Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-035 December 29, 2012

The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work).

Tomaso Poggio, Jim Mutch, Joel Leibo, Lorenzo Rosasco, and Andrea Tacchetti

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work).

December 30, 2012

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 3.0 of a report first published online on July 20, 2011 (npre.2011.6117.1). Much progress has been made since the previous version: most of the many changes are additions but some results that were partly wrong (e.g. the hierarchy of different types of invariances) have been corrected. As much material as possible has been moved to the appendices (which are available by emailing to TP).

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 and 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 robustness to diffeomorphisms. 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 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 technique 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

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
- *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. 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.

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 pre-

dicted to converge to one of the eigenvectors. We assume that the development of tuning in the cortical cells takes place in stages – one area, that we call often 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 ([71, 72]), including a patch (called AL) which contains mirror symmetric cells and is the input to the pose-invariant patch (AM, [13]) – 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 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

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

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 [37]) 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. [32, 33, 30, 51, 3, 31]); 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 [59, 64, 66, 65] 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 [59, 64, 66, 65] to more recent technical reports [38, 39, 35, 36]. See also the section *Background* in Supp. Mat. [55]. After writing previous versions of this report, TP found a few interesting and old references about transformations, invariances and receptive fields, see [53, 21, 28]. 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 3.5.1 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 noise (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 [71, 72], including a patch containing mirror symmetric cells before the pose-invariant patch [13] – again because of spectral properties of the face templates.

Remarks

- **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.

- **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).

- **Feedforward architecture as an idealized description** The architecture we propose is hierarchical; its most basic skeleton is feedforward. The architecture we advocate is however more complex, involving memory access from different levels of the hierarchy as well as top-down attentional effects, possibly driven by partial retrieval from an associative memory.

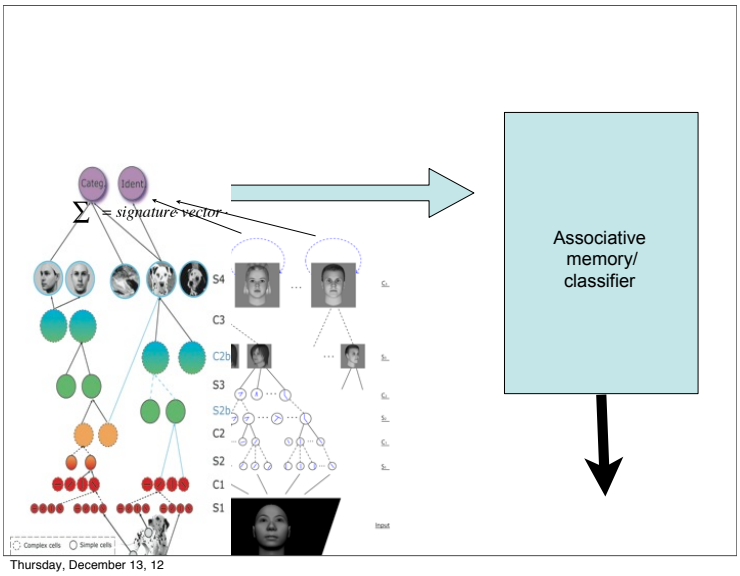


Figure 1: *Signatures from every level access associative memory modules.*

The neural implementation of the architecture requires local feedback loops within areas (for instance for normalization operations). The theory is most developed for the feedforward skeleton (probably responsible for the first 100 msec of perception/recognition).

- **Generic and class-specific transformations** We distinguish (as we did in past papers, see [56, 59]) 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 [36, 37].

- **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].

- **Spectral theory and receptive fields** Part II of the paper describes a *spectral theory* linking specific transformations and invariances to tuning properties of cells in each area. The most surprising implication is that the computational goals and some of the 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.

- **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: [10, 78, 69, 24, 43, 11]. There is also psychophysics and physiology evidence for these [5, 77, 41, 40]

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 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-

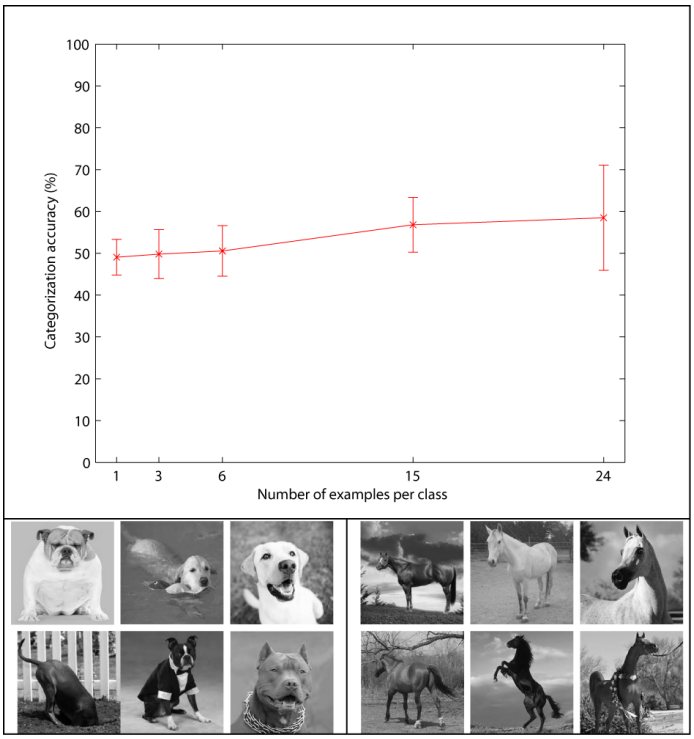


Figure 2: *Images of dogs and horses, in the wild, with arbitrary viewpoints (and clutter, eg background).The performance of a regularized least squares classifier (linear kernel, as in the next figure) is around chance. There are 60 images in total (30 per class) from Google. The x axis gives the number 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.*

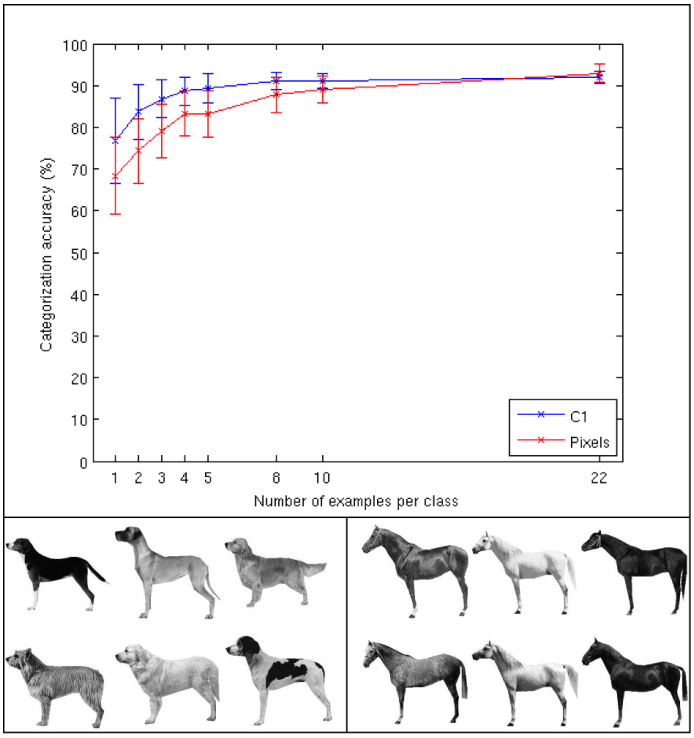


Figure 3: *Images of dogs and horses, 'normalized' with respect to image transformations. A regularized least squares classifier (linear kernel) tested on more than 150 dogs and 150 horses does well with little training. Error bars represent +/- 1 standard deviation computed over 100 train/test splits. 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 [74]). Recent techniques for normalizing for affine transformations are now well developed (see [80] for a review). Various attempts at learning transformations have been reported over the years (see for example [57, 30] and for additional references the paper by Hinton [20]).

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 hierarchical 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.

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.*

3.2 Templates and signatures

Summary. In this section we justify another assumption in the theory: a primitive computation performed by neurons is a 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 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 nonlinear 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.

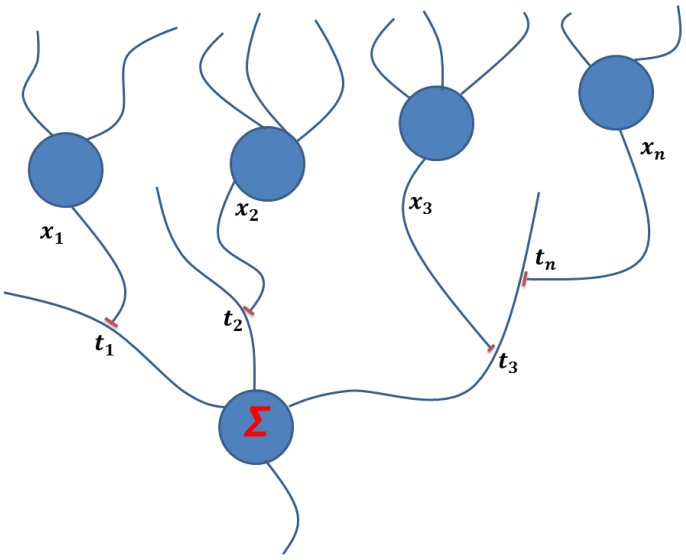


Figure 4: *A neuron receives on its dendritic tree in the order of* $10^3 - 10^4$ *synaptic inputs from other neurons. To a first approximation each synapse contributes a current which depends on the product of the input signal and the synapse. Since the soma of the neuron can be regarded as summating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

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 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

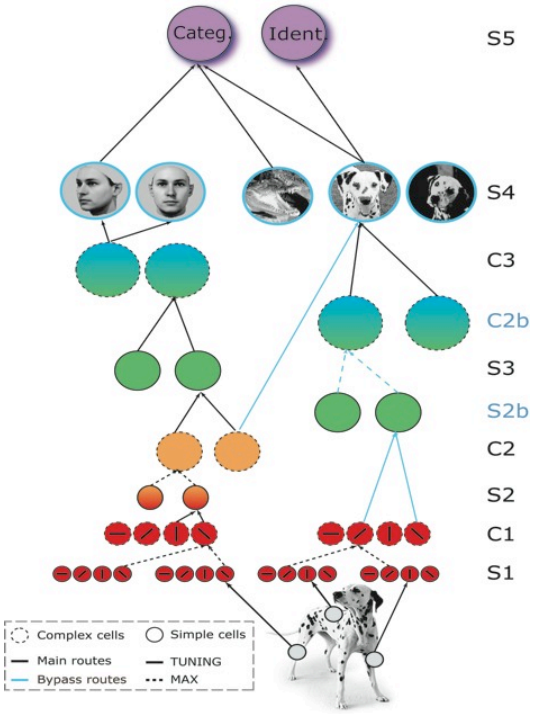


Figure 5: *Hierarchical feedforward model of the ventral stream – a modern interpretation of the Hubel and Wiesel proposal (see [58]). The theoretical framework proposed in this paper provides foundations for this model and how the synaptic weights may be learned during development (and with adult plasticity). It also suggests extensions of the model such as class specific modules at the top.*

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

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 OCR done via intersection of letters with a random, fixed set of lines and counting number of intersections (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*. Here it is just a suggestion that since there are no special conditions on the projections (though the assumption of randomness is 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 (and, of course, on the fact that they should be independent and informative).

