the direction of motion.

[4]This problem has also been considered in recent work from Andrew Ng's group [89].

- *The sum of Gaussian receptive fields is constant if their density is large enough*
  What is $\sum G(x-\xi_i)$? If $\sum G(x-\xi_i) \approx \int G(x-\xi)d\xi$ then we know that $\int G(x-\xi)d\xi = 1$ for normalized G and for $-\infty \leq x \leq \infty$.

4.3.1 Eigenvectors of the covariance of the template book for the translation group

As we mentioned, the linking conjecture connect the spectral properties to the tuning of the cells during development. We study here the spectral properties of the templatebook.
We consider a biologically realistic situation consisting of a layer of Gaussian "apertures". We characterize the spectral properties of the templatebook associated with each aperture (corresponding to the receptive field of a "neuron") resulting from translations of images "seen" through one of these Gaussian apertures. For the neuroscientist we are thinking about *a Gaussian distribution (wrt to image space) of synapses on the dendritic tree of a cortical cell in $V1$ that will develop into a simple cells*.

Thus the image seen through a Gaussian aperture is $t(x-s)g(x)$ when the image is shifted by s. In the discrete case we are led to the following (PCA) problem: find the eigenvectors of the symmetric matrix $T^T G^T G T$ where G is a diagonal matrix with the values of a Gaussian along the diagonal.

In the following we start with the continuous $1D$ version of the problem.

The $2D$ version of equation (37) (see remark 12) is an equation describing the development of simple cells in $V1$; we call it "cortical equation" because, as we will see later, according to the theory it describes development of other cortical layers as well.

Notice that equation (37)

$$\int dx g(y)g(x)\psi_n(x)t^{\circledast}(y-x) = \lambda_n \psi_n(y)$$

holds for all apertures defined by functions $g(x)$.

Remark 11. *Eq.* (37) *can be easily written in the case $x \in \mathcal{Y} = L^2(G, dg)$ being G a locally compact group*

$$[T^*TI](g') = \int dg\ K(g,g')I(g), \quad K(g,g') = a(g)a(g')t^{\circledast}(g^{-1}g'), \ \ I \in \mathcal{Y},\ g,g' \in G.$$

The convolution is now on the group G.

Remark 12. *In $2D$ the spectral problem is:*

$$\int d\xi d\eta g(x,y)g(\xi,\eta)t^{\circledast}(\xi-x,\eta-y)\psi_n(\xi,\eta) = \lambda_n \psi_n(x,y). \tag{38}$$

where $t^{\circledast} \equiv t \circledast t$.

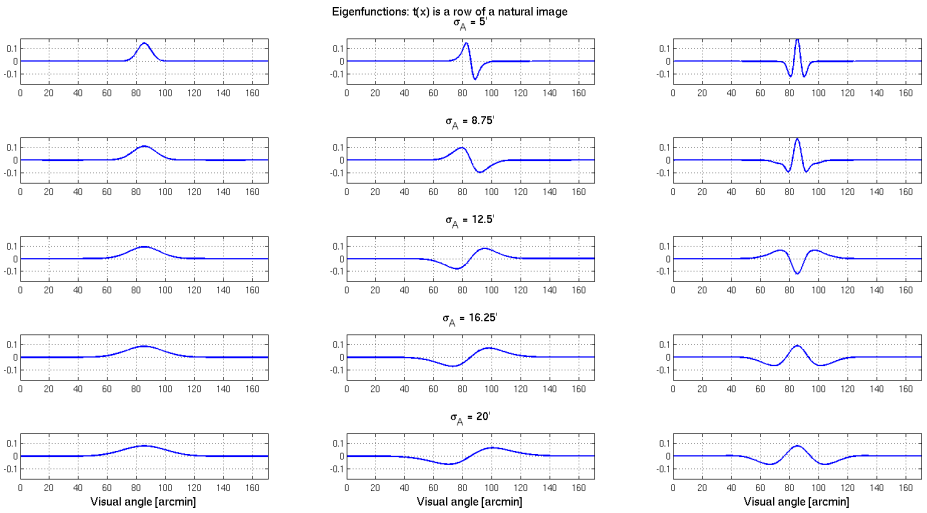


Figure 13: *Continuous spectrum of the covariance of the templatebook: Gabor-like eigenfunctions for different σ*

Numerical simulations in 1D show Gabor-like wavelets (see Figure 13) as eigenfunctions. This result is robust relative to the exact form of the correlation $t^{\circledast}(x)$. Other properties depend on the form of the spectrum (the Fourier transform of $t^{\circledast}(x)$). All the 1D simulations have been made (without any retinal processing) directly with natural images – which roughly have $t^{\circledast}(\omega) \propto \frac{1}{\omega^2}$.

In particular, the figures 14, 15 show that (in 1D) the eigenfunctions of the cortical equation *show the key signature of true gabor wavelets* in which the frequency is proportional to the σ. Figure 16 shows that the Gaussian envelope is smaller than the Gaussian aperture.

The following analysis of the eigenvalue equation provides some intuition behind the results of the numerical simulations.

1D: $t^{\circledast}(\omega_x)$ approximately piecewise constant

We represent the template as:

$$t^{\circledast}(x) = \frac{1}{\sqrt{2\pi}} \int d\omega \; t^{\circledast}(\omega) e^{i\omega x} \tag{39}$$

Let $\alpha = 1/\sigma_x^2$, $\beta = 1/\sigma_\psi^2$, and assume that the eigenfunction has the form $\psi_n(x) = e^{-\frac{\beta}{2}x^2} e^{i\omega_n x}$, where β and ω_n are parameters to be found.
With this assumptions eq. (37) reads:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\alpha}{2}y^2} \int dx \; e^{-\frac{x^2(\alpha+\beta)}{2}} \int d\omega \; t^{\circledast}(\omega) e^{i\omega(y-x)} e^{iw_n x} = \lambda(\omega_n) e^{-\frac{\beta y^2}{2}} e^{i\omega_n y}. \tag{40}$$

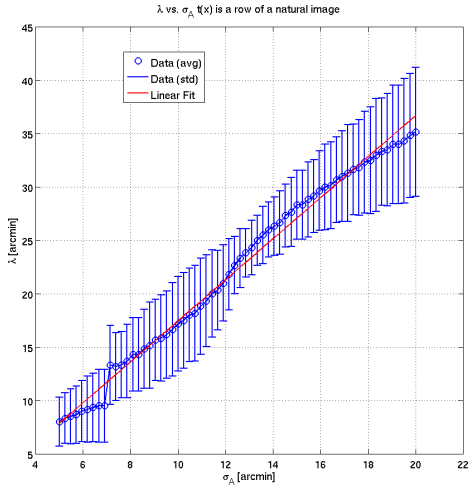


Figure 14: *Continuous spectrum: λ vs. σ_α for even symmetric patterns. The slope in this figure is k where $\lambda = k\sigma_\alpha$; $k \sim 2$ in this figure.*

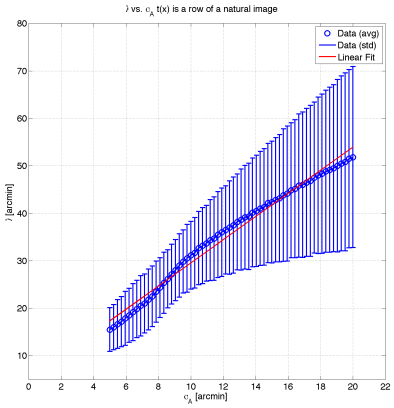


Figure 15: *Continuous spectrum: λ vs. σ_α for odd symmetric patterns. The slope is ~ 2.4. In 1D, odd symmetric eigenfunctions tend to have a lower modulating frequency.*

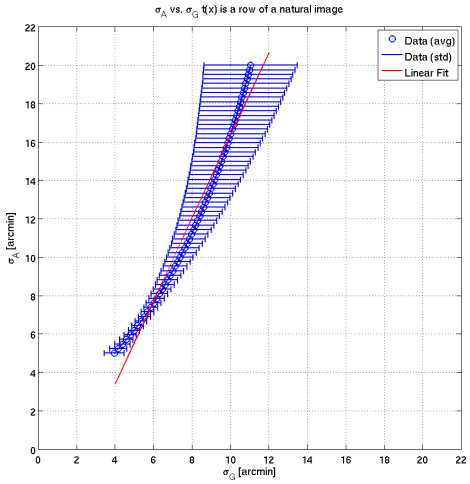


Figure 16: *Continuous spectrum: σ_α vs. σ_β. The slope is ~ 2. Though 1D, this is consistent with experimental data from [41] and [86] shown in fig. 22 where the slope is also roughly 2.*

Collecting the terms in x and integrating we have that the l.h.s becomes:

$$\sqrt{\frac{1}{\alpha+\beta}} e^{-\frac{\alpha}{2}y^2} \int d\omega \; t^{\circledast}(\omega) e^{i\omega y} e^{-\frac{(\omega-\omega_n)^2}{2(\alpha+\beta)}}. \tag{41}$$

With the variable change $\bar{\omega} = \omega - \omega_n$ and in the hypothesis that $t^{\circledast}(\bar{\omega}) \approx const$ over the significant support of the Gaussian centered in 0, integrating in $\bar{\omega}$ we have:

$$\sqrt{2\pi}\; const\; e^{-\frac{y^2\alpha}{2}} e^{i\omega_n y} e^{-\frac{y^2(\alpha+\beta)}{2}} \sim \lambda(\omega_n) e^{-\frac{y^2\beta}{2}} e^{i\omega_n y}. \tag{42}$$

Notice that this implies an upper bound on β since otherwise t would be white noise which is inconsistent with the diffraction-limited optics of the eye.
The condition is that the above holds approximately over the relevant y interval which is between $-\sigma_\psi$ and $+\sigma_\psi$. The approximate eigenfunctions ψ_1 (eg $n = 1$) has frequency ω_0. the minimum value of ω_0 is set by the condition that ψ_1 has to be roughly orthogonal to the constant (this assumes that the visual input does have a dc component, which implies that there is no exact derivative stage in the input filtering by the retina).

$$\langle \psi_0, \psi_1 \rangle = \int dx \; e^{-(\beta)x^2} e^{-i\omega_0 x} = 0 \;\Rightarrow\; e^{-\frac{(\omega_0)^2}{\beta}} \approx 0 \tag{43}$$

Using $2\pi f_0 = \frac{2\pi}{\lambda_0} = \omega_0$ the condition above implies $e^{-(\frac{2\pi\sigma_\psi}{\lambda_0})^2} \approx 0$ which can be satisfied with $\sigma_\psi \geq \lambda_0$; $\sigma_\psi \sim \lambda_0$ is enough since this implies $e^{-(\frac{2\pi\sigma_\psi}{\lambda_0})^2} \approx e^{-(2\pi)^2}$.
A similar condition ensures more in general orthogonality of any pair of eigen-

functions.

$$\int dx\ \psi_n^*(x)\psi_m(x) = \int dx e^{-(\beta)x^2} e^{in\omega_0 x} e^{-im\omega_0 x} \propto e^{-((m-n)\omega_0)^2\sigma_\psi^2},$$

which gives a similar condition as above. this also implies that λ_n should increase with σ_ψ of the Gaussian aperture, *which is a property of gabor wavelets!*.

2D $t^{\circledast}(\omega_x, \omega_y)$ approximately piecewise constant

We represent the template after retinal processing (but without motion) as:

$$t^{\circledast}(x,y) = \frac{1}{2\pi}\int d\omega_x d\omega_y t^{\circledast}(\omega_x,\omega_y) e^{i(\omega_x x + \omega_y y)} \tag{44}$$

and assume the following *ansatz*: the eigenfunctions have the form $\psi(x,y) = e^{-\frac{\beta}{2}x^2} e^{-\frac{\gamma}{2}y^2} e^{i\omega_g x}$, where β, γ and ω_g are parameters to be found.
With this assumptions eq. 38 reads:

$$\frac{1}{2\pi} e^{-\frac{\alpha}{2}(x^2+y^2)} \int d\xi d\eta\ e^{-\frac{\xi^2(\alpha+\beta)}{2}} e^{-\frac{\eta^2(\alpha+\gamma)}{2}} \int d\omega_x d\omega_y\ t^{\circledast}(\omega_x,\omega_y) \tag{45}$$

$$e^{i\omega_x(x-\xi)} e^{-i\omega_y\eta} e^{i\omega_\xi^g \xi} = \lambda(\omega_x^g,\omega_y^g) e^{-\frac{\gamma}{2}y^2} e^{-\frac{\beta x^2}{2}} e^{i\omega_x^g x} \tag{46}$$

Supposing $t^{\circledast}(\omega_x,\omega_y) = t^{\circledast}(\omega_x) t^{\circledast}(\omega_y)$ and $\lambda(\omega_x^g,\omega_y^g) = \lambda(\omega_x^g)\lambda(\omega_y^g)$ (which is the case if the spectrum is piecewise constant) we can separate the integral into the multiplication of the following two expressions:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\alpha}{2}x^2} \int d\xi\ e^{-\frac{\xi^2(\alpha+\beta)}{2}} \int d\omega_x\ t^{\circledast}(\omega_x) e^{i\omega(x-\xi)} e^{i\omega_\xi^g \xi} = \lambda(\omega_x^g) e^{-\frac{\beta x^2}{2}} e^{i\omega_x^g x}$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\alpha}{2}y^2} \int d\eta\ e^{-\frac{\eta^2(\alpha+\gamma)}{2}} \int d\omega_y\ t^{\circledast}(\omega_y) e^{-i\omega_y\eta} = \lambda(\omega_y^g) e^{-\frac{\gamma}{2}y^2}$$

The first equation is exactly the $1D$ problem analyzed in 4.3.1, meanwhile the second is satisfied if $\gamma = \alpha$.

Remark 13. *note that $\sigma_y \leq \sigma_\alpha$ and $\sigma_y \leq \sigma_x \leq \sigma_\alpha$, that is the "receptive fields" are elliptic Gaussians. This prediction is very robust wrt parameters and is clearly verified by the experimental data on simple cells across different species.*

4.4 Retina to V1: processing pipeline

Summary. The image is processed by the retina and the LGN before entering V1. Here we discuss how the spectrum of the image changes because of retinal processing. The main properties of the eigenvectors do not depend on it but some of the important quantitative properties – such as the linear relation between $lambda$ and σ – do. The question now is: what is the actual spectrum of t during development? Though the main qualitative properties of the eigenvectors of the cortical equation do

not depend on it, the quantitative relations do, since the kernel of the integral eigenvalue equation depends on t. In this section we describe models of processing in the retina up to V1 that affect the spectral properties of natural images and thereby determine the actual spectrum of t. We should also note that retinal waves may have a role in the development of cortex (c.f. [108]) in which case the spectrum of t during development (or part of development) may be independent of visual images and resemble more the simple case studied above of $t = t_0 + cos(\omega x)$. It may be possible to expose developing animals – for instance mice – to appropriately controlled artificial t,[25]. It is in any case interesting to check what various choice of t may yield.

4.4.1 Spatial and temporal derivatives in the retina

Let us start with the observation that the retina performs both a DOG-like spatial filtering operation as well as a high-pass filtering in time, roughly similar to a time derivative, probably to correct the slow signals provided by the photoreceptors. Natural images have a $\frac{1}{f}$ spatial spectrum, bandlimited by the optical point spread function at $60\frac{cycles}{degree}$ (in humans). Additional spatial low-pass filtering is likely to take place especially during development (in part because of immature optics).

This means that the spectrum of the patterns in the templatebook is spatially bandpass, likely with a DC component since the DOG derivative-like operation is not perfectly balanced in its positive and negative components. The temporal spectrum depends on whether we consider the faster *magno* or the slower *parvo* ganglion cells. The *parvo* or *midget* ganglion cells are likely to be input to the V1 simple cells involved in visual recognition. It is possible that the somewhat temporal high-pass properties of the retina and LGN (see [14]) simply correct *in the direction of motion* for the spatially low-pass components of the output of the retina (see Figure 17).
Consider as input to V1 the result $f(x, y; t)$ of an image $i(x, y)$ with a spatial power spectrum $\sim \frac{1}{\omega^2}$ filtered by the combination of a spatial low-pass filter $p(\omega)$ and then a bandpass dog. In this simple example we assume that we can separate a temporal filtering stage with a high-pass impulse response $h(t)$. Thus in the frequency domain

$$f(\omega_x, \omega_y; \omega_t) \sim i(\omega_x, \omega_y; \omega_t)p(\omega_x, \omega_y)dog(\omega_x, \omega_y).$$

Assume that $f(x, y, t)$ is then filtered through $h(t)$. For example, let us see the implications of $h(t) \sim \frac{d}{dt}$. Consider the effect of the time derivative over the signal generated by the translation of an image $f(x - vt)$, where x, v are vectors in $\setminus^2$:

$$\frac{\mathrm{d}\, I}{\mathrm{d}\, t} = \nabla I \cdot v. \tag{47}$$

assume for instance that the direction of motion is along the x axis, eg $v_y = 0$. Then

$$\frac{\mathrm{d}\, I}{\mathrm{d}\, t} = \frac{\partial I}{\partial x} v_x. \tag{48}$$

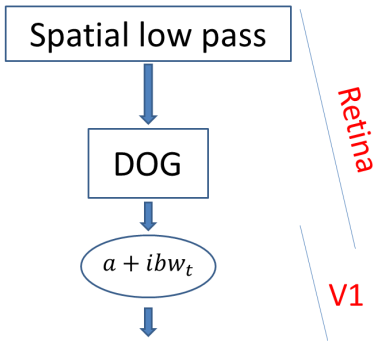


Figure 17: *The sequence of processing stage from the retina with spatial low-pass and bandpass (DOG) plus temporal d/dt derivative-like filtering to V1. Thus high-pass temporal filtering compensates for the spatial blurring in the direction of motion.*

Thus the prediction is that motion in the x direction suppresses spatial changes in y, eg spatial frequencies in ω_y, and enhances components orthogonal to its direction. This means that the time derivative of a pattern with a uniform spatial frequency spectrum in a bounded domain ω, as an effect of motion along x, gives a templatebook with a spectrum in ω which reflects the transformation and *not only the spectrum of the image and the filtering of the retina*: $i\omega_x f(\omega_x, \omega_y)$. Notice that spatial and temporal filtering commute in this linear framework, so their order (in the retina) is not important for the analysis. In particular, a high pass time-filtering may exactly compensate for the spatial-low pass operation *in the direction of motion* (but not in the orthogonal one). Interestingly, *this argument is valid not only for translations but for other motions on the plane.* From now on, we assume the pipeline of figure 17. The 2D simulations are performed with this pipeline using the low-pass filter of Figure 18.

Because of our assumptions, invariances to affine transformations are directly related to actual trajectories in $\mathbb{R}^2$ of the image while transforming. These are flows on the plane of which a classification exist (see Appendix18). We have the following result for the solution of the 2D eigenfunction equation in the presence of oriented motion:

Lemma 4. Selection rule
Assume that a templatebook is obtained after the $\nabla^2 g \circ \frac{\partial}{\partial t}$ filtering of a "video" generated by a transformation which is a subgroup of the affine group $Aff(2, \mathbb{R})$. Then the components in the image spectrum orthogonal to the trajectories of the transformations are preferentially enhanced.