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.

- 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*

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

$$\langle I, t \rangle = \int I(x,y)t(x,y)dxdy$$

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

$$\langle I, t_i \rangle, \quad i = 1, ..., T.$$

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

$$\Delta_I = (\langle I, t_1 \rangle, \langle I, t_2 \rangle, ..., \langle I, t_T \rangle)^T$$

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 23.1) for the n-dimensional space $\mathcal{X}$ spanned by n chosen images in R^d, that is the following holds

$$A||I||^2 \leq \sum_{k=1}^{T} | < I, t_k > |^2 \leq B||I||^2 \quad (1)$$

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 $u, v \in \mathcal{X}$ (where $\mathcal{X}$ is a n-dimensional subspace of R^d)

$$A \parallel u - v \parallel \leq \parallel Fu - Fv \parallel \leq B \parallel u - v \parallel .$$

If $A = 1 - \epsilon$ and $B \leq 1 + \epsilon$ where $\epsilon = \epsilon(T)$ the projections of u and v 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 $u, v \in V$*

$$(1 - \epsilon) \parallel u - v \parallel \leq \parallel Pu - Pv \parallel \leq (1 + \epsilon) \parallel u - v \parallel$$

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

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

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

Remarks

- The dimensionality of the measurement vector suggested 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.

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 $U(T) : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ transformations such that:

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

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 later, (e.g. in 3.4.2).
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 $\mathbb{R}^2$. Let us now define a key object of the paper:

Definition 4. *Suppose now we have a finite set of templates that are closed under the action of a group of transformations:*

$$G = (g_1, ..., g_{|G|}), \quad \mathcal{T} = (t_1, ..., t_T), \quad |G|, T < \infty$$

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

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

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 assumption 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 go back to ground zero: we consider not only the average of a function of the dot product but the full orbit – corresponding to the set of dot products. For compact groups if two orbits have a point in common then they are the same orbit. A distribution can be associated to each orbit and a distribution can be characterized in terms of its moments which are group averages of powers of the dot products. The overall logic is simple with some problems in the details. We also take somewhat of a detour in discussing sets of templates such as frames, random projections etc.

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 3.4.

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,

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

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, η, of the measurements which are the dot products of the image with all transformations of one of the templates, for each template:

$$\Sigma_t[I] = \frac{1}{|G|}\sum_{g\in G}\eta(\langle I, gt\rangle). \tag{2}$$

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

$$\Sigma_{t_i}(I) = \frac{1}{|G|}\sum_{j=1}^{|G|}|\langle I, g_j t_i\rangle|^2$$

Other examples of invariant group functionals are

- Max: $\Sigma_{t_i}(I) = max_j\langle I, g_j t_i\rangle$
- Average: $\Sigma_{t_i}(I) = \frac{1}{|G|}\sum_{j=1}^{|G|}\langle I, g_j t_i\rangle$

These functions are called *pooling or aggregation* functions. The original HMAX model uses a max of $I \circ g_j t_i$ over j or the average of $I \circ g_j t_i$ over j or the average of $(I \circ g_j t_i)^2$ over j. Often a sigmoidal function is used to describe the threshold operation of a neuron underlying spike generation. Such aggregation operations can 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 [29]). Notice that defining the p-norm of x with $||x||_p = (\sum |x_i|^p)^{\frac{1}{p}}$, it follows that $max(x) = ||x||_\infty$ and $energy-operation(x) = ||x||_2$. Therefore the invariant *signature*,

$$\Sigma(I) = (\Sigma_{t_1}(I), \Sigma_{t_2}(I), ..., \Sigma_{t_T}(I))^T$$

is a vector which is invariant under the transformations g_j.

Remarks

- *Group characters* As we will see later using templates that are the characters of the group is equivalent to performing the Fourier transform defined by the group. Since the Fourier transform is an isometry for all locally compact abelian groups, it turns out that the modulo or modulo square of the transform is an invariant.

- *Pooling functions* It is to be expected that different aggregation functions, all invariant, have different discriminability and noise robustness. For instance, the arithmetic average will support signatures that are invariant but are also quite similar to each other. On the other hand the max, also invariant, may be better at keeping signatures distinct from each other. This was the original reason for [58] to choose the max over the average.

- *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 Discrimination and invariance: distance between orbits

An invariant signature based on the arithmetic average is invariant but likely to be not discriminative enough. *Invariance is not enough: discrimination must also be maintained.* A signature can be however made more discriminative by using additional nonlinear functions of the same dot products. Next we discuss how group averages of a set of functions can characterize the whole orbit.

To do this, we go back to orbits as defined in equation 3.3.1. 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 3). As we mentioned, if two orbits intersect in one point they are identical everywhere. Thus equality of two orbits implies that at least one point eg image is in common.

The goal of this section is to provide a criterion that could be used in a biologically plausible implementation of when two empirical orbits are the same *irrespectively of the ordering of their points*. Ideally we would like to give meaning to a statement of the following type: if a set of invariants for u is ϵ close to the invariants associated with v, then corresponding points of the two orbits are ϵ close.

The obvious approach in the finite case is to rank all the points of the $\mathbf{u}$ set and do the same for the $\mathbf{v}$ set. Then a comparison should be easy (computationally). Another natural approach is to compare the distribution of numbers associated with the $\mathbf{u}$ set with the distribution associated with the $\mathbf{v}$ set. This is based on the following axiom (that we may take as a definition of equivalence between the orbits generated by G on the points u and v,

Definition 5.

$$p(u) = p(v) \iff u \sim v$$

where p is the probability distribution.

Thus the question focuses on how can probability distribution be compared. There are several metrics that can be used to compare probability distributions such as the Kolmogoroff-Smirnoff criterion and the Wasserstein distance. The K-S criterion is the simpler of the two. The empirical distribution function is defined as

$$F_n(u) = \frac{1}{n}\sum I_{U_i \leq u}$$

for n iid observations U_i, where I is the indicator function. The Kolmogoroff Smirnoff statistic for a given other cumulative empirical distribution function $G_n(u)$ is

$$D_n = sup_u |F_n(u) - G_n(u)|,$$

where sup_u is the supremum of the set of distances. By the GlivenkoCantelli theorem, if the samples come from the same distribution, then D_n converges to 0 almost surely.

An approach which seems possibly relevant, though indirectly, for neuroscience is related to the characterization of distributions in terms of moments. In fact, a sufficient (possibly infinite) number of moments uniquely characterizes a probability distribution. Consider an invariant vector $m^1(v)$ in $\mathbb{R}^d$ with components $m_i^1, i = 1, \cdots, d$ with $(m_i)^1(v) = \frac{1}{|G|}\sum_j ((v^j)_i)^1$ where $v^j = g_j v$ and $(v^j)_i$ is its i-component. Other similar invariant "moment" vectors such as $m_i^p(v) = \frac{1}{|G|}\sum_j ((v^j)_i)^p$ can be defined. Observe that intuitively a sufficient number ($p = 1, \cdots, P$) of moments $m_i^p(v)$ determines uniquely $(v^j)_i$ for all j and of course viceversa.

This is related to the *uniqueness of the distribution which is the solution of the moment problem* ensured by certain sufficient conditions such as the Carleman condition $\sum_{p=1}^{\infty} \frac{1}{(m^{2p})^{\frac{1}{2p}}}$ on the divergence of infinite sums of functions of the moments m^p. The moment problem arises as the result of trying to invert the mapping that takes a measure to the sequences of moments

$$m_p = \int x^p d\mu(x).$$

In the classical setting, μ is a measure on the real line. In this form the question appears in probability theory, asking whether there is a probability measure having specified mean, variance and so on, and whether it is unique. In the Hausdorff moment problem for a bounded interval, which without loss of generality may be taken as $[0, 1]$, the uniqueness of μ follows from the Weierstrass approximation theorem, which states that polynomials are dense under the uniform norm in the space of continuous functions on $[0, 1]$. In our case the measure μ is a Haar measure induced by the transformation group.

The following lemma follows:

Lemma 1. $p(u) = p(v) \iff m_i^p(u) = m_i^p(v), \forall i, p$

The lemma together with the axiom implies

Proposition 3. *If $min_g||gu - v||^2 = 0$ then $m_i^p(v) = m_i^p(u)$ for all p. Conversely, if $m_i^p(v) = m_i^p(u)$ for $i = 1, \cdots, d$ and for all p, then the set of gu for all $g \in G$ coincides with the set of all gv for all $g \in G$.*

We conjecture that a more interesting version of the proposition above should hold in terms of error bounds of the form:

Proposition 4. *For a given error bound ϵ it is possible (under weak conditions to be spelled out) to chose δ and K such that if $|m_i^p(v) - m_i^p(u)| \leq \delta$ for $p = 1, \cdots, K$ then $min_g||gu - v||^2 \leq \epsilon$.*

This would imply that moments computed in $\mathbb{R}^k$ can distinguish in an invariant way whether $\tilde{v}$ and $\tilde{u}$ are equivalent or not, in other words whether their orbits coincide or not. We now have to connect $\mathbb{R}^k$ to R^d. The key observation of course is that

$$\langle I, g_i^{-1} t \rangle = \langle g_i I, t \rangle$$

thus measurements of one of the images with "shifts" of the template are equivalent to measurements of (inverse) shifts of the image with a fixed template.

3.3.3 Frames and invariants

We know that with a sufficient number of templates it is possible to control ϵ and thus maintain distances among the points $\tilde{v}$ and $\tilde{u}$ and all of their $|G|$ translates (which do not need to be explicitly given: it is enough to have translates of the templates in order to have all the dot products between each random template and all the translates of each point, eg image). Thus the overall picture is that there is an embedding of n points and their translates in $\mathbb{R}^d$ into $\mathbb{R}^k$ that preserves approximate distances.

Consider now $\mathbb{R}^k$ and the projections in $\mathbb{R}^k$ of v and u and of their translates, that is $P(v) = \overline{v}$ and $P(u) = \overline{u}$ and $P(v^j) = \overline{v}^j$ etc. The same results wrt moments above also hold in $\mathbb{R}^k$. In other words, from a sufficient number of moments for each of the k coordinates, it is possible to estimate whether the orbits of $\overline{u}$ and $\overline{v}$ are the same or not. In particular, we conjecture that the following result should hold

Proposition 5. *For any given ϵ it is possible to chose δ and K such that if $|m_i^p(\overline{v}) - m_i^p(\overline{u})| \leq \delta$ for $p = 1, \cdots, K$ for all $i = 1, \cdots, k$ then $min_g||gu - v||^2 \leq \epsilon$.*

The resulting diagram is in Figure 7: images are on the left, the action of the group generates from each image a series of transformed images that form an orbit of the group. The real action is on the right side (in $\mathbb{R}^k$) where from a single image $\mathbf{u}$ on the left the orbit associated to the group is generated by the dot products of $\mathbf{u}$ with each template and the associate orbit. The orbits of $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ can be discriminated by the moments of each coordinate of the images providing two vectors of moments that discriminate between the two orbits but are invariant to the ordering of the transformed images. The diagram provides

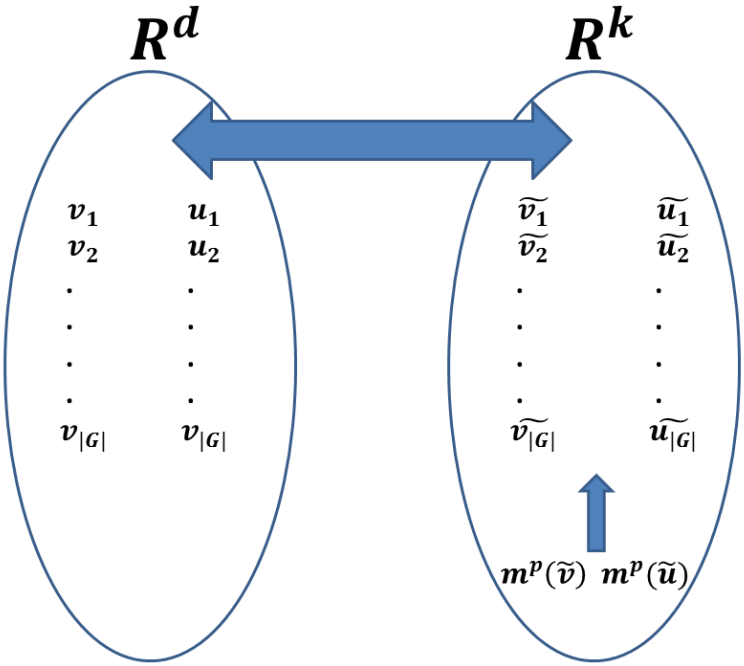


Figure 7: *Image space and feature space.*

an approach towards the characterization of the tradeoff between invariance and discriminability.

3.3.4 Random projections and invariants: an extension of J-L

Let us first consider how the JL result could be extended to the situation in which n points in $\mathbb{R}^d$ are projected in $\mathbb{R}^k$ using k random templates and their $|G|$ transformations induced by a group with $|G|$ elements.

Proposition 6. *For any set V of n points in $\mathbb{R}^d$ and for a group G of $|G|$ elements there exists k random templates and for each of the template and its $|G|$ transforms such that for all $u, v \in V$*

$$(1-\epsilon) \parallel u-v \parallel \leq \parallel P'u - P'v \parallel \leq (1+\epsilon) \parallel u-v \parallel$$

where the map P' includes the $|G|$ transforms of each of k random projection on $\mathbb{R}^k$ and

$$kC(\epsilon) \geq ln(n) + ln(|G|), \quad C(\epsilon) = \frac{1}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3})$$

The key point for biology is that the n vectors v and their $|G|$ transformations can be discriminated by random projections without the need to store explicitly the transformations of each vector: *a single image of an object is sufficient for invariant recognition of other views of the object!*

The previous result implies that by increasing the number of templates from $ln(n)$ to $ln(n)+ln(|G|)$ it is possible to maintain distances among the original n points and all of their $|G|$ translates (which do not need to be explicitly given: it is enough to have translates of the templates in order to have all the dot products between each random template and all the translates of each point, eg image). Thus the overall picture is that there is an embedding of n points and their translates in $\mathbb{R}^d$ into $\mathbb{R}^k$ that preserves approximate distances. The previous result guarantees that the n images and all of their transforms can be discriminated through their projections. The *selectivity-invariance tradeoff* is clear here: for a fixed number of templates (k) and a fixed accuracy (ϵ), there is an equivalent role for the number of discriminable objects, $ln(n)$ and number of transformations, $ln(|G|)$, and a direct tradeoff among them.
The key point for biology is that the n vectors v and their $|G|$ transformations can be discriminated by random projections without the need to store explicitly the transformations of each vector.
This observation allows to take measurements on the result of random projections to obtain signatures that are either selective to identity and invariant to transformations or selective to the transformation and invariant to identity.

The question is how – in addition to (non-invariant) distances in $\mathbb{R}^n$ – we may define invariants associated with each pattern which are the same for

every member of the set generated by the group. Notice that the Johnson-Lindenstrauss result implies that if u and v are very close (that is $||u - v|| \leq \eta$), their projections $P(u)$ and $P(v)$ are very close in every norm, in particular component-wise (that is $max_k|P(u)_k - P(v)_k| \leq \eta$).

Consider now $\mathbb{R}^k$ and the projections in $\mathbb{R}^k$ of v and u and of their translates, that is $P(v) = \overline{v}$ and $P(u) = \overline{u}$ and $P(v^j) = \overline{v}^j$ etc. The same results wrt moments above also hold in R^k. In other words, from a sufficient number of moments for each of the k coordinates, it is possible to estimate whether the orbits of $\overline{u}$ and $\overline{v}$ are the same or not. We can the use the extension in proposition 6 of the JL theorem to connect $\mathbb{R}^d$ to $\mathbb{R}^k$. In particular, a similar result should hold to frame proposition above:

Proposition 7. *For any given ϵ it is possible to chose δ and K such that if $|m_i^p(\overline{v}) - m_i^p(\overline{u})| \leq \delta$ for $p = 1, \cdots, K$ for all $i = 1, \cdots, k$ then $min_g||gu - v||^2 \leq \epsilon$.*

The diagram of Figure 7 should then describe the situation also for random projections.

Both random projections and frames behave like a quasi-isometry satisfying a frame-type bound. Of course random projections are similar to choosing random images as templates which are not natural images! Part II however considers templates which are Gabor wavelets (they emerge as the top eigenfunctions of templatebooks learned from "randomly" observed images undergoing an affine transformation). These templates are likely to be better characterized as randomly sampled frames than as random vectors! The most relevant situation is therefore when the templates are derived from a random subsampling from an overcomplete set. Corollary 5.56 of Vershynin "Introduction to non-asymptotic analysis of random matrices") gives conditions under which a random subset of size $N = O(nlogn)$ of a tight frame in $\mathbb{R}^n$ is an approximate tight frame ([76]).

Theorem 1. *Consider a tight frame $\{u_i, \quad i = 1, ..., M\}$ in $\mathbb{R}^n$ with frame bounds $A = B = M$. Let number m be such that all frame elements satisfy $||u_i||_2 \leq \sqrt{m}$. Let $\{v_i, \ i = 1, ..., N\}$ be a set of vectors obtained by sampling N random elements from the frame u_i uniformly and independently. Let $\epsilon \in (0, 1)$ and $t \geq 1$. Then the following holds with probability at least $12n^{-t^2}$: if $N \geq C(\frac{t}{\epsilon})^2 mlogn$ then v_i is a frame in $\mathbb{R}^n$ with bounds $A = (1 - \epsilon)N$ and $B = (1 + \epsilon)N$. Here C is an absolute constant. In particular, if this event holds, then every $x \in \mathbb{R}^n$ admits an approximate representation using only the sampled frame elements.*

3.3.5 Compact groups, probabilities and discrimination

In particular if G is a compact group, called $\mathcal{G}$, dg is a finite measure so that gI can be seen as a realization of a random variable with values in the signal space. A signature can be defined associating a probability distribution to each signal. Such a signature can be shown to be invariant and discriminant.

More precisely, if $\mathcal{G}$ is a compact group there is a natural Haar probability measure dg.
For any $I \in \mathcal{X}$, the space of signals, define the random variable,

$$Z_I : \mathcal{G} \to \mathcal{X}, \qquad Z_I(g) = gI.$$

Denote by P_I, the law (distribution) of Z_I, so that $P_I(A) = dg(Z_I^{-1}(A))$ for any borel set $A \subset \mathcal{X}$.
Let

$$\Phi_P : \mathcal{X} \to \mathcal{P}(\mathcal{X}), \qquad \Phi_P(I) = P_I,$$

where $\mathcal{P}(\mathcal{X})$ is the space of probability distribution on $\mathcal{X}$
We have the following fact.

Fact 1. *The signature Φ_P is invariant and discriminant 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 $\forall\ A \subseteq \mathcal{X}$

$$\int_A dP_I(s) = \int_A dP_{I'}(s)$$

This expression can be written equivalently as:

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

where

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

Now note that $\forall\ A \in \mathcal{X}$ if $gI \in A \Rightarrow g\bar{g}^{-1}\bar{g}I = g\bar{g}^{-1}I' \in A$, i.e. $g \in Z_I^{-1}(A) \Rightarrow g\bar{g}^{-1} \in Z_{I'}^{-1}(A)$. The inverse follows noticing that $g \in Z_{I'}^{-1}(A) \Rightarrow g\bar{g} \in Z_I^{-1}(A)$. Therefore $Z_I^{-1}(A) = Z_{I'}^{-1}(A)\bar{g},\ \forall A$. 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 variables on $\hat{g} = g\bar{g}^{-1}$ and the invariance property of the haar measure; this proves the implication.
To prove the implication $P_I = P_{I'} \Rightarrow I \sim I'$ note that $P_I - P_{I'} = 0$ is

equivalent to:

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

where with $\triangle$ we mean 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 $g', g'' \in \mathcal{G}$ such that $g'I = g''I'$. This implies $I = {g'}^{-1} g'' I' = \bar{g} I',\ \bar{g} = {g'}^{-1} g''$, i.e. $I \sim I'$. □

The signature Φ_P does not have a natural vector representation. One way to achieve such a representation is to consider the relationship between probability distribution and measurements.