4.5 Cortical equation: predictions for simple cells in V1

Summary. The numerical simulations predict surprisingly well, almost without any

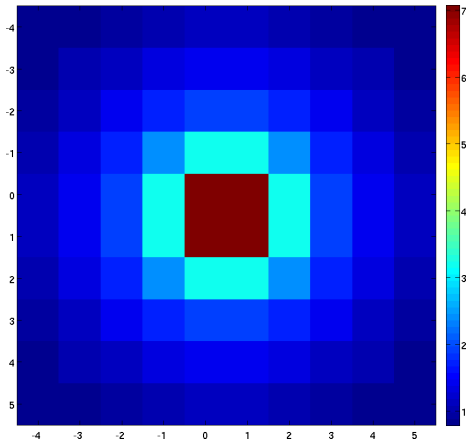


Figure 18: *Spatial lowpass filter* $\frac{1}{\sqrt{(x^2+y^2)}}$. *In the frequency domain the filter is* $1/\sqrt{\omega_x^2+\omega_y^2}$. *The filter is such that its effect is canceled in the direction of motion by the time derivative.*

parameter fitting, quantitative properties of the tuning of simple cells in V1 across different species.

Numerical simulations of the cortical equation in $2D$ using natural images moving in one direction and the pipeline of Figure 17 show that the top eigenvectors are oriented Gabor-like wavelets. We are mostly interested in the top three eigenvectors, since they are the ones likely to be relevant as solutions of a Oja-type equation. Figures 19 and 20 shows that the solutions are very close to actual Gabor wavelets. A number of other simulations (not shown here) together with the previous theoretical analysis suggests that the Gabor-like form of the solution is robust wrt large changes in the form of the signal spectrum.

Some of the other more quantitative properties however seem to depend on the overall shape of the effective spectrum though in a rather robust way. In this respect the simulations agree with the astonishing and little known finding that data from simple cells in several different species (see Figure 22) show very similar quantitative features.

The most noteworthy characteristics of the physiology data are:

- the tuning functions show a λ proportional to σ which is the signature of wavelets;
- in particular λ is always finite;
- $\sigma_y > \sigma_x$ always where x is the direction of motion and the direction of maximum modulation.

The 2D simulations with the pipeline described earlier reproduce these properties without any parameter fitting process. In particular, Figure 24 shows that $\sigma_y > \sigma_x$. Figure 25 summarizes the main quantitative properties of the simulations. Figure 26 shows that the simulations seem to be consistent with

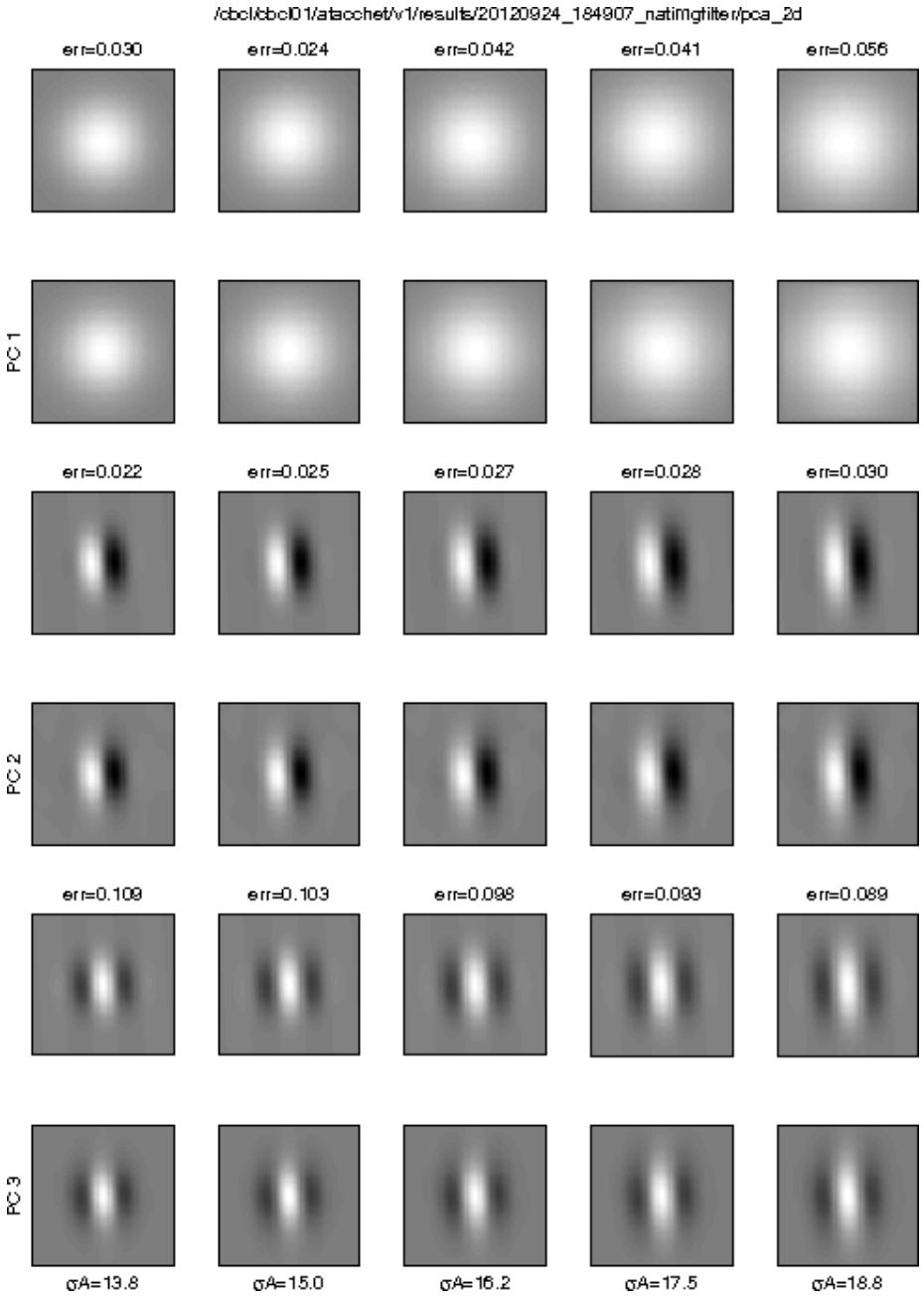


Figure 19: *Simulation results for V1 simple cells learning via PCA. Each "cell" receives as input all frames from a movie generated by a natural image patch undergoing a rigid translation along the horizontal axis. A Gaussian blur filter, a Difference of Gaussians filter and the spatial lowpass filter reported in Figure 18 are applied to every frame as a preprocessing step. A Gaussian mask is then overlaid on each frame to model a cell's receptive field. Lastly the weighted difference between two subsequent frames is fed to the learning stage, to simulate and imperfect temporal derivative (the weights we used are (-0.95, 1.00) so as not to suppress the DC component of the frames completely). Each movie is generated using 40 distinct patches go by one after the other. Each cell then "learns" its weight vector extracting the principal components of its input. For each row pairs: the top row shows the best Gabor fit (least squares) and the bottom row shows the actual principal component vector; different columns represent different values for the Gaussian mask aperture.*

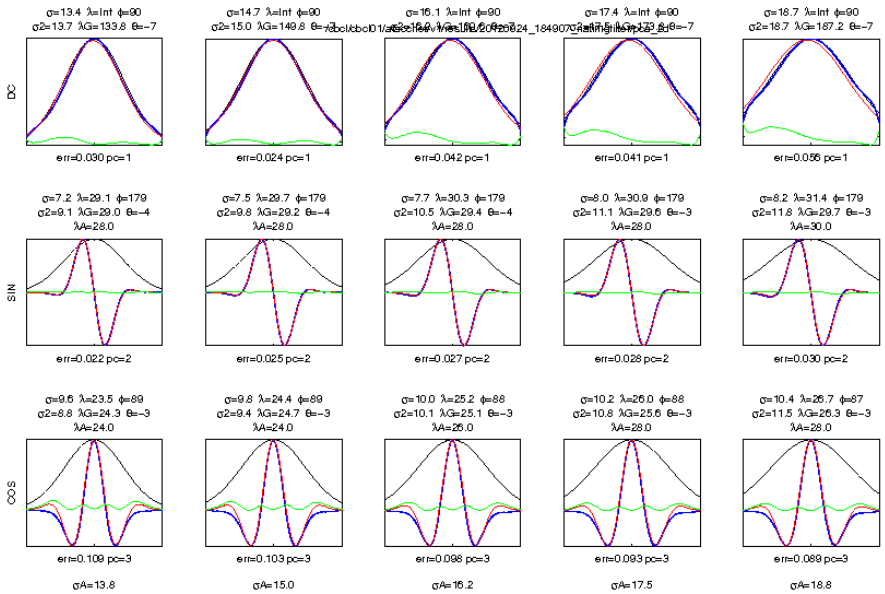


Figure 20: *1D sections of the principal components sorted by eigenvalue (row) for different Gaussian apertures (column). Red indicates best least square fit of a Gabor wavelet. The pipeline is the same described in Figure 19.*

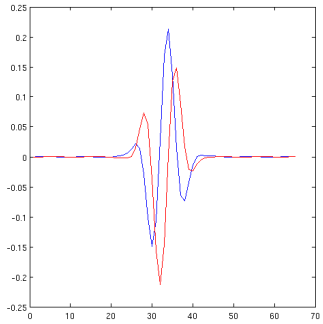


Figure 21: *A vertical slice through a quadrature pair (1st and 2nd eigenvector) from Figure 19*
.

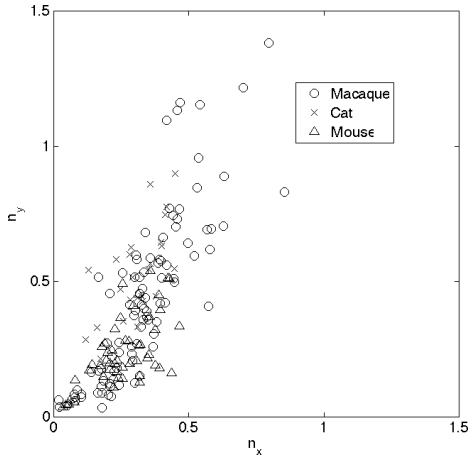


Figure 22: *Data from [41] (cat), [86] (macaque) and [68] (mouse). Here $n_x = \sigma_x f$ where σ_x is the standard deviation of the Gaussian envelope along the modulated axis and $f = \frac{2\pi}{\lambda}$ is the frequency of the Gabor's sinusoidal component. Likewise, $n_y = \sigma_y f$ where σ_y is the sigma of the Gaussian envelope along the unmodulated axis.*

the data across species. Notice that a better fitting may be obtainable with a minimum of parameter optimization.

The form of the low-pass filtering – a spatial average that cancels the time derivative in the direction of motion – seems to be important. When the filter is replaced by a Gaussian low pass filter, the slope of λ wrt σ becomes too small (see Appendix 16.8.2).

The image spectrum before the retinal processing matters. For instance, if instead of natural images a white noise pattern is moved, the key properties (see Figures 27 and 28) of the tuning functions are lost: λ is essentially constant, independent of σ.

An interesting question arises about the actual role of motion in the development of tuning in the simple cells. In our theoretical description, motion determines the orientation of the simple cells tuning. We cannot rule out however the possibility that motion is not involved and orientations emerge randomly (with orthogonal orientations for different eigenvectors, as in figure 29), in which different natural images, randomly chosen, were used as input to the eigenvector calculation, instead of a motion sequence. It would be interesting to examine experimentally predictions of these two possible situations. The first one predicts that all the eigenvectors generated for a simple cell during development have the same orientation; the second predicts orthogonal orientations during learning. Unfortunately, verifying this prediction is experimentally difficult. There is however another property – the relation between λ and σ – that distinguish these two mechanisms allowed by the theory. The prediction from our simulations is that motion yields finite λ (see Figure 25) whether absence of motion implies that some λ go to infinity (see Figures 29

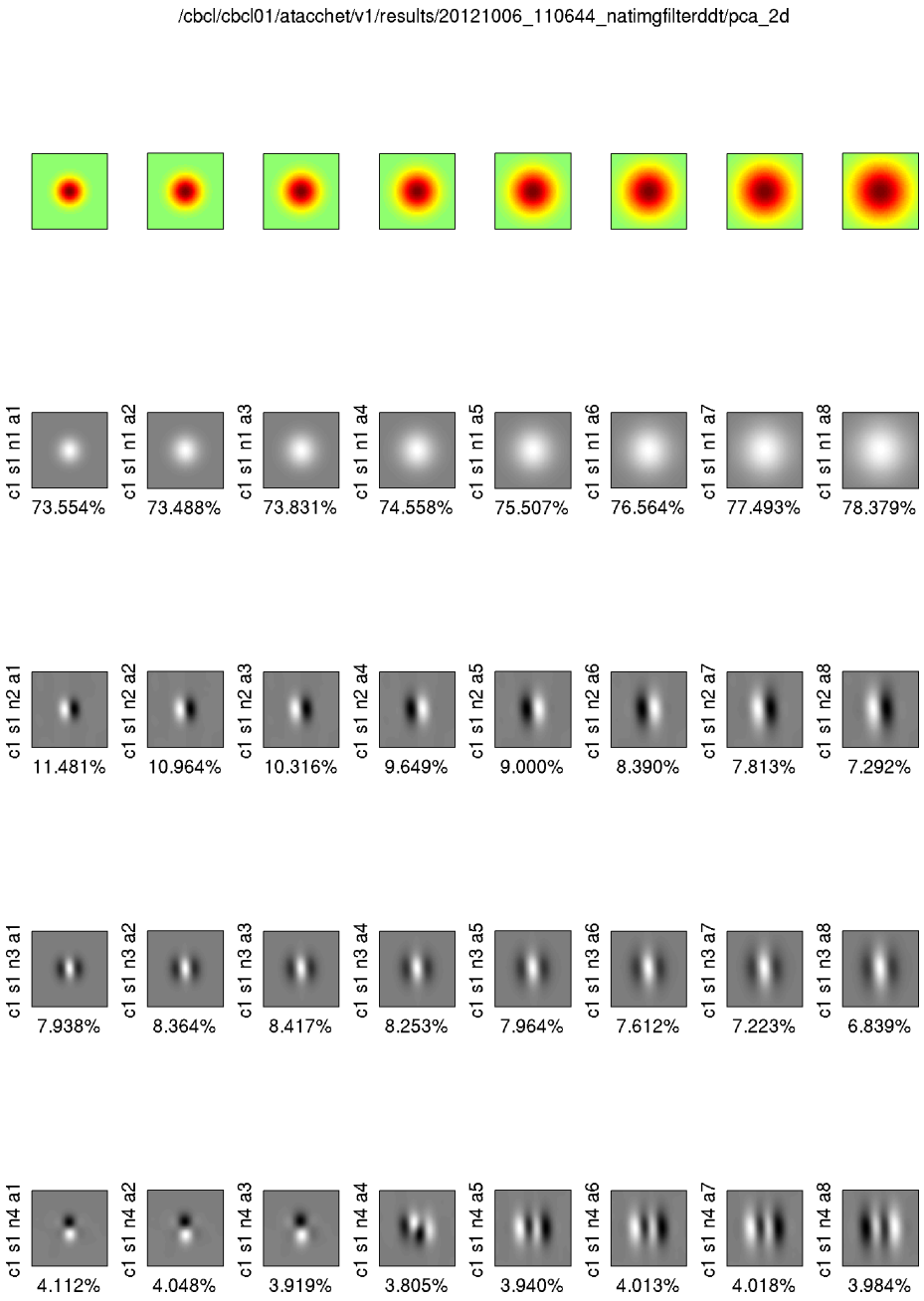


Figure 23: *Principal components of the template book. These are obtained using the pipeline described in Figure 19. The pipeline consists of a Gaussian blur, a DoG filter, a spatial low-pass filter* $1/\sqrt{\omega_x^2 + \omega_y^2}$ *and an imperfect temporal derivative. The principal components are sorted by eigenvalue (row), different columns refer to different apertures of the receptive field.*

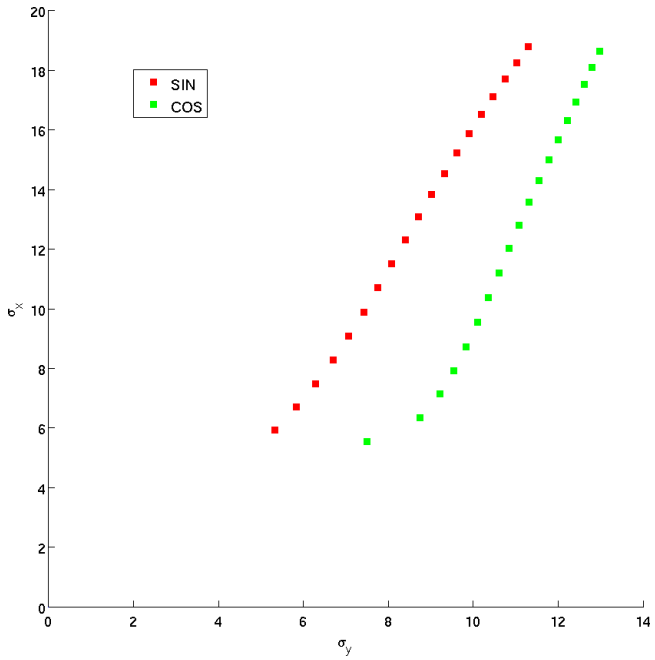


Figure 24: *Width of the Gaussian envelope for the modulated and unmodulated directions in cells learned using the pipeline described in Figure 19.*