3.3.6 Measurements and probability distributions

We begin by considering $\mathcal{X}$ to be a d dimensional space and $\mathcal{T} \subset \mathcal{X}$ an orthonormal basis (we will later relax this last assumption). Then we can define (with some abuse of notation) $\mathcal{T} : \mathcal{X} \to \mathbb{R}^d$, $\mathcal{T}(I) = \{(\langle I, t_i \rangle)_i,\ t_i \in \mathcal{T},\ \|\mathcal{T}(I)\|_d = \|I\|\}$, and the random variable

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

Denote by $P_I^{\mathcal{T}}$, the law (distribution) of $Z_I^{\mathcal{T}}$, so that $P_I^{\mathcal{T}}(A) = dg((Z_I^{\mathcal{T}})^{-1}(A))$ for any borel set $A \subset \mathbb{R}^d$ and let

$$\Phi_P^{\mathcal{T}} : \mathcal{X} \to \mathcal{G}(\mathbb{R}^d), \qquad \Phi(I) = P_I^{\mathcal{T}},$$

where $\mathcal{P}(\mathcal{X})$ be the space of probability distribution on $\mathbb{R}^d$

Fact 2. *If $\mathcal{T}$ is an orthonormal basis $\Phi_P^{\mathcal{T}}$ is discriminant and invariant.*

Proof. The proof follows the one in Fact 1 after noticing that $\mathcal{T} : \mathcal{X} \to \mathbb{R}^d$ is an isometry and therefore preserve the volumes which implies:

$$P_I^{\mathcal{T}}(A) = dg(Z_{\mathcal{T}I}^{-1}(A)) = dg(Z_I^{-1}(A)) = P_I(A),\ \forall A \in \mathcal{X}.$$

□

Fact 3. *If $\mathcal{T}$ is a frame $\Phi_P^{\mathcal{T}}$ (or Johnson Lindenstrauss), $\Phi_P^{\mathcal{T}}$ is discriminative and invariant.*

Proof. Suppose $P_I^{\mathcal{T}}(A) = P_{I'}^{\mathcal{T}}(A),\ \forall A \in \mathcal{X}$. Following the demonstration in Fact 1 we have that this implies

$$\mathcal{T} g I - \mathcal{T} g' I' = 0,\ \exists\ g' \in (Z_{I'}^{\mathcal{T}})^{-1}(A),\ \forall\ g \in\ (Z_I^{\mathcal{T}})^{-1}(A),\ A \in \mathcal{X}.$$

Therefore $\mathcal{T}(gI - g'I') = 0$; Being $\mathcal{T}$ a frame it implies $gI - g'I' = 0 \ \Rightarrow\ I' = \bar{g} I,\ \bar{g} = {g'}^{-1} g$, i.e., $I \sim I'$.
If $I \sim I'$ we have (following again the same reasoning done in Fact 1) that $g \in (Z_I^{\mathcal{T}})^{-1}(A) \ \Rightarrow\ g\bar{g}^{-1} \in (Z_{I'}^{\mathcal{T}})^{-1}(A)$ and $g \in (Z_{I'}^{\mathcal{T}})^{-1}(A) \ \Rightarrow\ g\bar{g} \in (Z_I^{\mathcal{T}})^{-1}(A)$. Therefore $(Z_I^{\mathcal{T}})^{-1}(A) = (Z_I^{\mathcal{T}})^{-1}(A)\bar{g}$ and using the Haar measure invariance we have proven the implication $I \sim I' \ \Rightarrow\ P_I^{\mathcal{T}} = P_{I'}^{\mathcal{T}}$. □

3.3.7 Moments

As we mentioned earlier, the moment problem concerns the question whether or not a probability distribution (or, equivalently, the associated random variable) is uniquely determined by the sequence of moments, all of which are supposed to exist, finite. The following result due to Papoulis [Probability, Random variables and Stochastic Processes, pg. $72-77,\ 1991$] is useful:

Fact 4. *Give two distribution $P, Q \in P(\mathcal{X})$, let*

$$\mu_n = \int_{\mathbb{R}^d} x^n Z(I)dx, \quad Z = P, Q \quad x \in \mathbb{R}^d$$

denote the $n-$th central moment. Then $P = Q$ if and only if $ch_P(t) = ch_Q(t)$ where $ch_Z(t)$ is the characteristic function of the distribution $Z(x)$ built from its moments μ_n:

$$ch_Z(t) = \mathfrak{F}[Z(x)](t) = 1 + it\mu_1 - \frac{1}{2!}t^2\mu_2 - \frac{1}{3!}t^3\mu_3 + \frac{1}{4!}t^4\mu_4 + ...$$

We can now define the map $M : \mathrm{P}(\mathbb{R}^d) \rightarrow \mathbb{R} \times \mathbb{N}^d$ and let

$$\Phi_M : \mathcal{X} \rightarrow \mathcal{P}(\mathbb{R}^d), \qquad \Phi_M^{\mathcal{T}}(x) = M(P_x^{\mathcal{T}}).$$

where M maps the probability distribution $P_x^{\mathcal{T}}$ into its moments (and cross-moments).

Fact 5. *The signature $\Phi_M^{\mathcal{T}}$ is invariant and discriminative, being $\mathcal{T}$ an orthonormal basis, Johnson Lindenstrauss or a frame.*

Proof. From above we have that the moments determine the characteristic function which uniquely determines the probability distribution. We then follow the proofs in the previous paragraph. □

3.4 Partially Observable Transformations (POTs)

Summary. The results described so far assume compactness of the transformation group and assume that all transformations are "visible". This is not fully realistic (though of course there are grids of cells and motion could be discretized). In this section we show how to relax the assumptions, though some work still needs to be done.

The invariance lemma is a key property that depends in our proof on the transformations having a compact group structure. As it turns out, it is possible to derive some invariance property under more general conditions. We describe here the following setup as one of the possible generalizations. Consider the case of translation of the image or image patch. Let us assume that the translation is on a torus (around the observer) but that only part of the torus is visible through a "window" interval, e.g an interval $I \in [-A, +A]$. Thus the transformations correspond to a compact group which is only *partially observable*.

3.4.1 Orbits and fragments

Notice that the prototypical group in the biological case consists of an abelian subgroup of the affine group. As a key case, consider translation which is abelian but only locally compact.

3.4.2 Invariance Lemma for POTs

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 $\mathcal{X}$ (see appendix 22).

Example 1. *Let $\mathcal{X} = L^2(\mathbb{R}^2)$, and $(T_g I)(x) = I(\sigma_g^{-1}(x))$, 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 $T_g^* I(r) = I(\sigma_g r)$ for all $g \in \mathcal{G}$ and $r \in \mathbb{R}^2$. Moreover, $T_g^* T_g = I$ so that $g \mapsto T_g$ is a unitary representation of $\mathcal{G}$ on $\mathcal{X}$.*

In the following we consider the continuous (and more general) version of eq. (2), the invariance lemma:

Invariance Lemma. Let $m, h : \mathcal{X} \to \mathbb{R}^2$ with $m(I) = \int h(T_g I)dg$. Then m is *invariant*:

$$m(T_{g'} I) = \int h(T_{g'} T_g I)dg = \int h(T_{g'g} I)dg = m(I), \tag{4}$$

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

Observation. Note that in the case the group of transformations is discrete and $h(I) = \eta(\langle I, t\rangle)$ for $t, I \in \mathcal{X}$, $\eta : \mathbb{R}^2 \to \mathbb{R}^2$ measurable, and T is a unitary representation, then

$$\eta(\langle T_g I, t\rangle) = \eta(\langle I, T_{g^{-1}} t\rangle).$$

and eq. (4) is exactly (2) in the continuous case.

Invariance Lemma for POTs. Let $\mathcal{G}_0 \subset \mathcal{G}$ and $m_0, h : \mathcal{X} \to \mathbb{R}^2$ with $m_0(I) = \int_{\mathcal{G}_0} h(T_g I)dg$.
Clearly, in this case m_0 is not invariant,

$$m_0(I) - m_0(T_{g'} I) = \int_{\mathcal{G}_0} h(T_g I)dg - \int_{\mathcal{G}_0} h(T_{g'} T_g I)dg \neq 0, \quad g, g' \in \mathcal{G} \tag{5}$$

The second integral of (5) can be written, with the variable change $\tilde{g} = g'g$ as

$$\int_{\mathcal{G}_0} h(T_{g'} T_g)I)dg = \int_{\mathcal{G}_0} h(T_{g'g} I)dg = \int_{g'^{-1}\mathcal{G}_0} h(T_{\tilde{g}} I)d\tilde{g}$$

where the last equality is true since we renormalize the operator T_g w.r.t. its Jacobian (see 3.2.3). With abuse of notation calling $\tilde{g} = g$:

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

where with Δ we indicate the symmetric difference. 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,

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

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

3.5 Hierarchical architectures: global invariance and local stability

Summary. In this section we find that the one-layer network, though globally invariant, cannot provide robust signatures for parts of the image. A hierarchical architecture of the same modules overcomes these limitations. The section shows that a hierarchical architecture of layers of dot products and pooling operations can be at each layer stable for small perturbations, locally invariant, covariant and, finally at the top, globally invariant

3.5.1 Limitations of one layer architectures: one global signature only

The architecture described so far is a one layer architecture – a $2D$ array of memory-based modules which, for each image, compute a single signature which is invariant to a group of global, uniform transformations such as the affine group. We conjecture that the architecture is optimal for (Gaussian) clutter since *matched filters* are optimal in the L^2 sense and the module performs a normalized dot-product – effectively a correlation (a max aggregation function would then compute the max of the correlation, an algorithm known as *matched filter*). Evolution may have in fact discovered first the one layer architecture: there is some evidence of correlation-based, non-invariant recognition in insects such as bees. The one-layer architecture has, however, a few weaknesses:

- Fragility to image perturbations such as non uniform warping or small shifts of image parts
- Memory storage issues when many object classes have to be recognized with tolerance to local and global affine transformations

The first problem is due to the global nature of the implemented invariance. In the next paragraphs we are going to see how a hierarchical architecture may solve this problem.

3.5.2 The basic idea

Consider a hierarchical architecture such as in Figure 8. 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 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 speculate that this kind of properties may be related to the minimal distances involved in crowding?

3.5.3 A hierarchical architecture: one dimensional translation group

In the following we are going to focus, as an easy example, on the locally compact group of one dimensional translations implemented by the operator

$$T_\xi : \mathcal{X} \to \mathcal{X}, \;\; (T_\xi I)(\tau) = I(\tau - \xi), \; I \in \mathcal{X}, \xi \in \mathbb{R}\,.$$

We fix the following basic objects:

- $\mathcal{X} = L^2(\mathbb{R})$, space of images.
- $\mathcal{T} \subset \mathcal{X}$, $|\mathcal{T}| < \infty$, the template set.
- $\eta : \mathcal{X} \to \mathcal{X}$ a non-linear function.
- $K_n : \mathbb{R} \to \{0, 1\}$ the characteristic function of the interval $P_n \subseteq \mathbb{R}$ where $P_{n-1} \subseteq P_n, \;\; \forall\, n = 1, ..., N$ or Gaussian functions of σ_n width.

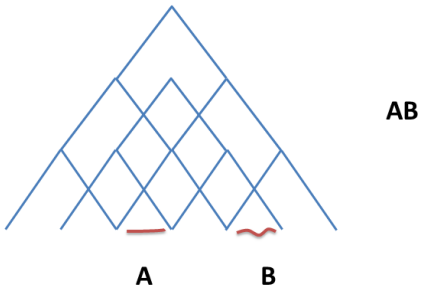


Figure 8: *Each of the nodes $\bigwedge$ is invariant for shifts of a pattern within its receptive field; we also assume that the output layer at each level is covariant.*

We are going to define a hierarchical construction where each interval P_n will be called *pooling range* and the index n *layer*; the basic building blocks of the construction are given by the following two operators:

Definition 6. Simple and complex response
The complex response operator $c_n : \mathcal{X} \to \mathcal{X}$, is iteratively defined as:

$$c_n(I)(\xi) = (K_n * \eta(s_n(I)))(\xi) = \langle K_n, T_\xi \eta(s_n(I)) \rangle \tag{6}$$

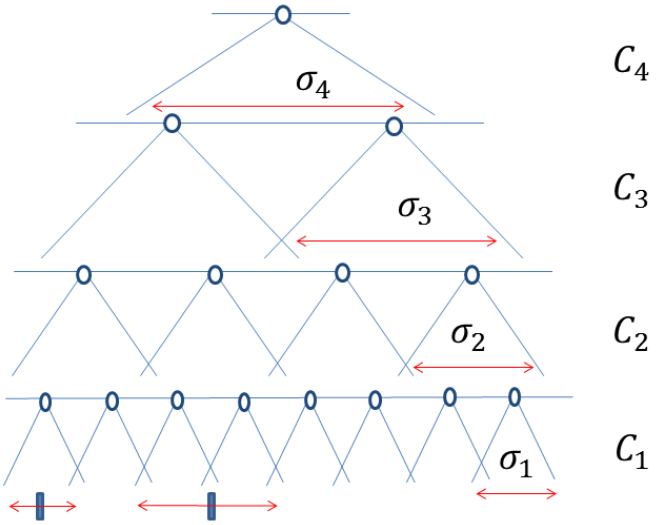
in terms of the simple response operator $s_n : \mathcal{X} \to \mathcal{X}$:

$$s_n(I)(\xi) = (c_{n-1}(I) * t)(\xi) = \langle c_{n-1}(I), T_\xi t \rangle \,,\ t \in \mathcal{T},\ I \in \mathcal{X}, \xi \in \mathbb{R} \tag{7}$$

where $c_0(I) \equiv I$.

Remark 1. *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 $\sigma^n_{eff} \geq \sigma^{n-1}_{eff}$. In general σ^n_{eff} increases (rapidly) with n where with σ_{eff} we mean the image part effectively seen by a complex response at layer n.

3.5.4 Properties of simple and complex responses

We provide now our definition of covariance and invariance for the set of responses c_n at each layer n of the architecture. We, for simplicity are focusing on translations in $1D$. However we can generalize the reasoning to translations in x, y which shifts "stuff" across receptive fields. In fact we can think of layers at each level indexed by orientation and scale. Thus our invariance and covariance definitions are for a given orientation and scale.
For $1D$ translations we are interested into studying the following two properties:

1. **Covariance:** the operator $f : \mathcal{X} \to \mathcal{X}$ is covariant with respect translations if:
$$f(T_\xi(I)) = T_\xi(f(I)), \ \ \forall\, I \in \mathcal{X},\ \xi \in \mathbb{R}$$

2. $r-$**invariance:** the operator $f : \mathcal{X} \to \mathcal{X}$ is $r-$invariant for $I \in \mathcal{X}$ with respect to translations if:
$$f(T_\xi(I)) = f(I), \ \ \xi \in [0, r], \ \ r \in \mathbb{R}$$

Remark 2. *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 3. *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).*

3.5.5 Property 1: covariance

The covariance of the complex response is a key ingredient in the analysis of the invariance; we prove:

Proposition 8. *The operator c_n is covariant with respect to translations.*
Let $c_n : \mathcal{X} \to \mathcal{X}$ the complex response at layer n and $I \in \mathcal{X}$ then:

$$c_n(T_{\bar\tau} I) = T_{\bar\tau}(c_n(I)), \ \ \forall \bar\tau \in \mathbb{R}\,.$$

Proof. We prove the proposition by induction. For $n = 1$ the covariance of the $s_1(I)$ function follows from:

$$s_1(T_{\bar\tau} I)(\tau) = \langle T_{\bar\tau} I, T_\tau t\rangle = \langle I, T_{\tau-\bar\tau} t\rangle = s_1(I)(\tau - \bar\tau) = (T_{\bar\tau} s_1(I))(\tau)$$

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

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

where on the second line we used the covariance property of $s_1(I)$ and on the third we used the change of variable $\hat{\tau} = \tilde{\tau} - \bar{\tau}$.
Suppose now the statement is true for n; by definition:

$$\begin{aligned} s_{n+1}(T_{\bar{\tau}}I)(\tau) &= \langle c_n(T_{\bar{\tau}}I), T_\tau t\rangle = \langle T_{\bar{\tau}}c_n(I), T_\tau t\rangle = \langle c_n(I), T_{\tau-\bar{\tau}}t\rangle \\ &= s_{n+1}(I)(\tau-\bar{\tau}) = (T_{\bar{\tau}}s_{n+1}(I))(\tau) \end{aligned}$$

Therefore

$$\begin{aligned} c_{n+1}(T_{\bar{\tau}}I)(\tau) &= \int K_{n+1}(\tau-\tilde{\tau})\eta(s_{n+1}(T_{\bar{\tau}}I)(\tilde{\tau}))d\tilde{\tau} \\ &= \int K_{n+1}(\tau-\tilde{\tau})\eta(s_{n+1}(I)(\tilde{\tau}-\bar{\tau}))d\tilde{\tau} \\ &= c_{n+1}(I)(\tau-\bar{\tau}) = (T_{\bar{\tau}}c_{n+1}(I))(\tau) \end{aligned}$$

where in the fourth line we used the change of variables $\tilde{\tau} - \bar{\tau} = \hat{\tau}$.
By induction, the statement is true for all layers. □

3.5.6 Property 2: partial and global invariance

We now prove that the functions c_n are approximately invariant (locally invariant) to translations within the range of the pooling. We further prove that the invariance increases from layer to layer in the hierarchical architecture.In the following, for reasons that will be clear later we choose as non linearity the modulus function, $\eta \equiv |\cdot|$.

Proposition 9. *Let $c_n : \mathcal{X} \to \mathcal{X}$ the complex response at layer n and $I \in \mathcal{X}$ then:*

$$c_n(T_{\bar{\tau}}I)(\tau) = c_n(I)(\tau), \quad I \in \mathcal{X}, \tag{8}$$

if

$$|s_n(I)(\tau)| = 0 \quad \tau \in P_n \Delta T_{\bar{\tau}} P_n$$

where Δ is the symmetric difference.

Proof. Let the pooling at layer n be achieved by a characteristic function on the interval P_n We have

$$\begin{aligned} &c_n(T_{\bar{\tau}}I)(\tau) - c_n(I)(\tau) = c_n(I)(\tau-\bar{\tau}) - c_n(I)(\tau) \\ &= \int_{\mathbb{R}} \Big(K_n(\tau-\bar{\tau}-\tilde{\tau}) - K_n(\tau-\tilde{\tau})\Big)|s_n(I)(\tilde{\tau})|d\tilde{\tau} \\ &= \int_{P_n\Delta T_{\bar{\tau}}P_n} |s_n(I)(\tilde{\tau})|d\tilde{\tau} \end{aligned}$$

where on the first line we used the covariance properties of the function $c_n(I)$.
□

Further the invariance is increasing from layer to layer since the effective pooling range at layer n will be

$$P_n^{eff} = P_n \times P_{n-1} \times \cdots \times P_0$$

Consequently for each transformation there will exists a layer such that the complex response is invariant to the transformation. We have

Theorem 2. *Let $I \in \mathcal{X}$, $c_n : \mathcal{X} \to \mathcal{X}$ the complex response at layer n. Let $\xi \in [0, \bar{\tau}],\ \bar{\tau} \in \mathbb{R}$, we have that c_n is $\bar{\tau}-$invariant for sone $\bar{n}$, i.e.*

$$c_{\bar{n}}(T_\xi I) = c_{\bar{n}}(I),\ \ \exists\, \bar{n}\ \ s.t.\ \ \forall m \geq \bar{n},\ \xi \in [0, \bar{\tau}]$$

Proof. The proof comes from the fact that for any $\xi \in [0, \bar{\tau}]$ there always exist an $n = \bar{n}$ such that $forall\tau \in P_{\bar{n}} \Delta T_{\bar{\tau}} P_{\bar{n}}$, $|s_{\bar{n}}(I)(\tau)| = 0$. The proof follows from 9. □

Remark 4. *If in 9 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 5. *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 images 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 the set of patches which has an invariant signature – at some level of the architecture – under affine transformations.*

3.5.7 Property 3: stability to perturbations

In the paragraphs above we assumed, for simplicity, to have one template at each layer. We will now suppose to have a set of templates at each layer that form a frame with good frequency localization properties (e.g. wavelets)

$$\mathcal{T}_n = \{t_n^i,\ i = 1, ..., T_n\}$$
$$a_n \left\| I \right\|_2 \leq \Big(\sum_{i=1}^{T_n} |\left\langle t_n^{i_n}, I \right\rangle|^2 \Big)^{\frac{1}{2}} = \left\| I \right\|_{\ell^2} \leq b_n \left\| I \right\|_2,\ \ a_n < b_n \in \mathbb{R}^+\ \ I \in \mathcal{X}.$$