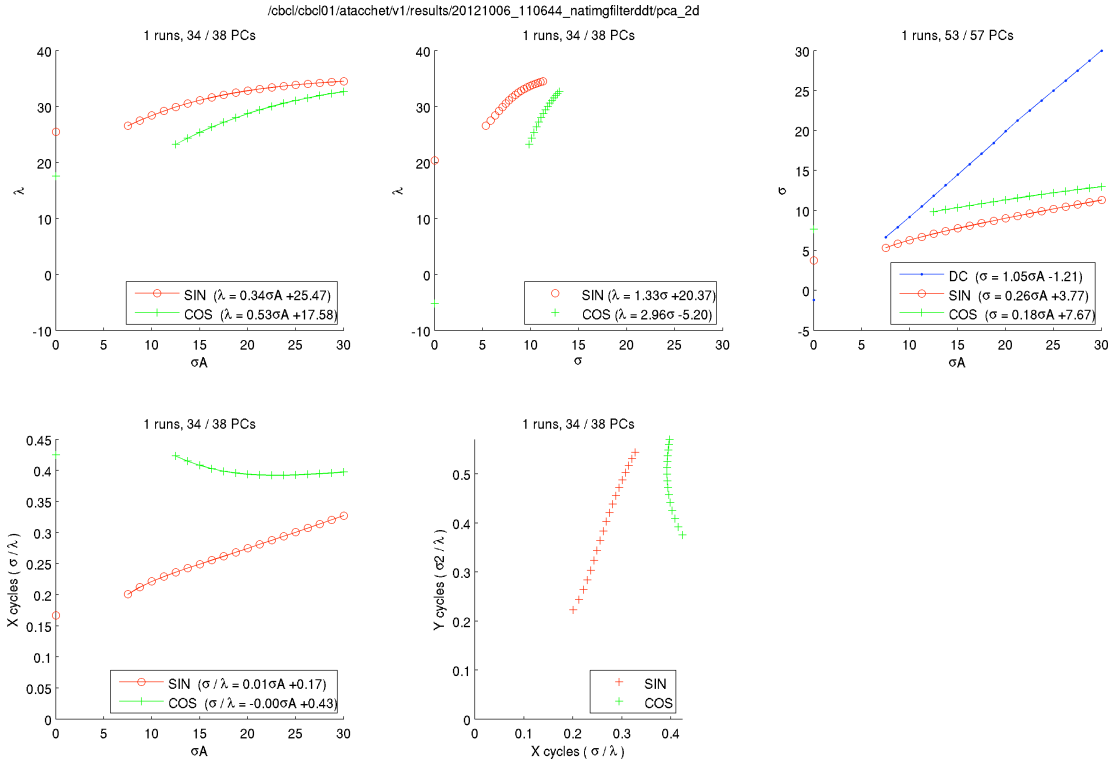


Figure 25: *Summary plots for 2D simulations of V1 cells trained according to the pipeline described in Figure 19. Figures from top left to bottom right: a) sinusoid wavelength (λ) vs. Gaussian aperture width (σ_α). b) Sinusoid wavelength (λ) vs. Gaussian envelope width on the modulated direction (σ). c) Gaussian envelope width for the modulated direction (σ) vs. Gaussian aperture width (σ_α). d) Ratio between sinusoid wavelength and Gaussian envelope width for the modulated direction (n_x) vs. Gaussian aperture width (σ_α). e) Ratio between sinusoid wavelength and Gaussian envelope width on the unmodulated direction (n_y) vs. ratio between sinusoid wavelength and Gaussian envelope width for the modulated direction (n_x). The pipeline consists of a Gaussian blur, a DOG filter, a spatial low-pass filter $1/\sqrt{\omega_x^2 + \omega_y^2}$ and an imperfect temporal derivative. Parameters for all filters were set to values measured in macaque monkeys by neurophysiologists.*

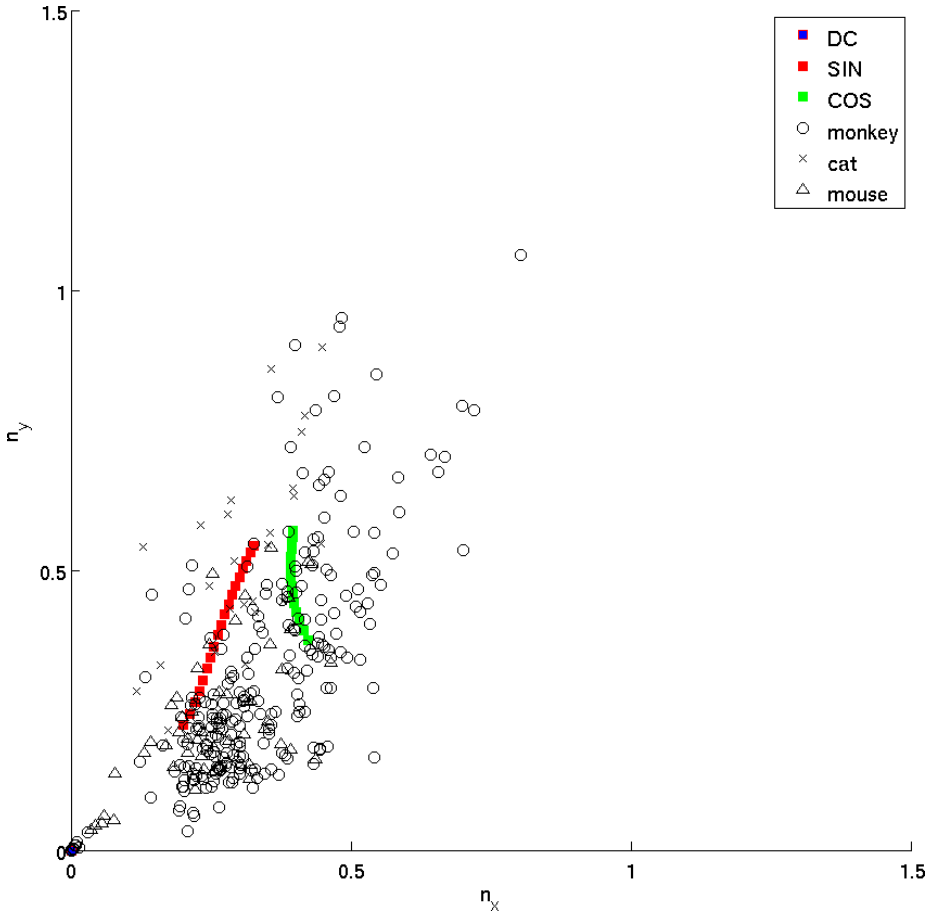


Figure 26: *This figure shows $n_y = \frac{\sigma_y}{\lambda}$ vs. $n_x = \frac{\sigma_x}{\lambda}$ for the modulated (x) and unmodulated (y) direction of the Gabor wavelet. Notice that the slope is $\frac{\sigma_y}{\sigma_x}$ – a robust finding in the theory and apparently also in the physiology data. Neurophysiology data from monkeys, cats and mice are reported together with our simulations. Simulated cells learn their weight vector according to the algorithm described in Figure 19.*

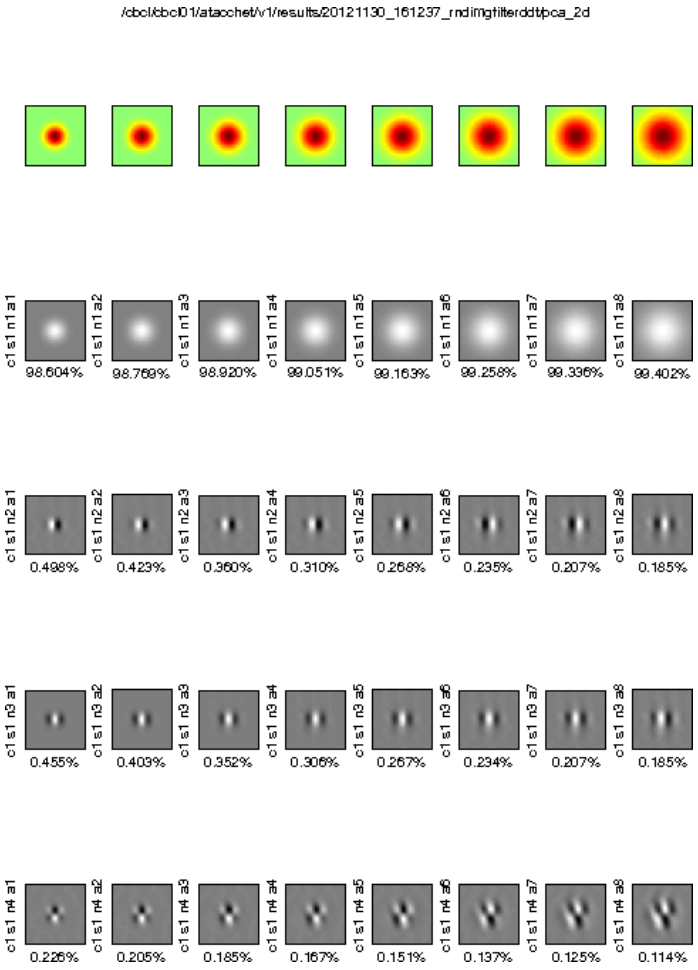


Figure 27: *Principal components of the template book. These are obtained using the pipeline described in Figure 19 on image patches that contain random noise instead of natural images. The pipeline consists of a Gaussian blur, a DoG filter, a spatial low-pass filter* $1/\sqrt{\omega_x^2 + \omega_y^2}$ *and an imperfect temporal derivative. The principal components are sorted by eigenvalue (row), different columns refer to different apertures of the receptive field.*

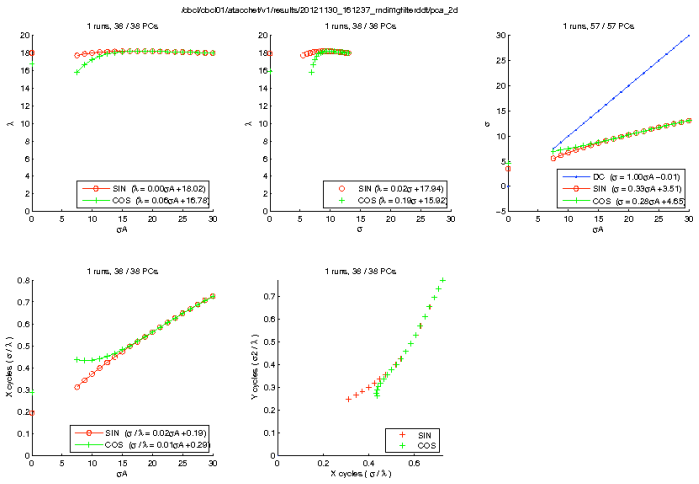


Figure 28: *Properties of the eigenfunctions of the template book obtained translating a white noise pattern and using the pipeline described in Figure 19, the template book learned for this case is in Figure 27. Plots are ordered as in Figure 25.*

and 30). Physiology data (see Figure 22) support then a key role of motion during development! Further checks show that without motion λ can be infinite even without spatial low pass filtering (see Appendix **??**).

Remarks

- *Gabor-like wavelets and motion* We have seen that motion is not necessary to obtain Gabor-like wavelets but is required for the right properties, such as finite λ.

  The story goes as follows. Originally the theory assumed that the covariance of the 2D input has the form $t^{\circledast}(x, y) = t^{\circledast}(y - x)$ with $x \in \mathbb{R}^2$ and $y \in \mathbb{R}^2$ *because* of shifts in the input images (that is because of motion of the recorded images).

  However, it turns out that the empirical estimate of the covariance of randomly sampled static images (assumed to be $\mathbb{E}[I(x)(y)]$ has the same, shift-invariant structure *without* motion. For images of natural environments (as opposed to images of cities and buildings) the covariance is approximately a *radial* function, eg $t^{\circledast}(x, y) \approx t^{\circledast}(||x - y||)$, therefore invariant for shifts and rotations. Scale invariance follows from the approximate $\frac{1}{\omega^2}$ power spectrum of natural images [99]. Further, natural images have a power spectrum $|I(\omega_x, \omega_y)|^2 \approx \frac{1}{\omega^2}$, where $\omega = (\omega_x^2 + \omega_y^2)^{-\frac{1}{2}}$. A power spectrum of this form is invariant for changes in scale of the image $I(x, y)$ and is an example of a power law. A related **open question** is whether these spectrum symmetries are reflected in the form of of the eigenfunctions.

- The Appendix (section **??**) collects a few notes about transformations and

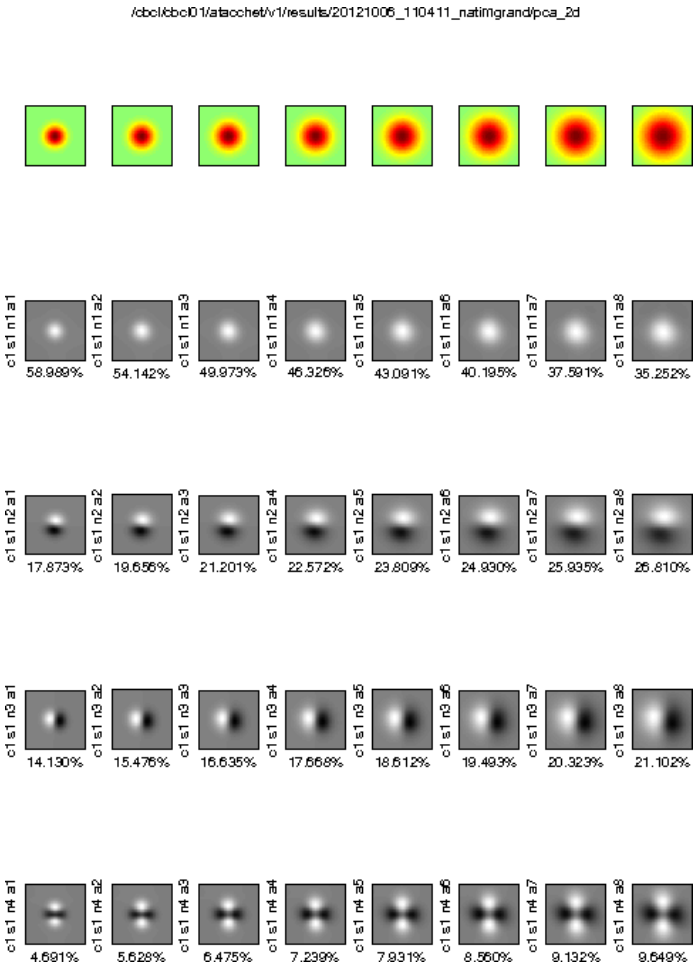


Figure 29: *Eigenvectors of covariance matrix of scrambled set of images. These are obtained using the pipeline described in Figure 19 but before applying the time derivative filter the frames are scrambled. This is the only difference in the pipeline, the exact same 40 preprocessed natural image patches are fed to the learning module.*

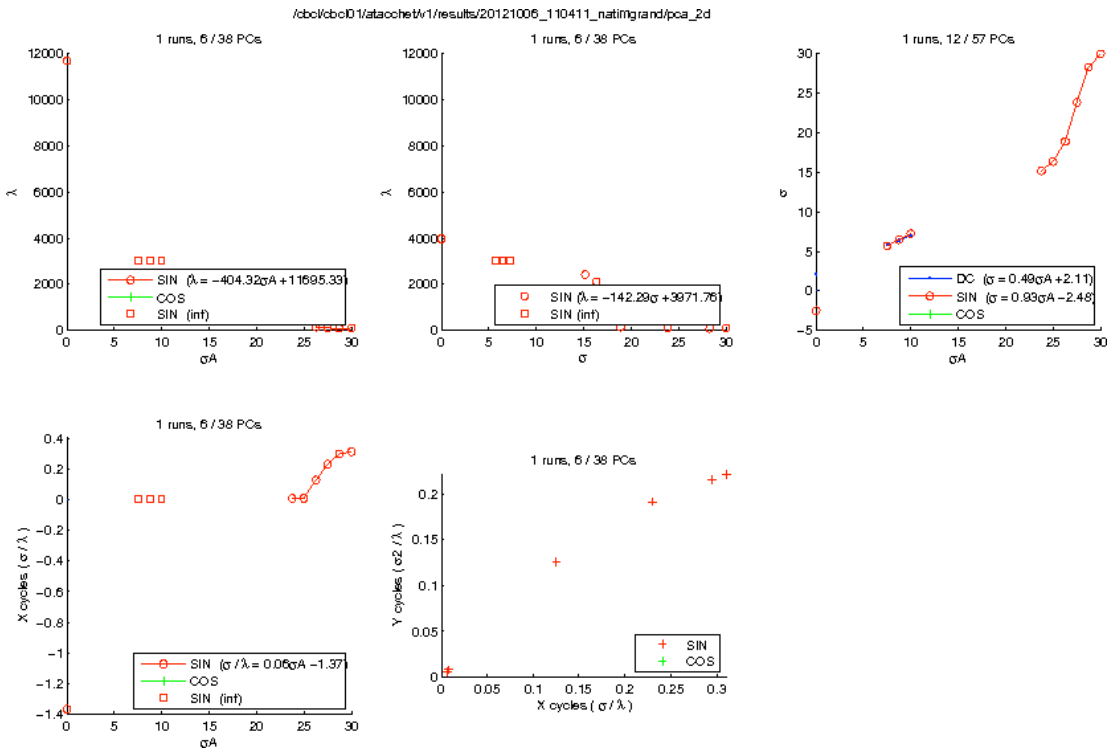


Figure 30: *Properties of the eigenfunctions the template book learned using the pipeline described in Figure 29. Plots are ordered as in Figure 25.*

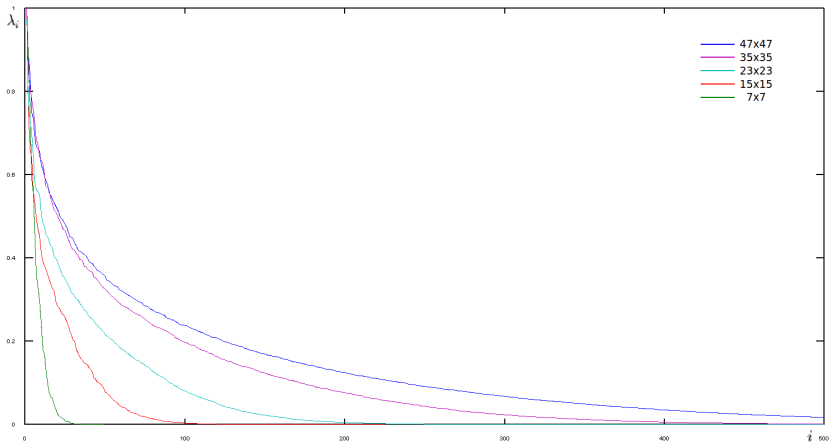


Figure 31: *Eigenvalues behavior as a function of the aperture for general Hilbert-Schmidt integral operators.*

spectral properties of them.

- The hypothesis explored here, given our pipeline containing a time derivative and PCA, is related to maximization of the norm of the time derivative of the input patterns (or more precisely a high-pass filtered version of it). This is related to – but almost the opposite of – the "slowness" principle proposed by Wiskott ([107, 18]) and made precise by Andreas Maurer.See also appendix 16.9.

- *Receptive fields size and eigenvalues distribution.* Simple properties of the eigenfunctions of integral operators of the Hilbert-Schmidt type imply two rather general properties of the receptive fields in different layers as a function of the aperture:

  Proposition 6. *(Anselmi, Spigler, Poggio)*

  – *Under the assumption of a power spectrum of the form* $t(\omega) \propto \frac{1}{\omega^2}$*, the eigenvalues obey the relation:*

  $$\frac{\lambda_i(\sigma)}{\lambda_i(\bar{\sigma})} \geq 1, \quad \sigma \geq \bar{\sigma}.$$

  This suggests that the top eigenvalues are closer to each other for large apertures, suggesting that in the presence of noise the eigenvector emerging as the result of Oja's flow may vary among the several top eigenvectors.

  – *The number of eigenfunctions depends on the size of the receptive field: this also suggests that the variety of tunings increases with the size of the RFs.*

4.6 Complex cells: wiring and invariance

Summary.We show that local PCA can substitute for templates in the sense that group averages over nonlinear functions of the PCA may be invariant. This is true in particular for modulo square nonlinearities. The section analyzes the connection between the simple complex cells stage of our theory and the first iteration of Mallat's scattering transform [63] (see also **??** *for more details).*

In the theory, complex cells are supposed to pool nonlinear functions of (shifted) templates over a small bounded domain in x, y, representing a partial group average. Clearly, pooling the modulo square of the top Gabor-like eigenvectors over a x, y domain is completely equivalent (since the eigenvectors are legitimate templates). Interestingly, pooling the modulo square of the top Gabor-like wavelets is also equivalent to a partial group average over a (small) domain. This can be seen (and proven) in a number of ways. The intuition is that the Gabor-like eigenvectors capture the transformations seen through the Gaussian windows (exact reconstructions of all the frames can be achieved by using all the eigenvectors; optimal L^2 approximation by using a smaller number). Thus pooling over the squares of the local eigenvectors is equivalent to pooling the squares of the templates (eigenvectors are orthogonal), assuming that the templates are normalized, over the aperture used for the eigenvector computation. This intuition shows that some invariance can be obtained locally. In fact, local pooling of the modulo square (of simple cells at the same x, y) increases invariance; extending the range of pooling to a domain in x, y of course increases the range of invariance. Thus pooling over eigenvectors In the case of Gabor wavelets the modulo square of the first quadrature pair is sufficient to provide quite a bit of invariance: this is shown by a reasoning similar to Mallat's [63]. The sum of the squares of the quadrature pair is equal to the modulo of each complex wavelet which maps a bandpass filter portion of the signal into a low-pass signal. In the Fourier domain the low pass signal is a Gaussian centered in 0 with the same σ_ω as the wavelet (which is roughly $\frac{1}{2}\omega_0$, the peak frequency of the Fourier transform of the wavelet). Thus a rapidly changing signal is mapped into a much slower signal in the output of the C cells. There is in fact an almost perfect equivalence between the simple complex stage of the theory here and the first iteration of the scattering transform ([63]). We discuss related issues next.

4.6.1 Complex cells invariance properties: mathematical description

Let $L^2(\mathcal{G}) = \{F : \mathcal{G} \to \mathbb{R} \mid \int |F(g)|^2 dg < \infty\}$, and

$$T_t : \mathcal{X} \to L^2(\mathcal{G}), \qquad (T_t f)(g) = \langle f, T_g t \rangle ,$$

where $t \in \mathcal{X}$. It is easy to see that T_t is a linear bounded and compact[5] operator, if $\|T_g t\| < \infty$. Denote by $(\sigma_i; u_, v_i)_i$ the singular system of T_t, where $(u_i)_i$ and $(v_i)_i$ are orthonormal basis for $\mathcal{X}$ and $L^2(\mathcal{G})$, respectively.

[5]In fact it is easy to see that T is Hilbert Schmidt, $\mathrm{Tr}(T_t^* T_t) = \int dg \, \|T_g t\|$

For $\sigma : \mathbb{R} \to \mathbb{R}$ measurable, define (complex response)

$$c : \mathcal{X} \to \mathbb{R}, \qquad c(I) = \sum_i \sigma(\langle I, u_i \rangle).$$

If $\sigma(a) = |a|^2$, $a \in \mathbb{R}$ and T_t/b_t is an isometry, where b_t is a constant possibly depending on t (see [30]), then c invariant. Indeed,

$$c(I) = \|I\|^2 = \frac{1}{b_t^2} \|T_t I\|^2_{L^2(\mathcal{G})} = \frac{1}{b_t^2} \int |\langle I, T_g t \rangle|^2 dg,$$

for all $I \in \mathcal{X}$, and $\|T_t I\|^2_{L^2(\mathcal{G})} = \|T_t T_{g'} I\|^2_{L^2(\mathcal{G})}$ for all $g' \in \mathcal{G}$.

Example 3 (Affine Group)**.** *If $\mathcal{G}$ is the affine group and $\mathcal{X} = L^2(\mathbb{R})$, then under the admissibility condition*

$$\int |\langle T_g t, t \rangle|^2 < \infty,$$

it is possible to take $b_t = \sqrt{C_t}$, with $C_t = 2\pi \int \frac{|\hat{t}(\omega)|^2}{\omega} d\omega$, where $\hat{t}$ denotes the Fourier transform of t.

4.6.2 Hierarchical frequency remapping

The theory so far does not provide information about the size of the receptive fields for the first layer S and C cells. Here we sketch an approach to this question which is related to section 10.4. A main difference is that we consider here the specific case of templates being Gabor wavelets and of pooling being energy pooling over a bounded interval. Thus we consider a partial group average of the squares.

We begin by considering one dimensional "images". Let the image $I(x) \in \mathcal{X}$. To analyze $I(x)$ we use a wavelet centered in ω_0, $x * \psi_{\omega_0,\sigma_0}$ where σ_0 is the width σ_{1s} of the wavelet Gaussian envelope, that is of the envelope of the simple cells impulse response at fist layer. There are several such channels centered on different frequencies and with corresponding σ resulting from Hebbian learning as described in previous sections such as 4.4.1. As an example the highest frequency channel may be centered on a frequency ω_0 that satisfies $\omega_{max} \leq \omega_0 + 3\hat{\sigma}_0$ with $max(supp(\hat{I})) = \omega_{max}$.

The signal I can be reconstructed exactly – apart from its DC and low frequencies around it – by combining a sufficiently large number of such *bandpass* filters according to the identity $\int G(\omega - \omega')d\omega' \hat{I}(\omega) = \hat{I}(\omega)$.

The pooling operation, from simple to complex cells, starts with taking the modulus square of the wavelet filtered signal. In Fourier space, the operation maps the support of he Fourier transform of $I * \psi_{\omega_0,\sigma_0}$ into one interval, centered in 0.

A one-octave bandwidth – that we conjecture is the maximum still yielding full information with a low number of bits (see Appendix 12.1)) – implies a certain size of the receptive field (see above) of simple cells. Complex cells

preserve information about the original image if the pooling region is in the order of the support of the simple cells (thus in the order of 6σ), since we assume that the sign of the signal is known (positive and negative parts of the signal are carried by different neural channels, see Appendix 12.1). The same reasoning can be also applied to higher order simple cells learned on the 4-D cube (see later) to obtain estimates of RF size at a fixed eccentricity. Interestingly, these arguments suggest that **if** information is preserved by pooling (which is not necessary in our case), then there the C cells pooling regions are very small (in order of $\sqrt{2}$ the simple cells receptive fields): most of the invariance is then due to the RF of simple cells and the pooling effect of the modulo square (sum over quadrature pairs).

4.7 Beyond V1

Summary. We show that the V1 representation – in terms of Gabor-like wavelets in x, y, θ, s – can locally approximate (within balls of radius r with $\frac{r}{R} \leq \delta$ where R is the retinal eccentricity) similitude transformations of the image as independent shifts in a 4-dimensional space (the subgroup of translations is a 2-parameter group (translations in x, y); the subgroup of rotations and dilations is also a two parameters group (ρ, θ)). Thus learning on the V1 representation can be such to generate 4-dimensional wavelets. This scenario – which is one of the several possible for processing stages above V1 – seems consistent with physiology data. Assuming that R is retinal eccentricity corresponds to assuming that most of the experienced and learned rotations and loomings are centered in the fovea.

4.7.1 Almost-diagonalization of non commuting operators

Let us start from the fact that if $(e_i, i = 1, ..., N)$ is an orthonormal basis in any finite Hilbert space, the matrix whose entries are $a_{i,j} = \langle Ae_i, e_j \rangle$ is diagonal if and only if each e_i is an eigenfunction of the operator A:

$$a_{i,j} = \langle Ae_i, e_j \rangle = \lambda_i \langle e_i, e_j \rangle = \lambda_i \delta_{i,j}$$

If another operator B acting on the Hilbert space is such that $[A, B] = 0$ the two operators share the same eigenfunctions and can therefore be simultaneously diagonalize. For example in the case of the Fourier basis $\{e^{ix\omega}$ we can say that the Fourier transform diagonalize any operator that commutes with translation.
What can we say if we have two commuting operators, A, B? In this case we cannot have simultaneous diagonalization but choosing a basis e_i of the Hilbert space we have

$$\langle Ae_i, e_j \rangle = a_{i,j} + \Delta(A)_{i,j}$$
$$\langle Be_i, e_j \rangle = b_{i,j} + \Delta(B)_{i,j}.$$

since the eigenvalues (the measurement results) cannot be determined with infinite precision at the same time. In this case we can speak of almost simultaneous diagonalization of the operators A, B if there exists a basis ψ_i that

minimize simultaneously $\Delta(A)_{i,j}, \Delta(B)_{i,j},\ i \neq j$. This corresponds to find the set of functions ψ that minimize the uncertainty principle

$$(\Delta_\psi A)(\Delta_\psi B) \geq \frac{1}{2}|[A, B]_\psi|$$

Example 4. *The Weyl-Heisenberg group in one dimension is generated by two non commuting operators, the translation in frequency and space. The minimizers of the associated uncertainty relations gives Gabor functions as solutions.*

Example 5. *The affine group in dimension two...*

4.7.2 Independent shifts and commutators

(From [17])

Theorem 11. *Given two Lie transformation groups, T_a and S_b, acting on an image $f(x, y) \in L^2(\mathbb{R}^2)$, there exists a representation of the image $g(u, v)$, $(u = u(x, y), v = v(x, y))$ such that*

$$\begin{aligned} \mathcal{L}_a u = 1, \quad \mathcal{L}_b v = 0 \\ \mathcal{L}_b u = 0, \quad \mathcal{L}_a v = 1 \end{aligned}$$

where $(\mathcal{L}_a, \mathcal{L}_b)$ are the lie generators of the transformations, if $\mathcal{L}_a$ and $\mathcal{L}_b$ are linearly independent and the commutator $[\mathcal{L}_a, \mathcal{L}_b] = 0$.

The last two equations state that, in the new coordinate system (u, v) the transformations T_a and S_b are translations along the u and v axes, respectively (and each translation is independent from the other).

Example 6. *In the case we consider dilation and rotation transformations we have that there exists a coordinate change such that, in that coordinate system rotations, and dilations are translations being $\mathcal{L}_r$ independent from $\mathcal{L}_r$ and $[\mathcal{L}_r, \mathcal{L}_d] = 0$*

4.7.3 Hierarchical wavelets: 4-cube wavelets

As a consequence of what found in the previous paragraphs a group transformation in the image space $\mathcal{X}$ is a shift in the space $L^2(SIM(2))$ where the function $c_n(I)$ is defined. In this approximation the transformations at the second layer can be written as direct product of translation group in the group parameters:

$$G = \mathbb{R}^2 \times \mathbb{S}_1 \times \mathbb{R} \tag{49}$$

The same reasoning applied at the first layer for the the translation group can be repeated: the eigenfunctions will be Gabor-like wavelets in the parameter group space.
The theoretical considerations above imply the following scenario. In the first layer, exposure to translations determines the development of a set of receptive

fields which are an overcomplete set of Gabor-like wavelets. The space of two-dimensional images – functions of x, y – is effectively expanded into a 4-cube of wavelets where the dimensions are x, y, θ, s, eg space, orientation and scale, (see fig. 4.7.3).

The same online learning at the level of the second layer (S2) with apertures "looking" at a Gaussian ball in x, y, θ, s will converge to Gabor-like wavelet after exposure to image translations, which induce translations in x, y of the 4-cube. Informally, the signature of a patch of image at the first layer within the aperture of a S2 cell will consist of the coefficients of a set of Gabor wavelets at different orientations and scales; after processing through the S2 second order wavelets and the C2 aggregation function it will be invariant for local translations within the aperture.

In the example above of x, y translation of the image, the second-order wavelets are wavelets parallel to the x, y plane of the 4-cube. For image motion that include rotations and looming, the resulting motion in the 4-cube is mostly still locally a shift – but in general along a diagonal in the 4-cube. Thus, in general, second-order wavelets are Gabor-like oriented along diagonals inhey are also x, y, θ, s (apart from a minority of polar wavelets near the fovea, see below).

Of course, the argument above are recursive with higher levels behaving as the second level. Not surprisingly, the tuning properties, seen from the image, of higher order wavelets is more complex: for instance shifts in scale correspond to receptive fields for which the preferred stimulus may be similar to concentric circles.

The theory predicts that pooling within the 4-cube takes place over relatively small balls in which rotations and expansions induce approximately uniform shifts in x, y together with uniform changes in orientations or scale. For this to happen the radius of the ball has to decrease proportionally to the distance from the center of rotation. If this is assumed to be the fovea then we derive the prediction that the size of receptive fields of complex cells should increase linearly with eccentricity – a prediction consistent with data (see [20]).

Remarks

- Mallat also considers wavelets of wavelets [63]. In his case all the wavelets are in x, y only with orientation and scale as parameters, whereas in the simple cells of V2 or higher we expect wavelets on x, y, orientation and scale.
- $V1$ (may be with $V2$) diagonalize the affine group: how can we check this prediction?

4.7.4 Predictions for V2, V4, IT

If during learning gaze is precisely maintained, then neurons which happen to contain the center of rotation and looming could develop wavelets in polar coordinates. The probability of this occurring is probably very low for any of

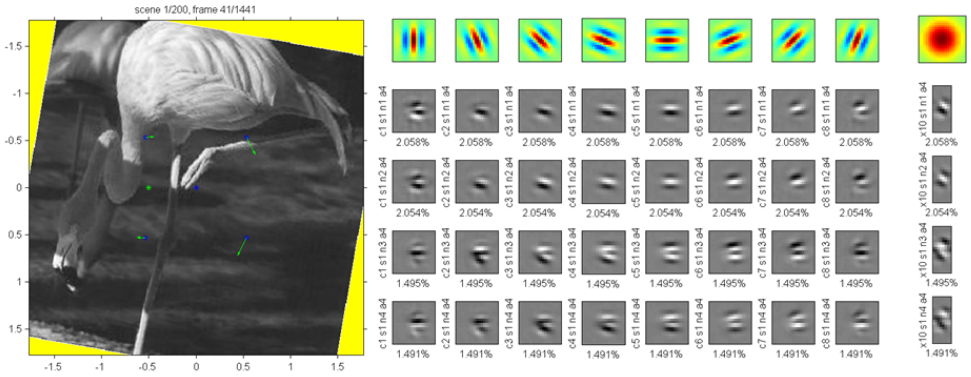


Figure 32: *Learning an S2 filter from C1 outputs (of a single scale only). Here the transformation is off-center rotation. The resulting S2 filters are Gabor filters in 3 dimensions: x; y, and orientation. Left: the receptive field center is in the middle (central blue asterisk) but the center of rotation is to the left (green asterisk). The green arrows show the speed of optical flow at various places. Middle: the learned filters. Each row represents a single filter; since the filters are 3D, we show a separate (x, y) plane for each orientation. However, in this view it is not easy to see shifting along the orientation dimension. Right: here we show that the 3D Gabors also have a sinusoidal component along the orientation dimension. We show a single slice, at the central X position, for each filter. The slices are planes in (y, orientation).*

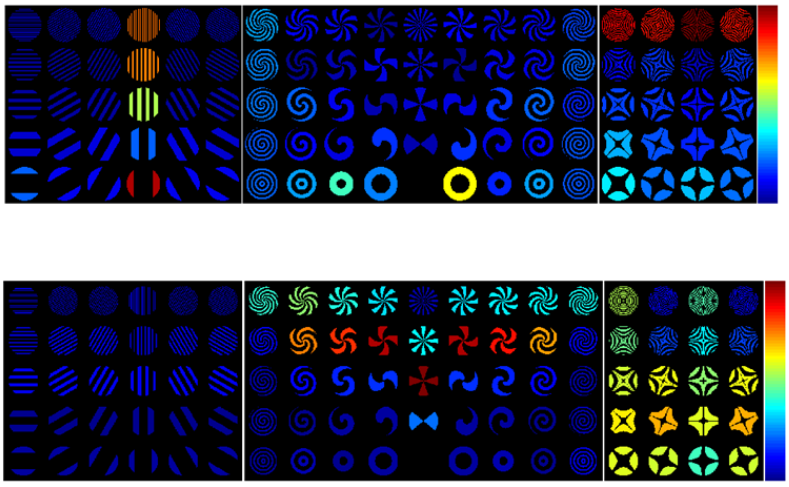


Figure 33: *It is not easy to predict the optimal stimulus – in pixel space – of a 3D Gabor filter computed in a space of 2D Gabor outputs. Here we test two model C2 units with stimuli used in [24] to test neurons in Macaque V4. Simply by changing the 3D orientation of the filter we are able to obtain a diversity of selectivities. Top: the S2 units' sinusoid is along x. In other words, it's just like an S1 filter (except that it spans 3 dimensions). Bottom: the S2 units have the sinusoid along θ.*

the small receptive fields in V1 but could be considerably higher for the larger receptive fields in areas such as V4—close to the very center of the fovea. In other words, in V2 and especially V4, some of the larger receptive fields could contain the center of rotation or the focus of expansion. The corresponding wavelets would be a mix of shifts in orientation and non-uniform translations in x, y (circles around the center of rotation) with respect to the previous layer. We expect quite a variety of wavelets – once projected back in image space. This could explain variety of receptive fields seen in Gallant's results [23].

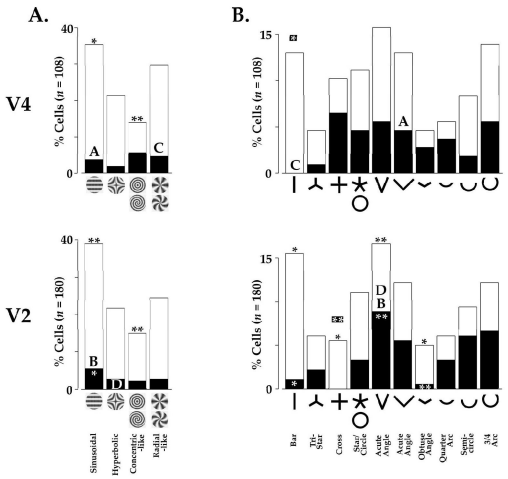


Figure 34: *The effectiveness of the various stimulus subclasses for V4 vs. V2. Each cell from either area was classified according the subclass to which its most effective grating or contour stimulus belonged. The resulting distributions are shown here for grating stimuli (panel A) or contour stimuli (panel B) for both V4 (top row) and V2. See [31].*

5 Part III: Class-specific Transformations and Modularity

Summary. *Part III shows that non-affine 2D image transformations can be well approximated by the template and dot-product module described in Part I and II for certain object classes, provided that the transformations of the templates capture class-specific transformation. The theory explains several properties of face patches in macaque cortex. It also suggests how pooling over transformations can provide identity-specific, pose-invariant representations whereas pooling over identities (templates) provides pose-specific, identity-invariant representations. Part III develops the theory of class-specific invariant recognition to faces; it then describes initial work in other recognition tasks involving bodies and words.*

5.1 Approximate invariance to non-generic transformations

Affine transformations are generic—invariance to them can be learned from any template objects and applied to any test objects. Many other important transformations do not have this property. Non-generic transformations depend on information that is not available in a single image. Perfect invariance to non-generic transformations is not possible. However, approximate invariance can still be achieved as long as the template objects transform similarly to the test objects. One view of this is to say that the missing information in the object's 2D projection is similar between template and test objects. For example, 3D rotation is a non-generic transformation—as a map between projected 2D images depends on the object's 3D structure. If the template and test objects have the same 3D structure then the transformation learned on the template will apply exactly to the test object. If they differ in 3D structure then the error incurred depends on the difference between their 3D structures.