Let the non$-$linearity $\eta : \mathcal{X} \to \mathcal{X}$ be the modulus square function $\sigma(\cdot) = |\cdot|$.
Using this formalism we can write the complex response $c_n(I)$ at layer n by components using the multi$-$index $\lambda \equiv (i_0, ..., i_{n-1})$ as follow.
Let $\mathcal{C}_n : \mathcal{X} \to \ell^2(\mathbb{R}^T)$, $T = T_0 + ... + T_n$, the map

$$I \longrightarrow c_n^\lambda(I) = c_n^{i_{n-1}, i_{n-2}, ..., i_1}(I) = K_n * |t_{n-1}^{i_{n-1}} * K_{n-1} * |t_{n-2}^{i_{n-2}} * \cdots * |t_1^{i_1} * I| \cdots |$$
$$I \in \mathcal{X},\ i_1 = 1, ..., T_1; ...; i_{n-1} = 1, ..., T_{n-1}$$

with norm in $\ell^2(\mathbb{R}^T)$,

$$\|c_n(I)\|_{\ell^2} = \sum_{i_{n-1},i_{n-2},...,i_1} \Big(|K_n * |t_{n-1}^{i_{n-1}} * K_{n-1} * |t_{n-2}^{i_{n-2}} * \cdots * |t_1^{i_1} * I| \cdots ||^2\Big)^{\frac{1}{2}}$$

We study now, starting from the top $c_n(I)$ complex response, the tolerance of the hierarchical representation to small perturbations $I \to I + \delta I$. As in the previous paragraph, let $\mathcal{T}_n = \{t_n^{i_n},\ i = 1, ..., T_n\}$ a frame at layer n.

Theorem 3. *Let $I \in \mathcal{X}$. We have*

$$\|c_n(I) - c_n(I + \delta I)\|_{\ell^2} \leq \prod_{i=1}^{n} b_i \|\delta I\|_\infty \tag{9}$$

with

$$\|c_n(I) - c_n(I + \delta I)\|_{\ell^2} = \Big(\sum_{i_1,...,i_n}^{T_1,...,T_n} |t^{in_n} * [c_n^{i_1,...,i_{n-1}}(I) - c_n^{i_1,...,i_{n-1}}(I + \delta I)|^2 \Big)^{\frac{1}{2}} \tag{10}$$

and

$$\|\delta I\|_\infty = sup_\tau |\delta I(\tau)|$$

The theorem says that the representation is continuous to perturbations from $L^2(\mathbb{R})$ to $\ell_2(\mathbb{R}^{T^n})$, $T = \sum_{i=1}^n T_i$

Proof. Being $\mathcal{T}_n = \{t_n^{i_n}, i = 1, ..., T_n\}$ a frame we have

$$\begin{aligned} \|c_n(I) - c_n(I + \delta I)\|_{\ell^2}^2 &= \sum_{i_1,...,i_n}^{T_1,...,T_n} |t_n^{i_n} * [c_n^{i_1,...,i_{n-1}}(I) - c_n^{i_1,...,i_{n-1}}(I + \delta I)]|^2 \\ &\leq b_n^2 \sum_{i_1,...,i_{n-1}}^{T_1,...,T_{n-1}} |c_n^{i_1,...,i_{n-1}}(I) - c_n^{i_1,...,i_{n-1}}(I + \delta I)|^2. \end{aligned}$$

Using the definitions of complex and simple response, $c_n^{i_n}(I) = K_n * |s_n^{i_n}(I)|$ and $s_n^{i_n}(I) = t_{n-1}^{i_{n-1}} * c_{n-1}(I)$

$$\|c_n(I) - c_n(I + \delta I)\|_{\ell^2}^2 \leq b_n^2 \sum_{i_1,...,i_{n-1}}^{T_1,...,T_{n-1}} |K_n * \Big(|t_{n-1}^{i_{n-1}} * c_n^{i_1,...,i_{n-2}}(I)| - |t_{n-1}^{i_{n-1}} * c_n^{i_1,...,i_{n-2}}(I + \delta I)| \Big)|^2$$

The convolution with K_n is a lowpass filter applied to the modulus square i.e. is decreasing the energy content of the signal by filtering. Further using the contraction properties of the modulus function i.e. $||a| - |b|| \leq |a - b|$

$$\|c_n(I) - c_n(I + \delta I)\|_{\ell^2}^2 \leq b_n^2 \sum_{i_1,...,i_{n-1}}^{T_1,...,T_{n-1}} |t_{n-1}^{i_{n-1}} * \Big(c_n^{i_1,...,i_{n-2}}(I) - c_n^{i_1,...,i_{n-2}}(I + \delta I)| \Big)|^2$$

The last equation has the same form of eq. (10) with the frame factor b instead of the sum over i_n. Repeating the reasoning for all layers

$$\|c_n(I) - c_n(DI)\|_{\ell^2} \leq \prod_{i=2}^{n} b_i^2 \sum_{i_1} |t_1^{i_1} * (I - (I + \delta I))|^2 \leq \prod_{i=1}^{n} b_i^2 \|\delta I\|_\infty^2$$

Taking the root square ends the proof. □

3.5.8 A hierarchical architecture: summary

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:

$$c_n(I)(\xi) = (K_n * \eta(s_n(I)))(\xi) = \langle K_n, T_g \eta(s_n(I)) \rangle \tag{11}$$

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

$$s_n(I)(\xi) = (c_{n-1}(I) * t)(\xi) = \langle c_{n-1}(I), T_\xi t \rangle \, , \; t \in \mathcal{T}, \; I \in \mathcal{X}, g \in G \tag{12}$$

where $c_0(I) \equiv I$.

Same kind of results obtained before for covariance, invariance and robustness to local perturbations can be obtained.

3.6 A short mathematical summary of the argument.

The theory just described has a simple mathematical structure, despite the mixed biological details. We summarize in this appendix.

3.6.1 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.

3.6.2 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, $A \|I\| \leq \|s(I)\|_2 \leq B \|I\|$, with $0 < A \leq B < \infty$.
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 $(1 - \epsilon) \|I\| \leq \|s(I)\|_2 \leq (1 + \epsilon) \|I\|$, as soon as $p \geq 8 \log n / \epsilon^2$.

3.6.3 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 $\mathcal{X}$.

Example 3. *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 $\mathcal{X}$.*

3.6.4 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.

3.6.5 Approximately invariant measurements via local group integration

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

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

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,

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

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,

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

Example 4. *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, ${g'}^{-1} \mathcal{G}_0 = [-\pi - \tau, \pi - \tau]$.*

3.6.6 Signature of approximately invariant measurements

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

3.6.7 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.

3.6.8 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.

3.6.9 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: Transformations, Apertures 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 makes a biologically plausible assumption: 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. 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. Since the size of the receptive fields in the hierarchy affects which transformations dominate, it follows that the level of the hierarchy determines the spectral properties and thus the tuning of the cells. Furthermore, invariance is now obtained by pooling nonlinear functions such as the mouduli of the dot products between the eigenfunctions (computed over the transformation of interest) and the new image.

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 mostly translations will be represented for "small" apertures with scale and orientation changes been relevant later in the hierarchy.

Let us be more specific. Suppose that the first layer consists of an array

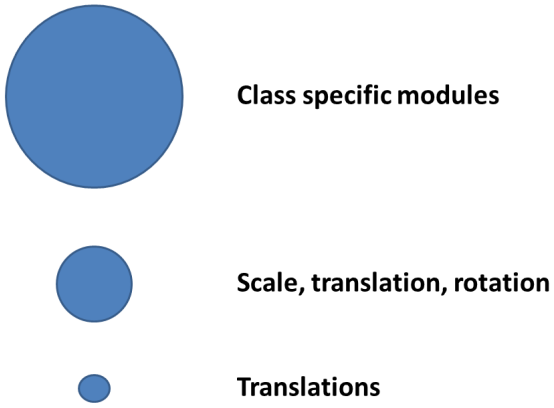


Figure 9: *The conjecture is that receptive field sizes affects not only the size but also the type of transformations that is learned and represented by the templates. In particular, small apertures (such as in V1) only "see" (small) translations.*

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.

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

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

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

Theorem 4. *(Poggio-Maurer) There exists $V \in \mathbb{R}^N$ such that for all $(t,x) \in I \times B$*

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

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

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

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

whence

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

□

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 4) – 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*.

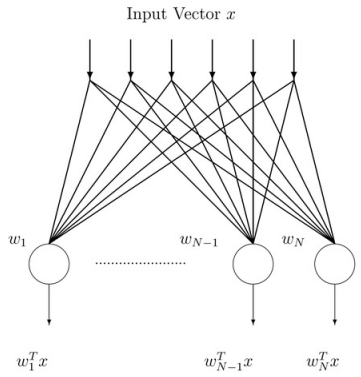
4.2 Linking conjecture: developmental memory is Hebbian

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.



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). Here we study a biologically more plausible online learning rule. Synapses change incrementally after each frame, effectively compressing information contained in the templates

and possibly making signatures more robust to noise. Plausible online learning rules for this goal are associative Hebb-like rules which may lead synapses at the level of simple-complex cells to match their tuning to the eigenvectors of the templatebooks (Hebb-like are known to be online algorithms for learning the PCA of a set of input patterns). Thus the receptive field at each layer would be determined by the transformations represented by the complex cells pooling at each layer. Later we will discuss in some detail Oja's rule[26] as an example. It is not the only one with the properties we need but it is a simple rule and variations of it are biologically plausible. The key point for now is that the conjecture links the spectral properties of the transformation being represented in each layer to the spectral properties of the templatebook at each level.

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 [47, 26] defines the change in presynaptic weights $\mathbf{w}$ given the output response $\mathbf{y}$ of a neuron to its inputs to be

$$\Delta \mathbf{w} \;=\; \mathbf{w}_{n+1} - \mathbf{w}_n \;=\; \eta\, y_n(\mathbf{x}_n - y_n \mathbf{w}_n) \tag{13}$$

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

$$\dot{\mathbf{w}} \;=\; \gamma\, y(\mathbf{x} - y\mathbf{w}) \tag{14}$$

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:

$$w_i(n+1) \;=\; \frac{w_i + \eta\, y(\mathbf{x}) x_i}{\left(\sum_{j=1}^{m} [w_j + \eta\, y(\mathbf{x}) x_j]^p\right)^{1/p}} \tag{15}$$

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 – ([73]) is in fact experimentally well established.

The above learning rules converge to the PCA with the largest eigenvalue (see Appendix 24). 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 10. *The Oja flow (Equation 13) 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 18).

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 [62, 48]. In the presence of noise, 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 18). 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 24.2) mabe responsible for wiring simple cells with the "same" orientation to the same complex cell.

It has been customary (for instance see **[?]** 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 hanges in the Oja equation give an online algorithm for computing ICAs instead of PCAs [23]. Which kind of plasticity is true biologically is 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 18 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 18. 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. 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

$$\mathbb{T}_t^* \mathbb{T}_t \psi_i = \lambda_i \psi_i, \ i = 1, ..., N \tag{16}$$

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 (16): in this case the basic observable given by the operator $T : \mathcal{X} \to \mathcal{X}$

$$(TI)(y) \equiv [t * M_a I](y) = \int dx t(y-x)a(x)I(x), \;\; t \in \mathcal{T}, \; a, I \in \mathcal{X} \qquad (17)$$

where

$$(M_a I)(x) = a(x)I(x), \; a \in \mathcal{X}$$

The equation (17) 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 6. *T is linear (from the properties of the convolution operator) and bounded (from $\|T\| = \|\mathfrak{F}(T)\| = \|t\| \, \|a\|$).*

Remark 7. *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

$$a(y) \int dx\ a(x)t^{\circledast}(y-x)\psi_n(x) = \lambda_n \psi_n(y) \tag{18}$$

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 12.

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.

[4]This problem has also been considered in recent work from Andrew Ng's group [63].

- *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 [61] (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.

- *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 $T^T G^T GT$ 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 (18) (see remark 9) 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 (18)

$$\int dx g(y)g(x)\psi_n(x)t^{\circledast}(y - x) = \lambda_n \psi_n(y)$$

holds for all apertures defined by functions $g(x)$.

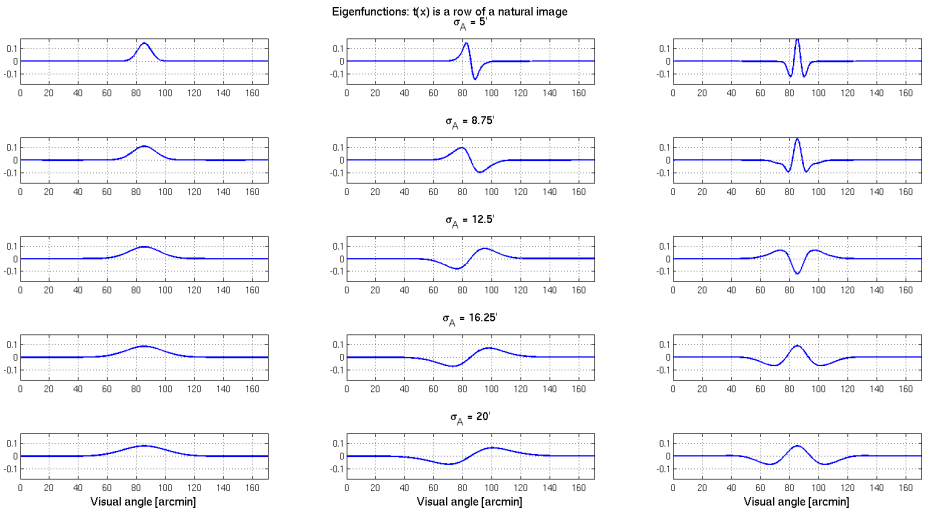


Figure 10: *Continuous spectrum of the covariance of the templatebook: Gabor-like eigenfunctions for different σ*

Remark 8. *Eq.* (18) *can be easily written in the case $x \in \mathcal{Y} = L^2(G, dg)$ being G a locally compact group*

$$[T^*TI](g') = \int dg\ K(g, g')I(g), \quad K(g, g') = a(g)a(g')t^{\circledast}(g^{-1}g'), \quad I \in \mathcal{Y},\ g, g' \in G.$$

The convolution is now on the group G.

Remark 9. *In $2D$ the spectral problem is:*

$$\int d\xi d\eta g(x, y)g(\xi, \eta)t^{\circledast}(\xi - x, \eta - y)\psi_n(\xi, \eta) = \lambda_n\psi_n(x, y). \tag{19}$$

where $t^{\circledast} \equiv t \circledast t$.

Numerical simulations in 1D show Gabor-like wavelets (see Figure 10) 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 11, 12 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 13 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.

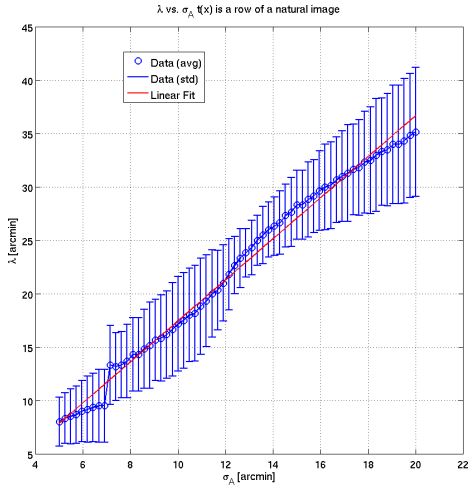


Figure 11: *Continuous spectrum: λ vs. σ_α for even symmetric patterns. The slope in this figure is k where $\lambda = k\sigma_\alpha$; $k \sim 2$ in this figure.*

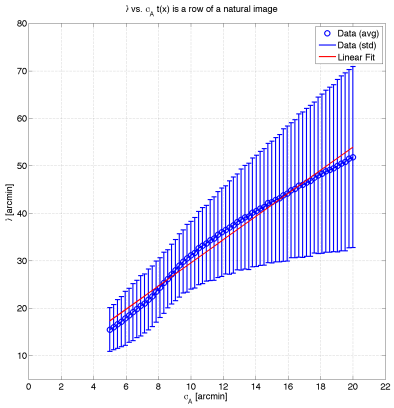


Figure 12: *Continuous spectrum: λ vs. σ_α for odd symmetric patterns. The slope is ~ 2.4. In 1D, odd symmetric eigenfunctions tend to have a lower modulating frequency.*

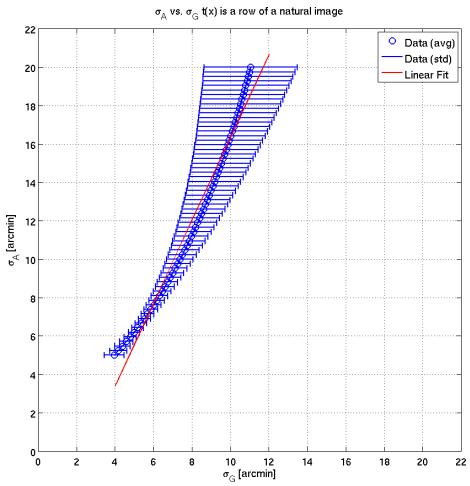


Figure 13: *Continuous spectrum: σ_α vs. σ_β. The slope is ~ 2. Though 1D, this is consistent with experimental data from [25] and [60] shown in fig. 19 where the slope is also roughly 2.*

1D: $t^{\circledast}(\omega_x)$ approximately piecewise constant

We represent the template as:

$$t^{\circledast}(x) = \frac{1}{\sqrt{2\pi}} \int d\omega \; t^{\circledast}(\omega) e^{i\omega x} \tag{20}$$

and assume that the eigenfunction has the form $\psi(x) = e^{-\frac{\beta}{2}x^2} e^{i\omega_g x}$, where β and ω_g are parameters to be found.
With this assumptions eq. (18) reads:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\alpha}{2}y^2} \int dx \; e^{-\frac{x^2(\alpha+\beta)}{2}} \int d\omega \; t^{\circledast}(\omega) e^{i\omega(y-x)} e^{iw_g x} = \lambda(\omega_g) e^{-\frac{\beta y^2}{2}} e^{i\omega_g y}. \tag{21}$$

Collecting the terms in x and integrating we have that the l.h.s becomes:

$$\sqrt{\frac{1}{\alpha+\beta}} e^{-\frac{\alpha}{2}y^2} \int d\omega \; t^{\circledast}(\omega) e^{i\omega y} e^{-\frac{(\omega-\omega_g)^2}{2(\alpha+\beta)}}. \tag{22}$$

With the variable change $\bar{\omega} = \omega - \omega_g$ and in the hypothesis that $t^{\circledast}(\bar{\omega}) \approx const$ over the significant support of the Gaussian centered in $\overline{\omega}$, integrating in $\bar{\omega}$ we have:

$$\sqrt{2\pi} \; const \; e^{-\frac{y^2\alpha}{2}} e^{i\omega_g y} e^{-\frac{y^2(\alpha+\beta)}{2}} \sim \lambda(\omega_g) e^{-\frac{y^2\beta}{2}} e^{i\omega_g y}. \tag{23}$$

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 $-\sigma_\beta$ and $+\sigma_\beta$. 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).

$$\langle \psi_0, \psi_1 \rangle = \int dx\ e^{-\beta x^2} e^{-i\omega_0 x},\ \rightarrow\ e^{-\frac{(\omega_0)^2}{\beta}} \approx 0 \tag{24}$$

Using $2\pi f_0 = \frac{2\pi}{\lambda_0} = \omega_0$ the condition above implies $e^{-(\frac{2\pi\sigma_\beta)}{\lambda_0})^2} \approx 0$ which can be satisfied with $\sigma_\beta \geq \lambda_0$; $\sigma_\beta \sim \lambda_0$ is enough since this implies $e^{-(\frac{2\pi\sigma_\beta)}{\lambda_0})^2} \approx e^{-(2\pi)^2}$.
A similar condition ensures more in general orthogonality of any pair of eigenfunctions.

$$\int dx\ \psi_n^*(x)\psi_m(x) = \int dx e^{-\frac{x^2}{\sigma_\beta^2}} e^{in\omega_0 x} e^{-im\omega_0 x} \propto e^{-((m-n)\omega_0)^2\sigma_\beta^2},$$

which gives a similar condition as above. this also implies that λ should increase with σ 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:

$$t^{\circledast}(x, y) = \frac{1}{2\pi} \int d\omega_x d\omega_y t^{\circledast}(\omega_x, \omega_y) e^{i(\omega_x x + \omega_y y)} \tag{25}$$

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. 19 reads:

$$\frac{1}{2\pi} e^{-\frac{\alpha}{2}(x^2+y^2)} \int d\xi d\eta\ e^{-\frac{\xi^2(\alpha+\beta)}{2}} e^{-\frac{\eta^2(\alpha+\gamma)}{2}} \int d\omega_x d\omega_y\ t^{\circledast}(\omega_x, \omega_y) \tag{26}$$

$$e^{i\omega_x(x-\xi)} e^{-i\omega_y \eta} e^{i\omega_\xi^g \xi} = \lambda(\omega_x^g, \omega_y^g) e^{-\frac{\gamma}{2}y^2} e^{-\frac{\beta x^2}{2}} e^{i\omega_x^g x} \tag{27}$$