Many non-generic transformations are class-specific. That is, there is a class of objects that are similar enough to one another that good (approximate) invariance can be achieved for new instances of the class by pooling over templates of the same type. Faces are the prototypical example of objects that have many class-specifc transformations. Faces are all similar enough to one another that prior knowledge of how a small set of faces transform can be used to recognize a large number of new faces invariantly to non-generic transformations like 3D rotations or illumination changes. We can extend our notion of a non-generic transformation even further and consider transformations that are difficult to parameterize like facial expressions or aging.

5.2 3D rotation is class-specific

There are many non-generic transformations. As an illustrative example we consider 3D rotation and orthographic projection along the z-axis of 3-space with the center of projection Cp at the origin (see figure 35). In homogenous

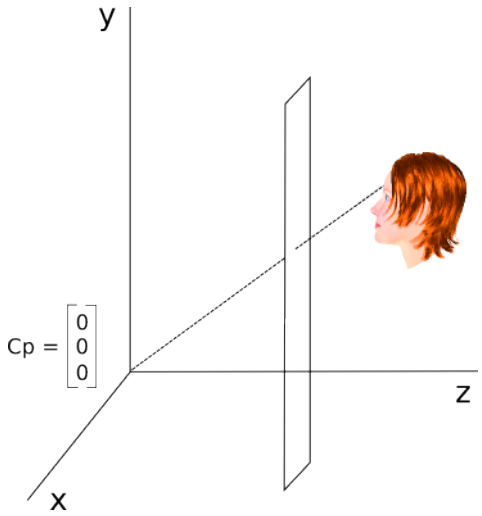


Figure 35: *Consider a situation where the center of projection Cp is at the origin in $\mathbb{R}^3$ and the projection is along the z-axis.*

coordinates this projection is given by

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{50}$$

In 3D homogenous coordinates a rotation around the y-axis is given by

$$R_\theta = \begin{pmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{51}$$

A homogenous 4-vector $X = (x, y, z, 1)^\intercal$ representing a point in 3D is mapped to homogenous 3-vector $\tilde{x} = (x, y, 1)^\intercal$ representing a point on the image plane by $\tilde{x} = PX$. The composition of 3D rotation and orthographic projection is

$$PR_\theta X = \begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \\ 1 \end{pmatrix} \tag{52}$$

Let $t_{\theta,z} : \mathbb{R}^2 \to \mathbb{R}^2$ be the function that describes the 2D transformation of the projection of one point undergoing a 3D rotation. Note: It depends on the z-coordinate which is not available in the 2D image.

$$t_{\theta,z} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \end{pmatrix} \tag{53}$$

Let $\tau = \{(x_\tau^i, y_\tau^i, z_\tau^i, 1)^\intercal\}$ be the set of homogenous 4-vectors representing points on a 3D template object. Likewise, define the test object $f = \{(x^i, y^i, z^i, 1)^\intercal\}$. Assume that the two objects are in correspondence—every point in τ has a corresponding point in f and vice-versa.

Just as in part 1, we use the stored images of the transformations of τ to create a signature that is invariant to transformations of f. However, in this case, the invariance will only be approximate. The transformation of the template object will not generally be the same as the transformation of the test object. That is, $t_{\theta,z_\tau} \neq t_{\theta,z}$ unless $z_\tau = z$.

If $|z - z_\tau| < \epsilon$ is the difference in z-coordinate between two corresponding points of τ and F. The error associated with mapping the point in 2D using t_{θ,z_τ} instead of $t_{\theta,z}$ is given by

$$\|t_{\theta,z}\begin{pmatrix}x\\y\end{pmatrix} - t_{\theta,z_\tau}\begin{pmatrix}x\\y\end{pmatrix}\| < (\epsilon \sin(\theta))^2 \tag{54}$$

5.2.1 The 2D transformation

So far this section has only been concerned with 3D transformations of a single point. We are actually interested in the image induced by projecting a 3D object (a collection of points). We define the *rendering operator* $\mathbb{P}_q[f]$ that takes a set of homogenous points in 3D and a *texture vector* q and returns the image map that puts the corresponding gray value at each projected point.

Definition: Let $f = \{(x^i, y^i, z^i, 1)^\intercal\}$ be a set of N homogenous 4-vectors representing points on a 3D object. Use the notation f^i to indicate the i-th element of f. Let $q \in \mathbb{R}^N$ with $q^i \in [0, 1]$ for $i = 1, ..., N$ be the vector of texture values for each point of f. Let P be the orthographic projection matrix. Define the map $\mathbb{P}_q[f] : \mathbb{R}^2 \to \mathbb{R}$ such that $\forall v \in \mathbb{R}^2$:

$$\mathbb{P}_q[f](v) = \begin{cases} q^i & \text{if } \; v = Pf^i \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

Remark 1: This definition of the rendering function assumes uniform lighting conditions. To address the general case that would allow for variations in gray value over the rendered image arising from the lighting direction this function would also have to depend on the object's material properties as well as other properties of the scene's lighting.

Remark 2: This definition leaves ambiguous the case where more than one point of the object projects to the same point on the image plane (the case where $Pf^i = Pf^j$ for some $i \neq j$). For now we assume that we are only considering objects for which this does not happen. We will have additional comments on the case where self-occlusions are allowed below.

Analogously to the single point case, we can write the 2D transformation $T_{\theta,\vec{z}} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ that maps an image of a 3D object to its image after a 3D rotation. It depends on a vector of parameters $\vec{z} \in \mathbb{R}^N$.

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]] = \mathbb{P}_q[\{R_\theta g^i \ \ : i = 1, ..., N\}] \tag{56}$$

Where g^i is obtained by replacing the z-component of f^i with $\vec{z^i}$. Thus

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]] = \mathbb{P}_q[\{R_\theta f^i \ \ : i = 1, ..., N\}] \qquad \text{if } \vec{z^i} = \text{ the z-component of } f^i \ \ (\forall i) \tag{57}$$

$T_{\theta,\vec{z}}$ transforms individual points in the following way:

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]] \begin{pmatrix} x \\ y \end{pmatrix} \ = \ \mathbb{P}_q[f] \begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \end{pmatrix} \tag{58}$$

We can bound the error arising from mapping the image using $\vec{z_\tau}$ obtained from a template object $\tau = \{(x^i_\tau, y^i_\tau, z^i_\tau, 1)^\intercal\}$—different from the test object f.

If $|z^i_\tau - z^i_f| < \epsilon \ \ (\forall i)$ then

$$\|T_{\theta,\vec{z_\tau}}[\mathbb{P}_q[f]] - \mathbb{P}_q[\{R_\theta f^i \ \ : i = 1, ..., N\}]\| \ < \ \Sigma_{i=1}^N |z^i_\tau \sin(\theta) - z^i_f \sin(\theta)|^2 \ = \ N(\epsilon \sin(\theta))^2 \tag{59}$$

5.2.2 An approximately invariant signature for 3D rotation

We now consider a range of transformations $T_{\theta,\vec{z_\tau}}$ for $\theta \in [-\pi, \pi]$. As in part 1 we define the *template response* (the S-layer response) as the normalized dot product of an image with all the transformations of a template image.

$$\Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[\tau]}(\mathbb{P}_q[f]) = \begin{pmatrix} < T_{-\pi,\vec{z_\tau}}[\mathbb{P}_q[\tau]] \ , \ \mathbb{P}_q[f] > \\ \vdots \\ < T_{\pi,\vec{z_\tau}}[\mathbb{P}_q[\tau]] \ , \ \mathbb{P}_q[f] > \end{pmatrix} \tag{60}$$

In the affine case we have that $\Delta_{G,\tau}(f) = \Delta_{G,f}(\tau)$ up to the ordering of the elements. In that case this fact implies that the signature is invariant.

However, in the case of 3D rotation/projection the template response is defined with respect to the 2D transformation that uses the parameters $\vec{z_\tau}$ obtained from the z-coordinates of τ. Therefore the analogous statement to the invariance lemma of part 1 is false.

In the case of 3D rotation / projection there is only approximate invariance. The closeness of the approximation depends on to which extent the template and test object share the same 3D structure. We believe a statement like the following can be proven:

If for all stored views of the template τ, the difference between the z-coordinate of each point and its corresponding point in the test object f is less than ϵ. That is, if

$$|z^i_\tau - z^i_f| < \epsilon \ \ (\forall i). \tag{61}$$

Then there exists a permutation function S such that

$$S(\Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[\tau]}(\mathbb{P}_q[f])) - \Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[f]}(\mathbb{P}_q[\tau]) < N(\epsilon \sin(\theta))^2 \vec{1} \tag{62}$$

This statement is not mathematically precise (we haven't said how to define the permutation function), but it is the approximate analog of the statement in part I. From this it will follow that we can define an approximately invariant signature. The approximate invariance of the signature defined in this way depends on how similar the 3D structure of the template objects is to the 3D structure of the test object. We will verify this claim empirically in the next section.

Remark: On self-occlusions. Many 3D objects have multiple points that project to the same point on the image plane. These are the places where one part of the object occludes another part e.g. the back of a head is occluded by its front. Since 3D rotation brings different points into view it immediately follows that invariance to 3D rotation from a single 2D example image can never be perfect. Consider: It is never possible to predict a tattoo on someone's left cheek from a view of the right profile. On the other hand, this does not necessarily impact the approximate invariance obtained from templates aquired from similar objects. For example, a lot can be said about the likely appearance of the back of someone's head from a view of the front—e.g. the hair and skin color remain the same. This makes it difficult to precisely formulate an approximate version of the invariance lemma (except for the unrealistic case of objects with no self-occlusions).

5.3 Empirical results on class-specific transformations

Class-specific transformations, like 3D rotation, can be learned from one or more exemplars of an object class and applied to other objects in the class. For this to work, the object class needs to consist of objects with similar 3D shape and material properties. Faces, as a class, are consistent enough in both 3D structure and material properties for this to work. Other, more diverse classes, such as "automobiles" are not.

Figure 36 depicts an extension of the HMAX model that we used to empirically test this method of building signatures that are approximately invariant to non-affine transformations. The signature at the top of the usual HMAX model (C2 in this case) is not invariant to rotation in depth. However, an additional layer (S3 and C3) can store a set of class-specific template transformations and provide class-specific approximate invariance (see Figures 37 and 38).

Figures 37 and 38 show the performance of the extended HMAX model on viewpoint-invariant and illumination-invariant within-category identification tasks. Both of these are one-shot learning tasks. That is, a single view of a target object is encoded and a simple classifier (nearest neighbors) must rank test images depicting the same object as being more similar to the encoded target than to images of any other objects. Both targets and distractors were presented under varying viewpoints and illuminations. This task models the common situation of encountering a new face or object at one viewpoint and then being asked to recognize it again later from a different viewpoint.

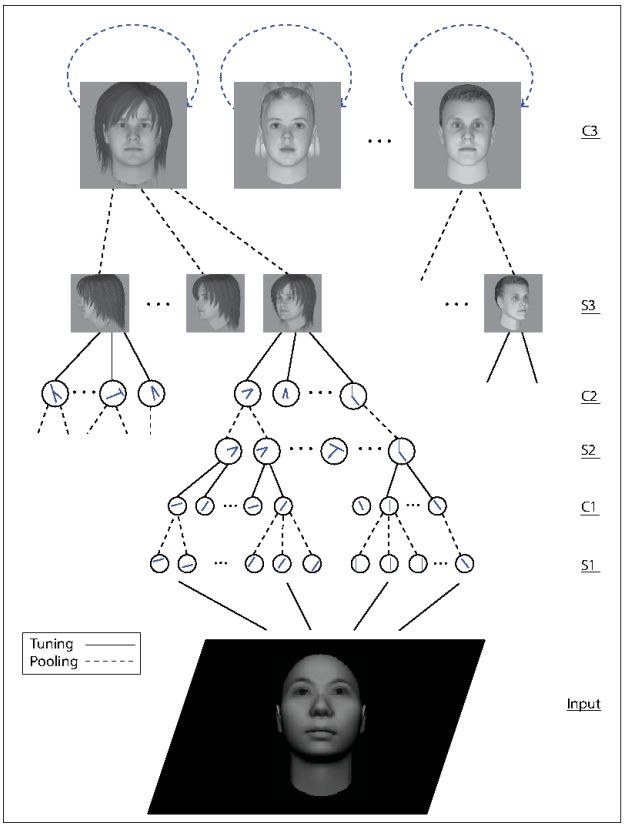


Figure 36: *Illustration of an extension to the HMAX model to incorporate class-specific invariance to face viewpoint changes. Note: All simulations with this model (Figures 37, 38) use a Gaussian radial basis function to compute the S2 and S3 layers as opposed to the normalized dot product that is used in its S1 layer and elsewhere in this report.*

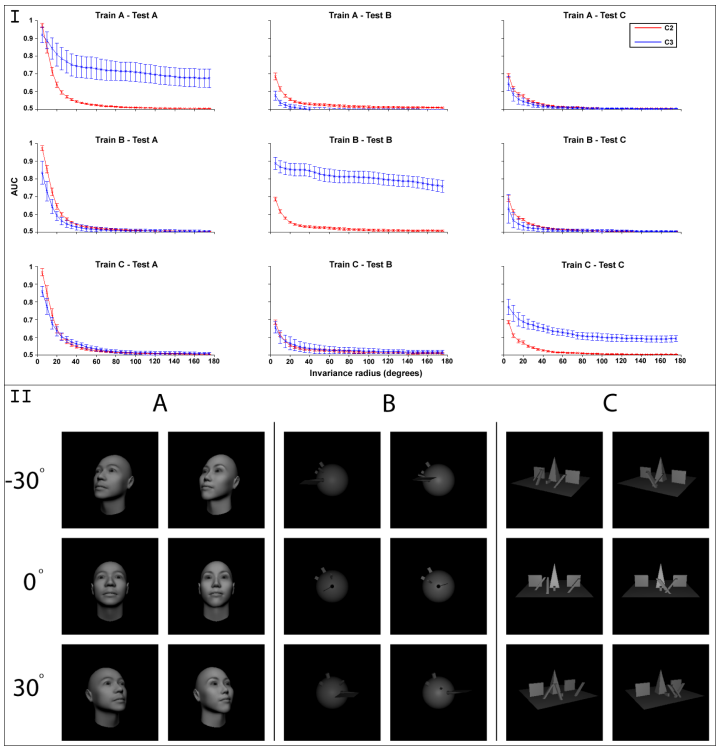


Figure 37: *Viewpoint invariance. Bottom panel (II): Example images from three classes of stimuli. Class A consists of faces produced using FaceGen (Singular Inversions). Class B is a set of synthetic objects produced using Blender (Stichting Blender Foundation). Each object in this class has a central spike protruding from a sphere and two bumps always in the same location on top of the sphere. Individual objects differ from one another by the direction in which another protrusion comes off of the central spike and the location/direction of an additional protrusion. Class C is another set of synthetic objects produced using Blender. Each object in this class has a central pyramid on a flat plane and two walls on either side. Individual objects differ in the location and slant of three additional bumps. For both faces and the synthetic classes, there is very little information to disambiguate individuals from views of the backs of the objects. Top panel (I): Each column shows the results of testing the model's viewpoint-invariant recognition performance on a different class of stimuli (A,B or C). The S3/C3 templates were obtained from objects in class A in the top row, class B in the middle row and class C in the bottom row. The abscissa of each plot shows the maximum invariance range (maximum deviation from the frontal view in either direction) over which targets and distractors were presented. The ordinate shows the AUC obtained for the task of recognizing an individual novel object despite changes in viewpoint. The model was never tested using the same images that were used to produce S3/C3 templates. A simple correlation-based nearest-neighbor classifier must rank all images of the same object at different viewpoints as being more similar to the frontal view than other objects. The red curves show the resulting AUC when the input to the classifier consists of C2 responses and the blue curves show the AUC obtained when the classifier's input is the C3 responses only. Simulation details: These simulations used 2000 translation and scaling invariant C2 units tuned to patches of natural images. The choice of natural image patches for S2/C2 templates had very little effect on the final results. Error bars (+/- one standard deviation) show the results of cross validation by randomly choosing a set of example images to use for producing S3/C3 templates and testing on the rest of the images. The above simulations used 710 S3 units (10 exemplar objects and 71 views) and 10 C3 units.*

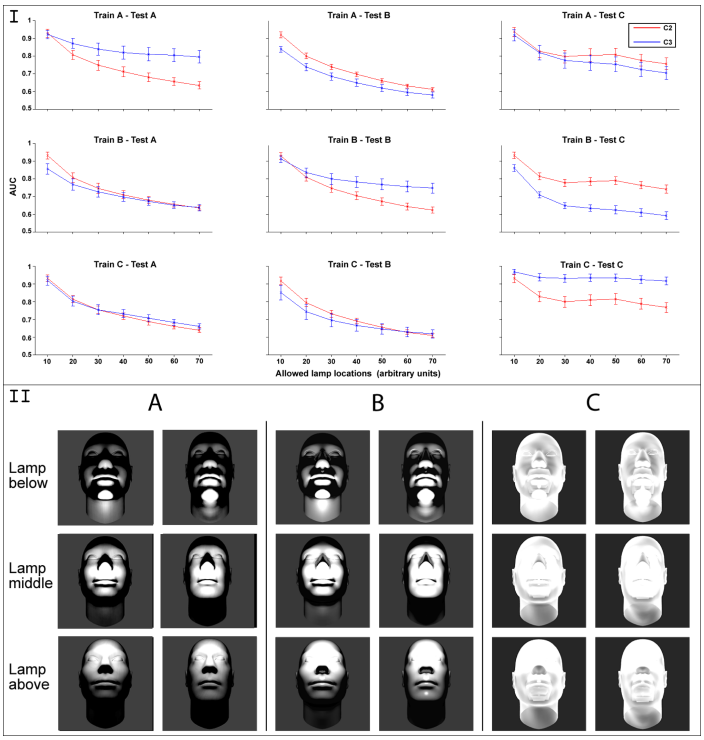


Figure 38: *Illumination invariance. Same organization as in figure 3. Bottom panel (II): Example images from three classes of stimuli. Each class consists of faces with different light reflectance properties, modeling different materials. Class A was opaque and non-reflective like wood. Class B was opaque but highly reflective like a shiny metal. Class C was translucent like glass. Each image shows a face's appearance corresponding to a different location of the source of illumination (the lamp). The face models were produced using FaceGen and modified with Blender. Top panel (I): Columns show the results of testing illumination-invariant recognition performance on class A (left), B (middle) and C (right). S3/C3 templates were obtained from objects in class A (top row), B (middle row), and C (bottom row). The model was never tested using the same images that were used to produce S3/C3 templates. As in figure 3, the abscissa of each plot shows the maximum invariance range (maximum distance the light could move in either direction away from a neutral position where the lamp is even with the middle of the head) over which targets and distractors were presented. The ordinate shows the AUC obtained for the task of recognizing an individual novel object despite changes in illumination. A correlation-based nearest-neighbor "classifier" must rank all images of the same object under each illumination condition as being more similar to the neutral view than other objects. The red curves show the resulting AUC when the input to the classifier consists of C2 responses and the blue curves show the AUC obtained when the classifier's input is the C3 responses only. Simulation details: These simulations used 80 translation and scaling invariant C2 units tuned to patches of natural images. The choice of natural image patches for S2/C2 templates had very little effect on the final results. Error bars (+/- one standard deviation) show the results of cross validation by randomly choosing a set of example images to use for producing S3/C3 templates and testing on the rest of the images. The above simulations used 1200 S3 units (80 exemplar faces and 15 illumination conditions) and 80 C3 units.*

The original HMAX model [92], represented here by the red curves (C2), shows a rapid decline in performance due to changes in viewpoint and illumination. In contrast, the C3 features of the extended HMAX model perform significantly better than C2. Additionally, the performance of the C3 features is not strongly affected by viewpoint and illumination changes (see the plots along the diagonal in Figures 37I and 38I).

The C3 features are class-specific. Good performance on within-category identification is obtained using templates derived from the same category (plots along the diagonal in figures 37I and 38I). When C3 features from the wrong category are used in this way, performance suffers (off-diagonal plots). In all these cases, the C2 features which encode nothing specifically useful for taking into account the relevant transformation perform as well as or better than C3 features derived from objects of the wrong class. It follows that in order to accomplish within-category identification, then the brain must separate the circuitry that produces invariance for the transformations that objects of one class undergo from the circuitry producing invariance to the transformations that other classes undergo.

Object classes that are important enough to require invariance to non-generic transformations of novel exemplars must be encoded by dedicated circuitry. Faces are clearly a sufficiently important category of objects to warrant this dedication of resources. Analogous arguments apply to a few other categories; human bodies all have a similar 3D structure and also need to be seen and recognized under a variety of viewpoint and illumination conditions, likewise, reading is an important enough activity that it makes sense to encode the visual transformations that words and letters undergo with dedicated circuitry (changes in font, viewing angle, etc). We do not think it is coincidental that, just as for faces, brain areas which are thought to be specialized for visual processing of the human body (the extrastriate body area [15]) and reading (the visual word form area [9, 5]) are consistently found in human fMRI experiments (See section 5.5).

5.4 The macaque face-processing network

In macaques, there are 6 discrete face-selective regions in the ventral visual pathway, one posterior lateral face patch (PL), two middle face patches (lateral- ML and fundus- MF), and three anterior face patches, the anterior fundus (AF), anterior lateral (AL), and anterior medial (AM) patches [100, 101]. At least some of these patches are organized into a feedforward hierarchy. Visual stimulation evokes a change in the local field potential ~ 20 ms earlier in ML/MF than in patch AM [21]. Consistent with a hierarchical organization involving information passing from ML/MF to AM via AL, electrical stimulation of ML elicited a response in AL and stimulation in AL elicited a response in AM [65]. In addition, spatial position invariance increases from ML/MF to AL, and increases further to AM [21] as expected for a feedforward processing hierarchy.

Freiwald et al. (2010) found that the macaque face patches differ qualitatively in how they represent identity across head orientations. Neurons in the

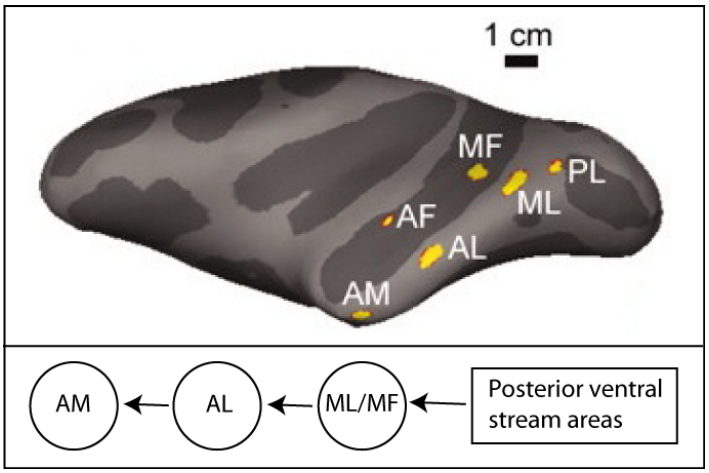


Figure 39: *Layout of face-selective regions in macaque visual cortex, adapted from [21] with permission.*

middle lateral (ML) and middle fundus (MF) patches were view-specific; while neurons in the most anterior ventral stream face patch, the anterior medial patch (AM), were view invariant. Puzzlingly, neurons in an intermediate area, the anterior lateral patch (AL), were tuned identically across mirror-symmetric views. That is, neurons in patch AL typically have bimodal tuning curves e.g., one might be optimally tuned to a face rotated $45°$ to the left and $45°$ to the right[6] (see figure 40).

In Part II of this paper, we argued that Hebbian plasticity at the synapses in visual cortex causes the tuning of the cells to converge to the eigenvectors of their input's covariance. In this section we demonstrate that the same theory, when applied to class-specific layers, yields cells with properties that closely resemble those of the cells in the macaque face-processing network.

Suppose that AL receives neural representations of face images in different poses during a developmental stage. This may require a neural "gate" , possibly in the posterior lateral face patch (PL), that is "on" only for face-like images. If the synapses onto neurons in patch AL are updated by Oja's rule then they will converge to the eigenvectors of the covariance matrix of their inputs. In this sense, the AL neurons receiving inputs from ML/MF are analogous to simple cells in V1.

[6]Freiwald and Tsao (2010) found that 92 of the 215 AL cells in their study responded at least twice as strongly to one of the two full-profiles as to frontal faces. These profile-selective cells responded very similarly to both profiles. A subsequent test using face stimuli at more orientations found that 43 of 57 cells had view tuning maps with two discrete peaks at mirror symmetric positions.

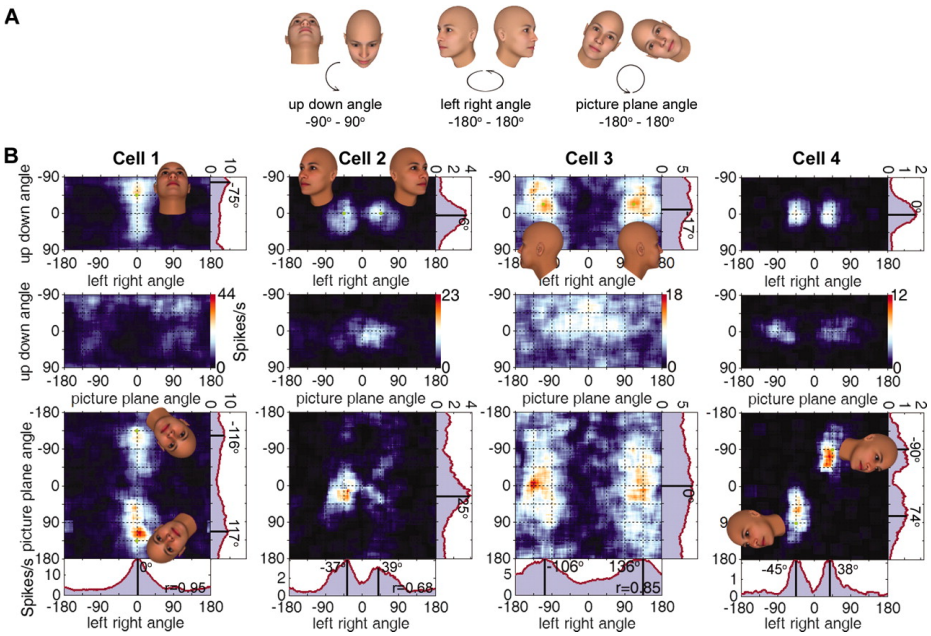


Figure 40: *Figure from [21]. Tuning of AL cells to a head randomly rotated in three dimensions. (A) Illustration of stimulus head and three axes of rotation. (B) View tuning in four typical cells. Tuning to up-down angle versus left-right angle (responses averaged across picture-plane angle). (Middle) Tuning to up-down angle versus picture-plane angle (responses averaged across left-right angle). (Bottom) Tuning to picture-plane angle versus left-right angle (responses averaged across up-down angle). Marginal tuning curves are also shown (vertical lines indicate tuning peak position).*

5.4.1 Principal components and mirror-symmetric tuning curves

Define $\tau^*_{n,i}$ as the i-th principal component (PC) of the templatebook obtained from a single base template. For the following, assume that the templatebook $\mathbb{T}$ is centered (we subtract its mean as a preprocessing step). The $\tau^*_{n,i}$ are by definition the eigenvectors of $\mathbb{T}^\intercal\mathbb{T}$: $\tau^*_{n,1}$ is the first PC acquired from the n-th base pattern's transformation, $\tau^*_{n,2}$ the second PC, and so on.

A frontal view of a face is symmetric about its vertical midline. Thus equal rotations in depth (e.g., 45° to the left and 45° to the right) produce images that are reflections of one another. Therefore, the templatebook $\mathbb{T}$ obtained from a face's 3D rotation in depth must have a special structure. For simplicity, consider only "symmetric transformation sequences", e.g., all the neural frames of the rotation from a left 90° profile to a right 90° profile. For each neural frame $\tau_{n,t}$ there must be a corresponding reflected frame in the templatebook that we will indicate as $\tau_{n,-t}$. It will turn out that as a consequence of its having this structure, the eigenfunctions of the templatebook will be even and odd. Therefore, the templates obtained from compressing the templatebook as though they were neural frames, are symmetric or anti-symmetric images (see figure 43).

Properties of the spectrum of the covariance of faces and their reflections

Let $R : \mathbb{R}^d \to \mathbb{R}^d$ denote the reflection operator. It is an involution i.e. $R^2(x) = x \ \forall x$. For simplicity, consider a cell that has been exposed to just one template τ and its reflection $R\tau$.

Let $Z = (\tau, R\tau)$ denote the matrix of transformed images (the templatebook), in this case consisting just of τ and its reflection. In part 1 we also defined the operators $\Delta_Z(x) = Z^\intercal x$—called the template response—and $\Sigma_Z = P(\Delta_Z(x))$—the signature with aggregation function P.

We are interested in the eigenfunctions of the covariance operator $ZZ^\intercal$. Note: $\Delta_{ZZ^\intercal}$ acts on images by right-multiplying with its transpose, but this is irrelevant since $ZZ^\intercal$ is symmetric ($\Delta_{ZZ^\intercal}$ is self-adjoint). We show that $ZZ^\intercal$ commutes with the reflection operator R. Then, applying the fact that whenever two operators commute they must have the same eigenfunctions, we show that $ZZ^\intercal$'s eigenfunctions must be even or odd functions.

We want to show that $RZZ^\intercal = ZZ^\intercal R$.

Notice that we can write the covariance operator as the sum of outer products of Z's columns. $ZZ^\intercal = \tau\tau^\intercal + (R\tau)(R\tau)^\intercal$. Thus:

$$\begin{aligned} RZZ^\intercal &= R\tau\tau^\intercal + RR\tau(R\tau)^\intercal \\ &= R\tau\tau^\intercal + \tau\tau^\intercal R \end{aligned}$$

and

$$\begin{aligned} ZZ^\intercal R &= \tau\tau^\intercal R + R\tau(R\tau)^\intercal R \\ &= \tau\tau^\intercal R + R\tau\tau^\intercal \end{aligned}$$

Therefore $RZZ^\intercal = ZZ^\intercal R$, the covariance operator commutes with the reflection operator. Thus they must have the same eigenfunctions. Since the eigenfunctions of R are even and odd, the eigenfunctions of $ZZ^\intercal$ (and of $\Delta_{ZZ^\intercal}$) must also be even and odd.

5.4.2 Models of the macaque face recognition hierarchy

We have shown that models of the ventral stream that compute a signature relative to the principal components of the templatebooks acquired from rotation of template faces must contain an intermediate step with identical tuning to symmetric face faces. We propose to identify patch AL with the the projection onto principal components and patch AM with the subsequent pooling stage.

These considerations alone do not completely constrain a model of the ventral stream. In order to demonstrate the working of these models and perform virtual electrophysiology experiments to test the properties of the simulated cells, we must make some other parameter and architectural choices. We investigated several model architectures. Each one corresponds to different choices we made about the processing of the visual signal prior to face patch AL (see figure 41).

At run time, cells in the S-PCA layer compute the absolute value of the normalized dot product of their stored PC with the input. Each cell in the C-PCA layer pools over all the cells in the S-PCA layer with PCs from the same templatebook.

In the developmental phase, the S-PCA templates are acquired by PCA of the templatebooks. Each templatebook contains all the (vectorized) images of the rotation (in depth) of a single face. All the 3D models used to produce training and testing images were produced by FaceGen[7] and rendered with Blender[8]. Images of each face were rendered every 5 degrees, Each templatebook covered nearly the full range of orientations ($0 - 355^\circ$).

Each experiment used 20 faces (templatebooks) in the developmental phase, and 20 faces for testing. These training and testing sets were always independent. No faces that appeared in the developmental phase ever appeared in the testing phase.

Figure 44 compares three of these models to two different layers of the HMAX model on a viewpoint-invariant face identification task. The proposed model is considerably better able to recognize new views of a face despite viewpoint changes. The results shown here use all the principal components of each templatebook. In analogous simulations we showed that roughly the same level of performance is achieved when only the first 5-10 PCs are used.

5.5 Other class-specific transformations: bodies and words

Many objects besides faces are nice in the sense that they have class-specific transformations. Within the ventral stream there are also patches of cortex that

[7] Singular Inversions Inc.
[8] The Blender foundation

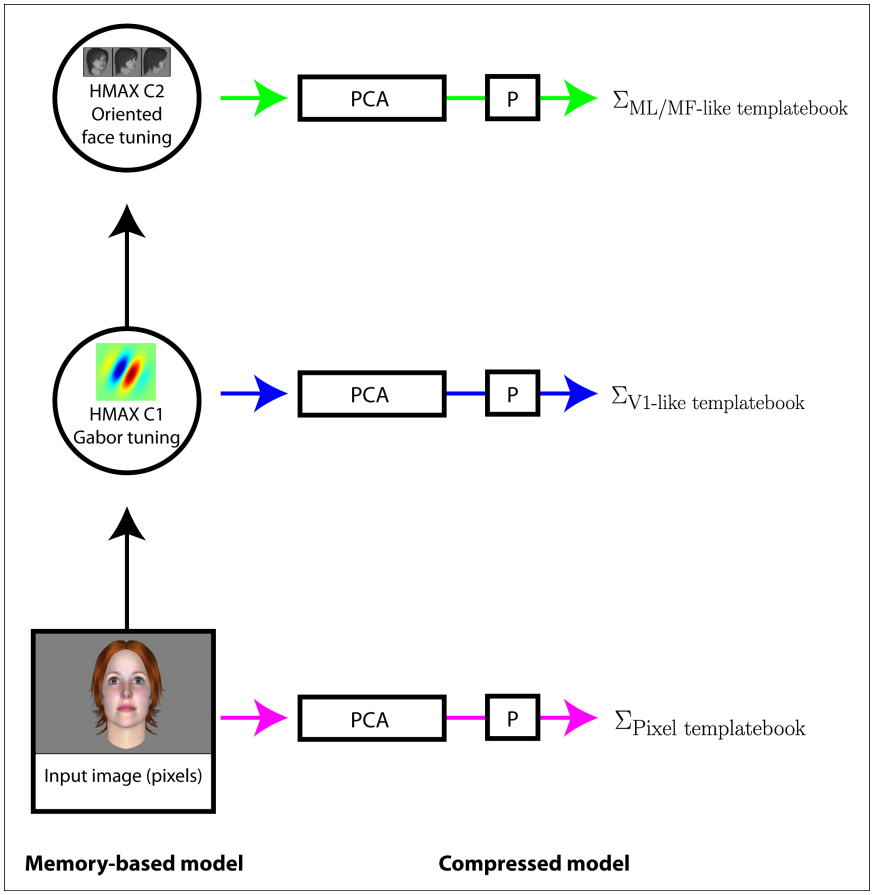


Figure 41: *Schematic of three possible models. Magenta: A model where the templatebooks were raw pixels with no preprocessing. Blue: A model where the templatebooks were encoded in an HMAX C1 layer (preprocessing with Gabor filtering and some limited pooling over position). Green: A model where the templatebooks are encoded in the responses of an HMAX C2 layer with large—nearly global—receptive fields and optimal tuning to specific views of faces.*

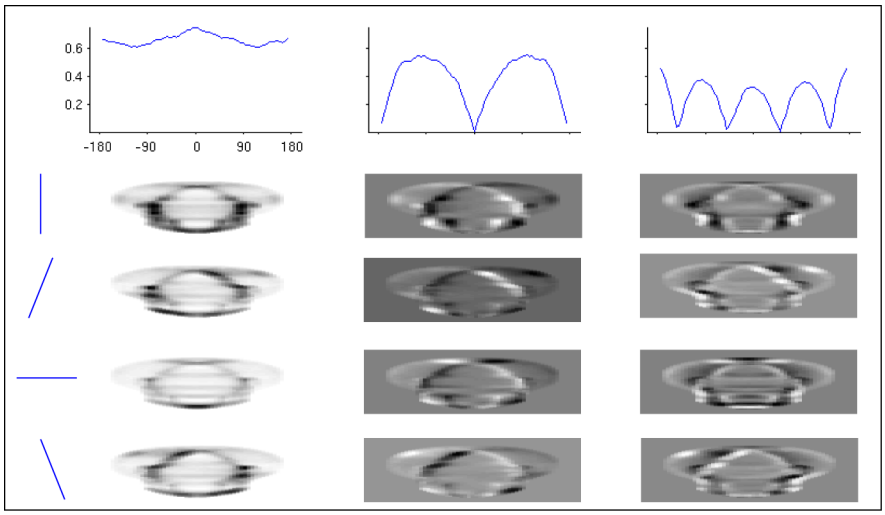


Figure 42: *Sample tuning curves and principal components for the model that encodes all inputs as HMAX C1 responses (the blue model in figures 41 and 44). Top row: the responses of S-PCA layer cells to systematically varying the orientation of a randomly-chosen test face. Below each tuning curve are 4 "slices" from the PC encoded by that cell. There are 4 slices corresponding to each of the 4 orientations we used in the C1 layer (orientations shown in far left column). The first and third PCs are clearly symmetric (even functions) while the second is anti-symmetric (an odd function). These 3 PCs all came from the same templatebook (other templatebooks give very similar results). They are ordered by their corresponding eigenvalue with the largest eigenvalue on the left.*

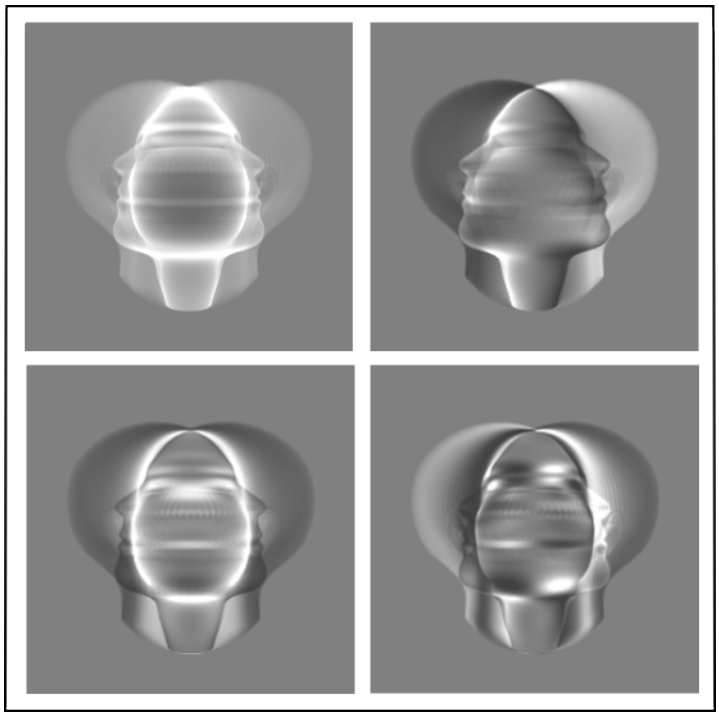


Figure 43: *More sample principal components. These were obtained from a model that does PCA directly on pixel inputs. They are the first 4 PCs obtained from the rotation of one head from* -90° *to* 90°.

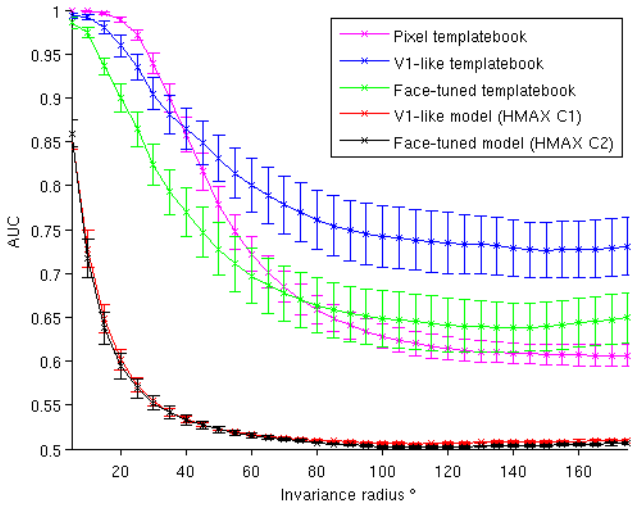


Figure 44: *Results from a test of viewpoint-invariant face identification. Test faces were presented on a black background. The task was to correctly categorize images by whether or not they depict the same person shown in a reference image—despite changes in viewpoint. This is a test of generalization from a single example view. The abscissa shows the maximum invariance range (maximum deviation from the frontal view in either direction) over which targets and distractors were presented. The ordinate shows the area under the ROC curve (AUC) obtained for the task of recognizing an individual despite changes in viewpoint (nearest neighbor classifier). The model was never tested with any of the images that went into the templatebooks in the developmental phase. We averaged the AUC obtained from experiments on the same model using all* 20 *different reference images and repeated the entire simulation (including the developmental phase)* 10 *times with different training/test splits (for cross validation). The error bars shown on this figure are* 1 *standard deviation, over cross validation splits. Magenta, blue and green curves: results from the models that encoded templatebooks and inputs as raw pixels, HMAX* $C1$ *responses, HMAX* $C2$ *(tuned to faces at different views) respectively. These are the same models depicted in Figure 41. Red and black curves: Performance of the HMAX* $C1$ *and HMAX C2 layers on this task (included for comparison).*

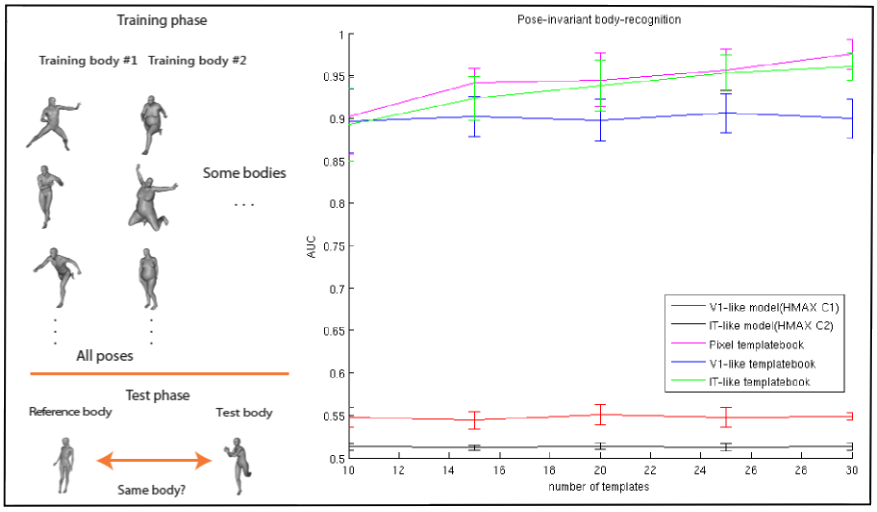


Figure 45: *Left: Images of human bodies in various poses were used to train and test the model. 1280 3D object models of human body were created with DAZ 3D Studio and one 256*256 pixel greyscale image was rendered from each object automatically with blender. The 1280 objects consisted of 40 differently shaped human bodies in 32 poses. The 40 bodies were either male or female, had varying degrees of fatness, muscularity, and limb proportion. The 32 poses were natural, commonly encountered poses such as waving, running, leaning, and clinging. Right: Performance of class-specific models and HMAX control models on a pose-invariant body recognition task. 10 bodies were used for testing. The abscissa is the number of bodies used to train the model. Performance was averaged over 10 cross-validation runs. The error bars correspond to standard deviations of AUC values over the cross-validation runs.*

show BOLD responses for non-face objects. These include regions that respond to scenes—the parahippocampal place area (PPA) [16]—written words—the visual word form area (VWFA) [9], and bodies—the extrastriate body area (EBA) and the fusiform body area (FBA) [15, 72]. Many of these regions were shown to be necessary for recognition tasks with the objects they process by lesion studies ([53, 66]) and TMS ([104, 74]. We have begun to study transformations of two of these: bodies (different poses, actions) and printed words (changes in font, viewing angle, etc.) (See also the preliminary report of our work on scenes: [43]).

Figures 45 and 46 show the results of class-specific invariant recognition tasks for bodies—identification of a specific body invariantly to its pose—and words—font-invariant word recognition. In both cases, the models that employ class-specific features (they pooling over templates depicting different bodies or different fonts) outperform control HMAX models. Additional details on these models will soon be available in forthcoming reports from our group.

Remark: Throughout this report we have held temporal contiguity to be

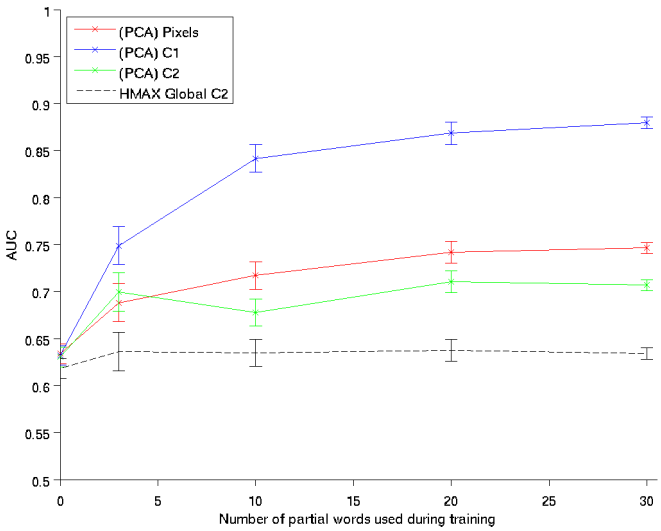


Figure 46: *Words (4-grams) were chosen from a fixed alphabet of 4 letters. A nearest-neighbor classifier ranked each image—of a word in a particular font—by its similarity to the image of a reference word. Templatebooks were obtained from translated and font-transformed images of single letters, bigrams and trigrams. Red, blue and green curves: These used a version of the compression-based model described in part II of this report. Black curve: An HMAX C2 model with global pooling (for comparison). The S2 dictionary consisted of 2000 patches of natural images. The abscissa is the number of partial words (bigrams and trigrams) used in the templatebook. Error bars are +/- 1 standard deviation, over 5 runs of the simulation using different randomly chosen bigrams, trigrams and testing words. This simulation used 4 different fonts.*

an important cue for associating the frames of the video of an object's transformation with one another. That approach cannot be taken to learn these body/word recognition models. The former model requires the association of different bodies under the same pose and the latter requires the same words (rather: partial words) to be associated under a variety of fonts. A temporal-contiguity based learning rule could not be used to learn the pooling domains for these tasks. Additionally, in other sensory modalities (such as audition) recognizing temporally extended events is common. It is not clear how temporal contiguity-based arguments could apply in those situations.

5.6 Invariance to X and estimation of X

So far we have discussed the problem of recognition as estimating identity or category invariantly to a transformation X – such as translation or pose or illumination. Often however, the key problem is the complementary one, of estimating X, for instance pose, possibly independently of identity. The same neural population may be able to support both computations as shown in IT recordings [37] and model simulations [90]. We are certainly able to estimate position, rotation, illumination of an object without eye movements, though probably not very precisely. In the ventral stream this may require the use of lower-level signatures, possibly in a task-dependent way. This may involve attention.

Figure 47 shows the results on the task of recognizing the pose—out of a set of 32 possibilities—of a body invariantly to which body is shown. Notice that low-level visual features (HMAX C1) work just as well on this task as the class-specific features.

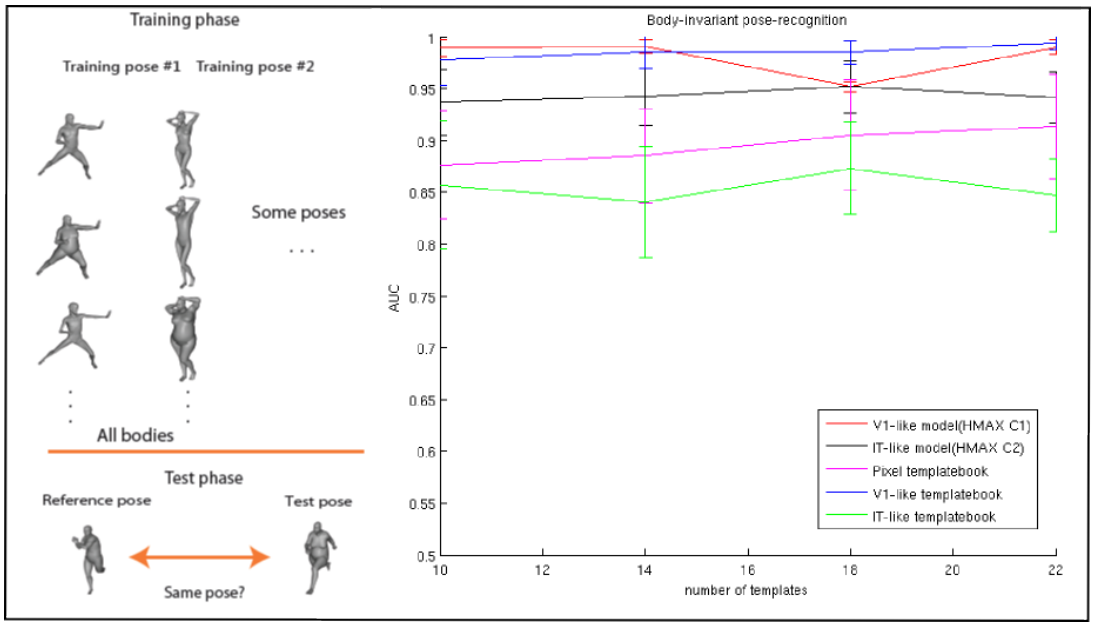


Figure 47: *Left: These simulations used the same images as the one in figure 45. Right: Performance of class-specific models and HMAX control models on a body-invariant pose recognition task. 10 poses were used for testing. The abscissa is the number of poses used to train the model. Performance was averaged over 10 cross-validation runs. The error bars correspond to standard deviations of AUC values over the cross-validation runs.*

6 Discussion

This section gives first an overview of the various parts of the theory and then summarizes some of its main ideas. We also discuss the new theory with respect to the old model, list potential problems and weaknesses and finally discuss directions for future research.

- *Part I presents a theory in which transformations are learned during development by storing a number of templates and their transformations. Invariant signatures can be obtained by pooling dot products of a new image with the transformed templates over the transformations for each template. A hierarchical architecture of these operations provides global invariance and stability to local deformations.*
- *Part II assumes that the storing of templates during biological development is based on Hebbian synapses effectively computing the eigenvectors of the covariance of the transformed templates. A cortical equation is derived which predicts the tuning of simple cells in V1 in terms of Gabor-like wavelets. The predictions agree with physiology data across different species. Instead of pooling a template across its transformations, the system pools nonlinear functions, such as modulo, of eigenfunctions. Furthermore, we show that the V1 representation diagonalizes the local representation (within balls of radius r with $\frac{r}{R} \leq k$) of similitude transformations of the image as independent shifts in a 4D space. Thus learning at higher layers generates 4D wavelets. The prediction may be consistent with physiology data*
- *Part III shows that non-affine transformations on the image plane (such as the image changes induced by 3D rotations of an object) can be well approximated by the template and dot-product module described in Part I and II for certain object classes, provided that the transformed templates capture class-specific transformation. The theory explains several properties of faces patches in macaque cortex. It also suggests how pooling over transformations can provide identity-specific, pose-invariant representations whereas pooling over identities (templates) provides pose-specific, identity-invariant representations.*

6.1 Some of the main ideas

There are several key ideas in the theoretical framework of the paper. We recount here ideas already mentioned in the paper.

1. We conjecture that the sample complexity of object recognition is mostly due to geometric image transformations (e.g. different viewpoints) and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations.
2. The most surprising implication of the theory emerging from these specific assumptions is that the computational goals and detailed properties

of cells in the ventral stream follow from *symmetry properties* of the visual world through a process of correlational learning. The obvious analogy is physics: for instance, the main equation of classical mechanics can be derived from general invariance principles. In fact one may argue that a Foldiak-type rule together with the physics of the world is all that is needed by evolution to determine through developmental learning the hierarchical organization of the ventral stream, the transformations that are learned and the tuning of the receptive fields in each visual area.

3. Aggregation functions such as the modulo square or approximations of it or the max (as in HMAX or in [47]) ensure that signatures of images are invariant to affine transformations of the image and that this property is preserved from layer to layer.

4. The theory assumes that there is a hierarchical organization of areas of the ventral stream with increasingly larger receptive apertures of increasing size determining a stratification of the range of invariances. At the smallest size there are effectively only translations.

5. Memory-based invariances are related to the spectral properties of transformed templates recorded by a memory-based recognition architecture such as an (extended) HMAX.

6. Spectral properties of the input determine receptive field tuning via Hebbian-like online learning rules that converge to the principal components of the inputs.

7. Signatures from all layers access the associative memory or classifier module and thus control iterations in visual recognition and processing. Of course, at lower layers there are many signatures, each one in different complex cell layer locations, while at the top layer there are only a small number of signatures – in the limit only one.

The theory of this paper starts with this central computational problem in object recognition: identifying or categorizing an object after looking at a single example of it – or of an exemplar of its class. To paraphrase Stu Geman, the difficulty in understanding how biological organisms learn – in this case how they recognize – is not the usual $n \to \infty$ but $n \to 0$. The mathematical framework is inspired by known properties of neurons and visual cortex and deals with the problem of how to learn and discount invariances. Motivated by the Johnson-Lindenstrauss theorem, we introduce the notion of a *signature* of an object as a set of similarity measurements with respect to a small set of template images. An *invariance lemma* shows that the stored transformations of the templates allow the retrieval of an invariant signature of an object for any uniform transformation of it such as an affine transformation in 2D. Since any transformation of an image can be approximated by local affine transformations, corresponding to a set of local receptive fields, the invariance lemma provides a solution for the problem of recognizing an object after experience

with a single image – under conditions that are idealized but hopefully capture a good approximation of reality. Memory-based hierarchical architectures are much better at learning transformations than non-hierarchical architectures in terms of memory requirements. This part of the theory shows how the hierarchical architecture of the ventral stream with receptive fields of increasing size (roughly by a factor of 2 from V1 to V2 and again from V2 to V4 and from V4 to IT) could implicitly learn during development different types of transformations starting with local translations in V1 to a mix of translations and scales and rotations in V2 and V4 up to more global transformations in PIT and AIT (the *stratification conjecture*).

Section 4 speculates on how the properties of the specific areas may be determined by visual experience and continuous plasticity and characterizes the spectral structure of the templatebooks arising from various types of transformations that can be learned from images. A conjecture – to be verified with simulations and other empirical studies – is that in such an architecture the properties of the receptive fields in each area are mostly determined by the underlying transformations rather than the statistics of natural images. The last section puts together the previous results into a detailed hypothesis of the plasticity, the circuits and the biophysical mechanisms that may subserve the computations in the ventral stream.

In summary, some of the broad predictions of this theory-in-fieri are:

- each cell's tuning properties are shaped by visual experience of image transformations during developmental and adult plasticity;
- the mix of transformations – seen from the retina – learned in each area influences the tuning properties of the cells – oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (e.g. face tuned cells);
- during evolution, areas above V1 should appear later than V1, reflecting increasing object categorization abilities and the need for invariances beyond translation;
- an architecture based on signatures that are invariant (from an area at some level) to affine transformations may underly *perceptual constancy* against small eye movements and other small motions[9].
- invariance to affine transformations (and others) can provide the seed for evolutionary development of "conceptual" invariances;
- the *transfer of invariance* accomplished by the machinery of the templatebooks may be used to implement high level abstractions;

[9]There may be physiological evidence (from Motter and Poggio) suggesting invariance of several minutes of arc at the level of V1 and above.

- the preceding sections stressed that the statistics of natural images do not play a primary role in determining the spectral properties of the templatebook and, via the *linking theorem* the tuning of the cells in specific areas. This is usually true for the early areas under normal development conditions. It is certainly not true if development takes place in a deprived situation. The equations show that the spectrum of the images averaged over the presentations affects the spectral content, e.g. the correlation matrix and thus the stationary solutions of Hebbian learning.

- In summary, from the assumption of a hierarchy of areas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation represented in an area determines the tuning of the neurons in the area; and that class-specific transformations are learned and represented at the top of the hierarchy.

6.2 Extended model and previous model

So far in this paper, existing hierarchical models of visual cortex – eg HMAX – are reinterpreted and extended in terms of computational architectures which evolved to discount image transformations learned from experience. From this new perspective, I argue that a main goal of cortex is to learn equivalence classes consisting of patches of images (that we call templates), associated together since they are observed in close temporal contiguity – in fact as a temporal sequence – and are therefore likely to represent physical transformations of the same object (or part of the same object). I also conjecture that the hierarchy – and the number of layers in it - is then determined by the need to learn a group of transformations – such as the affine group. I prove that a simple memory-based architecture can learn invariances from the visual environment and can provide invariant codes to higher memory areas. I also discuss the possibility that the size of the receptive fields determines the type of transformations which are learned by different areas of cortex from the natural visual world – from local translations to local rotations and image-plane affine transformations up to almost global translations and viewpoint/pose/expression transformations. Earlier layers would mostly represent local generic transformations such as translation and scale and other similitude transformations. Similar considerations imply that the highest layers may represent class-specific transformations such as rotations in depth of faces or changes in pose of bodies.

- The present HMAX model has been hardwired to deal with 2 generic transformations: translation and scale. The model performance on "pure" translation tasks is perfect (apart from discretization noise), while it declines quickly with viewpoint changes (± 20 degrees is roughly the invariance range).

- As mentioned several times, the theory assumes that signatures from several layers can be used by the associative memory- classifier at the top, possibly under attentional or top-down control, perhaps via cortical-pulvinar-cortical connections.
- What matters for recognition is not the strong response of a population of neurons (representing a signature) but the invariance of the response in order to provide a signal, invariant as possible, to the classifier.
- *Untangling invariance* Getting invariance is easy if many examples of the specific object are available. What is difficult is getting invariance from a single example of an object (or very few). Many of the discussions of invariance are confused by failing to recognize this fact. Untangling invariance is easy[10] when a sufficiently large number of previously seen views of the object are available, by using smooth nonlinear interpolation techniques such as RBFs.

6.3 What is under the carpet

Here is a list of potential weaknesses, small and large, with some comments:

- "The theory is too nice to be true". One of the main problems of the theory is that it seems much too elegant – in the sense of physics – for biology.
- Backprojections are not taken into account and they are a very obvious feature of the anatomy, which any real theory should explain. Backprojections and top-down controls are however implied by the present theory. The most obvious limitation of feedforward architectures is recognition in clutter and the most obvious way around the problem is the attentional masking of large parts of the image under top-down control. More in general, a realistic implementation of the present theory requires top-down control signals and circuits, supervising learning and possibly fetching signatures from different areas and at different locations in a task-dependent way. An even more interesting hypothesis is that backprojections update local signatures at lower levels depending on the scene class currently detected at the top (an operation similar to the top-down pass of Ullman). In summary, the output of the feedforward pass is used to retrieve labels and routines associated with the image; backprojections implement an attentional focus of processing to reduce clutter effects and also run spatial visual routines at various levels of the hierarchy.

[10] apart from self-occlusions and uniqueness problems. Orthographic projections in 2D of the group $Aff(3, \mathbb{R})$ are not a group; however the orthographic projections of translations in x, y, z and rotations in the image plane are a group.

- Subcortical projections, such as, for instance, projections to and from the pulvinar, are not predicted by the theory. The present theory still is (unfortunately) in the "cortical chauvinism" camp. Hopefully somebody will rescue it.

- Cortical areas are organized in a series of layers with specific types of cells and corresponding arborizations and connectivities. The theory does not say anything at this point about this level of the circuitry.

6.4 Directions for future research

6.4.1 Associative memories

In past work on HMAX we assumed that the hierarchical architecture performs a kind of preprocessing of an image to provide, as result of the computation, a vector (that we called "signature" here) that is then input to a classifier. This view is extended in this paper by assuming that *signature vectors* not only at the top of the hierarchy but at every complex cell level are input to an associative memory. In this way a number of properties of the image (and associations) can be recalled. Parenthetically we note that old *associative memories* can be regarded as vector-valued classifiers – an obvious observation.

An associative architecture for retrieval: dot products primitives and matrix computation by neurons

- a neuron and its $10^3 - 10^4$ synapses provide the basic computational operation: a *dot product*: the input f to the neuron gives the scalar $f\dot{t}$ as the output where t is the vector of synaptic weights.

- a set of K neurons (simple cells) computes the matrix operation $M^i f$, where

$$M^i = \begin{pmatrix} g_0 t^i \\ \cdots \\ g_K t^i \end{pmatrix}$$

- Then the output of each complex cell c_i is the average over i of $|M^i \cdot f|$ which can be written as the dot product $c = e^T |M^i \cdot f|$, where

$$e = \begin{pmatrix} 1 \\ \cdots \\ 1 \end{pmatrix}$$

- The signature vector c is used to access an associative memory represented as a matrix A. Let us denote a specific vector c as c^j and assume that the element of the matrix M have stored by associating the j signature with a set of properties given by the vector p^j: $M_{k,l} = \sum_j p_k^j c_l^j$. If c^n is noiselike then $Mc^n \sim p^n$.

There are interesting simulations and estimate of memory capacity, generalization should be done, especially when associative retrieval is performed at different levels for a single image.

Retrieving from an associative memory: optimal sparse encoding and recall There are interesting estimates of optimal properties of codes for associative memories, including optimal sparsness (see [71, 76]). It would be interesting to connect these results to estimated capacity of visual memory (Oliva, 2010).

Weak labeling by association of video frames Assume that the top associative module associates together images in a video that are contiguous in time (apart when there are clear transitions). This idea (mentioned to TP by Kai Yu) relies on smoothness in time to label via association. It is a very biological semi-supervised learning, essentially identical to a Foldiak-type rule. It is thus very much in tune with our proposal of the S:C memory-based module for learning invariances to transformations and with the ideas above about an associative memory module at the very top.

Space, time, scale, orientation Space and time are in a sense intrinsic to images and to their measurement. It seems that the retina is mainly dealing with those three dimensions (x, y, t), though x, y are sampled according to the sampling theorem in a way which is eccentricity-dependent forcing in later cortical layers the development of receptive field with a size which increases with eccentricity (spacing in the lattice and scale of receptive fields increase proportionally to $\sim logr$).

The theory assumes that at each eccentricity a set of receptive fields of different size (eg σ) exist during development at the level of developing simple cells, originating a set of *scales*. It is an open question what drove evolution to discover multiresolution analysis of the image. Given finite channel resources – eg bandwidth, number of fibers, number of bits – there is a tradeoff between size of the visual field and scale (defined as the resolution in terms of spatial frequency cutoff). Once multiple scales are superimposed on space (eg a lattice of ganglion cells in each x, y) by a developmental program, our theory describes how the orientation dimension is necessarily discovered by exposure to moving images.

6.4.2 Invariance and Perception

Other invariances in visual perception may be analyzed in a parallel way. An example is color constancy. Invariance to illumination (and color opponent cells) may emerge during development in a similar way as invariance to affine transformations. Thus we have a

Color constancy conjecture. The theory of Part I should be able to learn invariance to illumination by observing during development transformations in the appearance of the same scene under changes of the illuminant – direction and spectral composition. A natural conjecture emerging from the approach of Part II is that eigenvectors of the covariance matrix of such transformations of natural images may provide the spatial-chromatic tuning of different types of color opponent cells in V1 and other areas.

The idea that the key computational goal of visual cortex is to learn and exploit invariances extends to other sensory modalities such as hearing of sounds and of speech. It is tempting to think of music as an abstraction (in the sense of information compression and PCA) of the transformations of sounds. Classical (western) music would then emerge from the transformations of human speech (the roots of western classical music were based in human voice – Gregorian chants).

6.4.3 The dorsal stream

The ventral and the dorsal streams are often portrayed as *the what and the where* facets of visual recognition. It is natural to ask what the theory described here implies for the dorsal stream.

In a sense the dorsal stream seems to be the dual of the ventral stream: instead of being concerned about the invariances under the transformation induced by a Lie algebra it seems to represent (especially in MST) the orbits of the dynamical systems corresponding to the same Lie algebra.

6.4.4 Visual "concepts"

- *"Concept" of parallel lines* Consider an architecture using signatures. Assume it has learned sets of templates that guarantee invariance to all affine transformations. The claim is that *the architecture will appear to have learned the concept of parallel lines from a single specific example of two parallel lines*. According to the theorems in the paper, the signature of the single image of the parallel lines will be invariant to affine transformations (within some range).

- *Number of items in an image* A classifier which learns the number five in a way which is invariant to scale should be able to recognize five objects independent of class of objects.

- *Line drawings conjecture* The memory-based module described in this paper should be able to generalize from real images to line drawings when exposed to illumination-dependent transformations of images. This may need to happen at more than one level in the system, starting with the very first layer (eg V1). Generalizations with respect to recognition of objects invariant to shadows may also be possible.

6.4.5 Is the ventral stream a cortical mirror of the invariances of the physical world?

It is somewhat intriguing that Gabor frames - related to the "coherent" states and the *squeezed states* of quantum mechanics - emerge from the filtering operations of the retina which are themselves a mirror image of the position and

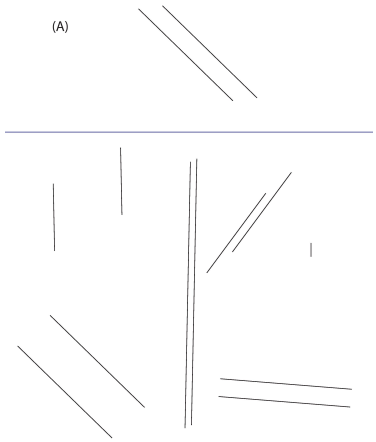


Figure 48: For a system which is invariant to affine transformations a single training example (A) allows recognition of all other instances of parallel lines – never seen before.

.

momentum operator in a Gaussian potential. It is even more intriguing that invariances to the group $SO2 \times \mathbb{R}^2$ dictate, according to our theory, the computational goals, the hierarchical organization and the tuning properties of neurons in visual areas. In other words: it did not escape our attention that the theory described here implies that the brain function, structure and properties reflect in a surprising direct way the physics of the visual world.

Acknowledgments We would like to especially thank Steve Smale, Leyla Isik, Owen Lewis, Steve Voinea, Alan Yuille, Stephane Mallat, Mahadevan, S. Ullman for discussions leading to this preprint and S. Soatto, J. Cowan, W. Freiwald, D. Tsao, A. Shashua, L. Wolf for reading versions of it. Andreas Maurer contributed the argument about small apertures in section 4.1.1. Giacomo Spigler, Heejung Kim, and Darrel Deo contributed several results including simulations. Krista Ehinger and Aude Oliva provided to J.L. the images of Figure 3 and we are grateful to them to make them available prior to publication. In recent years many collaborators contributed indirectly but considerably to the ideas described here: S. Ullman, H. Jhuang, C. Tan, N. Edelman, E. Meyers, B. Desimone, T. Serre, S. Chikkerur, A. Wibisono, J. Bouvrie, M. Kouh, M. Riesenhuber, J. DiCarlo, E. Miller, A. Oliva, C. Koch, A. Caponnetto, C. Cadieu, U. Knoblich, T. Masquelier, S. Bileschi, L. Wolf, E. Connor, D. Ferster, I. Lampl, S. Chikkerur, G. Kreiman, N. Logothetis. This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain and Cognitive Sciences, and which is affiliated with the Computer Sciences and Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO),

National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation.

References

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

[2] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

[3] J. Antoine, R. Murenzi, P. Vandergheynst, and S. Ali. *Two-dimensional wavelets and their relatives*. Cambridge University Press, Cambridge, UK, 2004.

[4] D. Arathorn. Computation in the higher visual cortices: Map-seeking circuit theory and application to machine vision. In *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop*, AIPR '04, pages 73–78, Washington, DC, USA, 2004. IEEE Computer Society.

[5] C. Baker, J. Liu, L. Wald, K. Kwong, T. Benner, and N. Kanwisher. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 104(21):9087, 2007.

[6] Y. Bengio and Y. LeCun. Scaling learning algorithms towards ai. *Large-Scale Kernel Machines*, 34, 2007.

[7] J. Bruna and S. Mallat. Invariant scattering convolution networks. *CoRR, http://arxiv.org/abs/1203.1513*, 2012.

[8] S. S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, May 2010.

[9] L. Cohen, S. Dehaene, and L. Naccache. The visual word form area. *Brain*, 123(2):291, 2000.

[10] D. Cox, P. Meier, N. Oertelt, and J. DiCarlo. 'Breaking'position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–1147, 2005.

[11] H. Cramer and H. Wold. Some theorems on distribution functions. *J. London Math. Soc.*, 4:290–294, 1936.

[12] J. Cuesta-Albertos. How many random projections suffice to determine a probability distribution? *IPMs sections*, 2009.

[13] J. Cuesta-Albertos, R. Fraiman, and R. T. A sharp form of the cramer–wold theorem. *Journal of Theoretical Probability*, 20:201–209, 2007.

[14] Y. Dan, A. J. J., and R. C. Reid. Effcient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. *The Journal of Neuroscience*, (16):3351 – 3362, 1996.

[15] P. Downing and Y. Jiang. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470, 2001.

[16] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[17] M. Ferraro and T. M. Caelli. Relationship between integral transform invariances and lie group theory. *J. Opt. Soc. Am. A*, 5(5):738–742, 1988.

[18] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

[19] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, 23(9):2289–2323, 2011.

[20] J. Freeman and E. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14:11951201, 2011.

[21] W. Freiwald and D. Tsao. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, 330(6005):845, 2010.

[22] K. 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, Apr. 1980.

[23] J. Gallant, J. Braun, and D. V. Essen. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, 1993.

[24] J. Gallant, C. Connor, S. Rakshit, J. Lewis, and D. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology*, 76:2718–2739, 1996.

[25] L. Galli and L. Maffei. Spontaneous impulse activity of rat retinal ganglion cells in prenatal life. *Science (New York, NY)*, 242(4875):90, 1988.

[26] S. Geman. Invariance and selectivity in the ventral visual pathway. *Journal of Physiology-Paris*, 100(4):212–224, 2006.

[27] D. George and J. Hawkins. A hierarchical bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 3, pages 1812–1817, 2005.

[28] L. Glass and R. Perez. Perception of Random Dot Interference Patterns. *Nature*, 31(246):360–362, 1973.

[29] K. Groechenig. Multivariate gabor frames and sampling of entire functions of several variables. *Appl. Comp. Harm. Anal.*, pages 218 – 227, 2011.

[30] A. Grossman, J. Morlet, and T. Paul. Transforms associated to square integrable group representations. ii: Examples. In *Annales de l'IHP Physique théorique*, volume 45, pages 293–309. Elsevier, 1986.

[31] J. Hegde and D. Van Essen. Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, 20(5):61, 2000.

[32] A. Heppes. On the determination of probability distributions of more dimensions by their projections. *Acta Mathematica Hungarica*, 7(3):403–410, 1956.

[33] G. Hinton and R. Memisevic. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 22:14731492, 2010.

[34] W. Hoffman. The Lie algebra of visual perception. *Journal of Mathematical Psychology*, 3(1):65–98, 1966.

[35] K. Hornik. Approximation capabilities of muitilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[36] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106, 1962.

[37] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866, Nov. 2005.

[38] A. Hyvrinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[39] L. Isik, J. Z. Leibo, and T. Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. *Frontiers in Computational Neuroscience*, 6, 2012.

[40] L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio. The timing of invariant object recognition in the human visual system. *Submitted*, 2013.

[41] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[42] J. Karhunen. Stability of Oja's PCA subspace rule. *Neural Comput.*, 6:739–747, July 1994.

[43] E. Y. Ko, J. Z. Leibo, and T. Poggio. A hierarchical model of perspective-invariant scene identification. In *Society for Neuroscience Annual Meeting Abstracts (486.16/OO26)*, Washington DC, USA, 2011.

[44] J. Koenderink. The brain a geometry engine. *Psychological Research*, 52(2):122–127, 1990.

[45] A. Koloydenko. Symmetric measures via moments. *Bernoulli*, 14(2):362–390, 2008.

[46] M. Kouh and T. Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural computation*, 20(6):1427–1451, 2008.

[47] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[48] Q. V. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. *CoRR,http://arxiv.org/abs/1112.6209*, abs/1112.6209, 2011.

[49] Y. LeCun. Learning invariant feature hierarchies. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 496–505. Springer, 2012.

[50] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pages 255–258, 1995.

[51] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[52] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–97. IEEE, 2004.

[53] A. Leff, G. Spitsyna, G. Plant, and R. Wise. Structural anatomy of pure and hemianopic alexia. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(9):1004–1007, 2006.

[54] J. Z. Leibo, J. Mutch, F. Anselmi, W. Freiwald, and T. Poggio. Part III:View-invariance and mirror-symmetric tuning in the macaque face-processing network. *in preparation*, 2013.

[55] J. Z. Leibo, J. Mutch, and T. Poggio. How can cells in the anterior medial face patch be viewpoint invariant? *MIT-CSAIL-TR-2010-057, CBCL-293; Presented at COSYNE 2011, Salt Lake City*, 2011.

[56] J. Z. Leibo, J. Mutch, and T. Poggio. Learning to discount transformations as the computational goal of visual cortex. *Presented at FGVC/CVPR 2011, Colorado Springs, CO.*, 2011.

[57] J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain, 2011.

[58] J. Z. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio. Learning Generic Invariances in Object Recognition: Translation and Scale. *MIT-CSAIL-TR-2010-061, CBCL-294*, 2010.

[59] J. Z. Leibo, J. Mutch, S. Ullman, and T. Poggio. From primal templates to invariant recognition. *MIT-CSAIL-TR-2010-057, CBCL-293*, 2010.

[60] N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–7, Sept. 2008.

[61] N. Li and J. J. DiCarlo. Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. *Neuron*, 67(6):1062–1075, 2010.

[62] B. Logan Jr. Information in the zero crossings of bandpass signals. *ATT Technical Journal*, 56:487–510, 1977.

[63] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[64] T. Masquelier, T. Serre, S. Thorpe, and T. Poggio. Learning complex cell invariance from natural videos: A plausibility proof. *AI Technical Report #2007-060 CBCL Paper #269*, 2007.

[65] S. Moeller, W. Freiwald, and D. Tsao. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, 320(5881):1355, 2008.

[66] V. Moro, C. Urgesi, S. Pernigo, P. Lanteri, M. Pazzaglia, and S. Aglioti. The neural basis of body form and body action agnosia. *Neuron*, 60(2):235, 2008.

[67] J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. *Computer Vision and Pattern Recognition 2006*, 1:11–18, 2006.

[68] C. Niell and M. Stryker. Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience*, 28(30):7520–7536, 2008.

[69] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[70] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992.

[71] G. Palm. On associative memory. *Biological Cybernetics*, 36(1):19–31, 1980.

[72] M. Peelen and P. Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005.

[73] N. Pinto, Z. Stone, T. Zickler, and D. Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 35–42. IEEE, 2011.

[74] D. Pitcher, L. Charles, J. Devlin, V. Walsh, and B. Duchaine. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, 19(4):319–324, 2009.

[75] W. Pitts, W.and McCulloch. How we know universals: the perception of auditory and visual forms. *Bulletin of Mathematical Biology*, 9(3):127–147, 1947.

[76] T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19(4):201–209, 1975.

[77] T. Poggio. Visual algorithms. *AIM-683*, 1982.

[78] T. Poggio. The computational magic of the ventral stream: Towards a Theory. *Nature Precedings, doi:10.1038/npre.2011.6117.1*, July 16 2011.

[79] T. Poggio. The computational magic of the ventral stream: Towards a Theory. Supplementary Material. *Nature Precedings, doi:10.1038/npre.2011.6117.1*, 2011.

[80] T. Poggio, J. Mutch, F. Anselmi, J. Z. Leibo, L. Rosasco, and A. Tacchetti. Invariances determine the hierarchical architecture and the tuning properties of the ventral stream. Technical Report available online, MIT CBCL, 2013. Previously released as MIT-CSAIL-TR-2012-035, 2012 and in Nature Precedings, 2011.