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 10. *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. [79]) 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,[15]. 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 [6]) simply correct *in the direction of motion* for the spatially low-pass components of the output of the retina (see Figure 14).
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

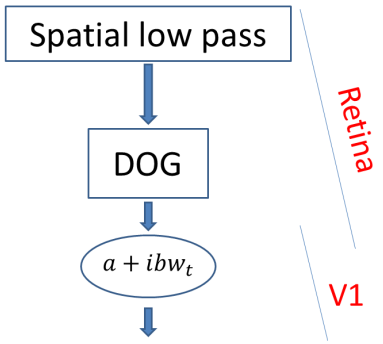


Figure 14: *The sequence of processing stage from the retina with spatial low-pass and bandpass (DOG) plus temporal d/dt derivative-like filtering to V1. Thus high-pass temporal filtering compensates for the spatial blurring in the direction of motion.*

signal generated by the translation of an image $f(x-vt)$, where x, v are vectors in $\setminus^2$:

$$\frac{\mathrm{d}\, I}{\mathrm{d}\, t} = \nabla I \cdot v. \tag{28}$$

assume for instance that the direction of motion is along the x axis, eg $v_y = 0$. Then

$$\frac{\mathrm{d}\, I}{\mathrm{d}\, t} = \frac{\partial I}{\partial x} v_x. \tag{29}$$

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 14. The 2D simulations are performed with this pipeline using the low-pass filter of Figure 15.

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 Appendix22). We have the following result for the solution of the 2D eigenfunction equation in the presence of oriented motion:

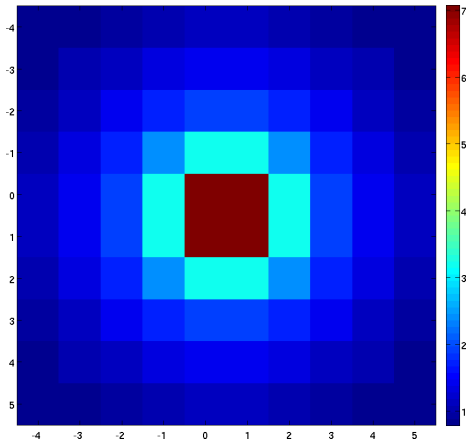


Figure 15: *Spatial lowpass filter* $1/\sqrt{\omega_x^2 + \omega_y^2}$ *as implemented in the 2D simulations.*

Lemma 2. 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 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 14 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 16 and 17 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 19) 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;

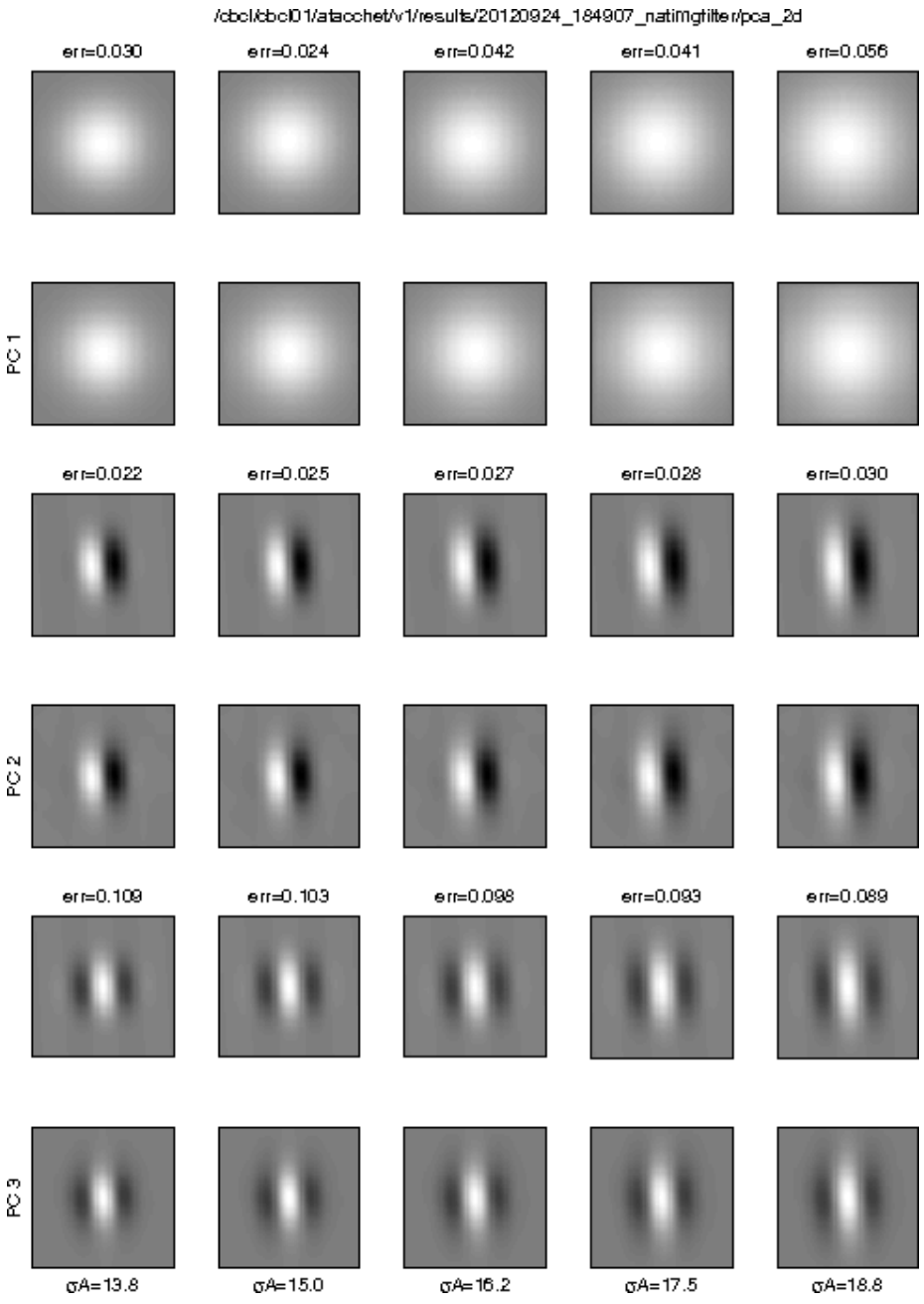


Figure 16: *For each row pair: the top row shows the Gabor fit, the bottom row shows the eigenvector.*

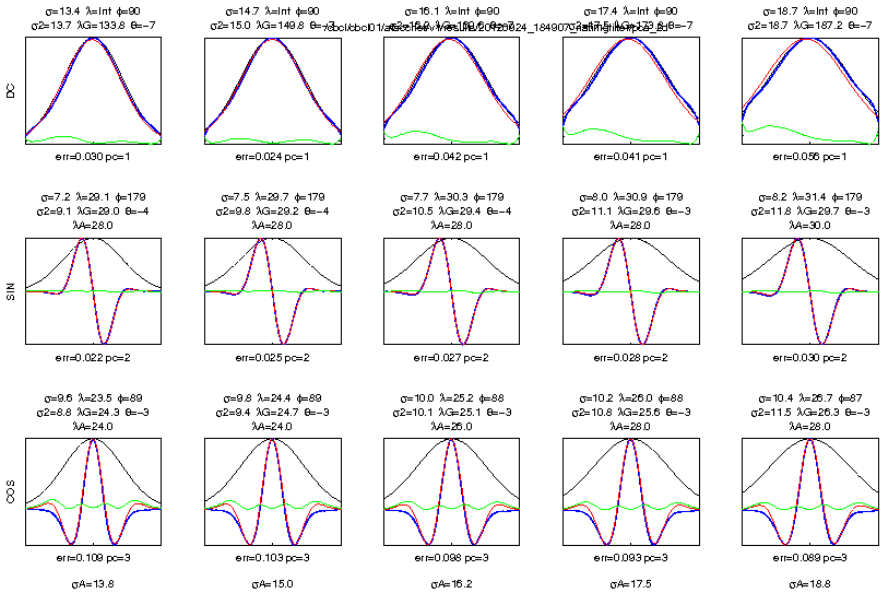


Figure 17: *1D sections of the principal components sorted by eigenvalue (row) for different Gaussian apertures (column). Red indicates best least square fit of a Gabor wavelet.*

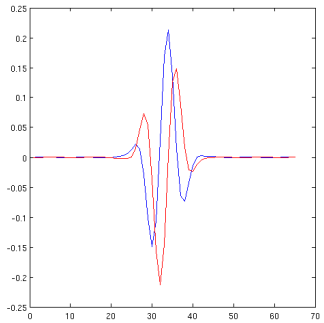


Figure 18: *A vertical slice through a quadrature pair (1st and 2nd eigenvector) from Figure 16*
.

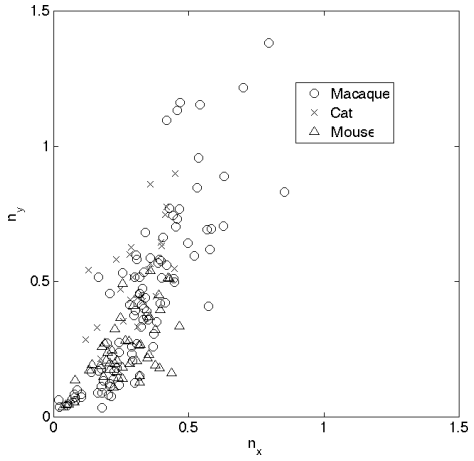


Figure 19: *Data from [25] (cat), [60] (macaque) and [46] (mouse). Here $n_x = \sigma_x f$ where σ_x is the standard deviation of the Gaussian envelope along the modulated axis and $f = \frac{2\pi}{\lambda}$ is the frequency of the Gabor's sinusoidal component. Likewise, $n_y = \sigma_y f$ where σ_y is the sigma of the Gaussian envelope along the unmodulated axis.*

- 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 21 shows that $\sigma_y > \sigma_x$. Figure 22 summarizes the main quantitative properties of the simulations. Figure 23 shows that the simulations seem to be consistent with 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 17).

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 24 and 25) 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 26), in which different natural images, randomly chosen, were used as input to the eigenvector calculation, instead of a motion sequence. It would be inter-

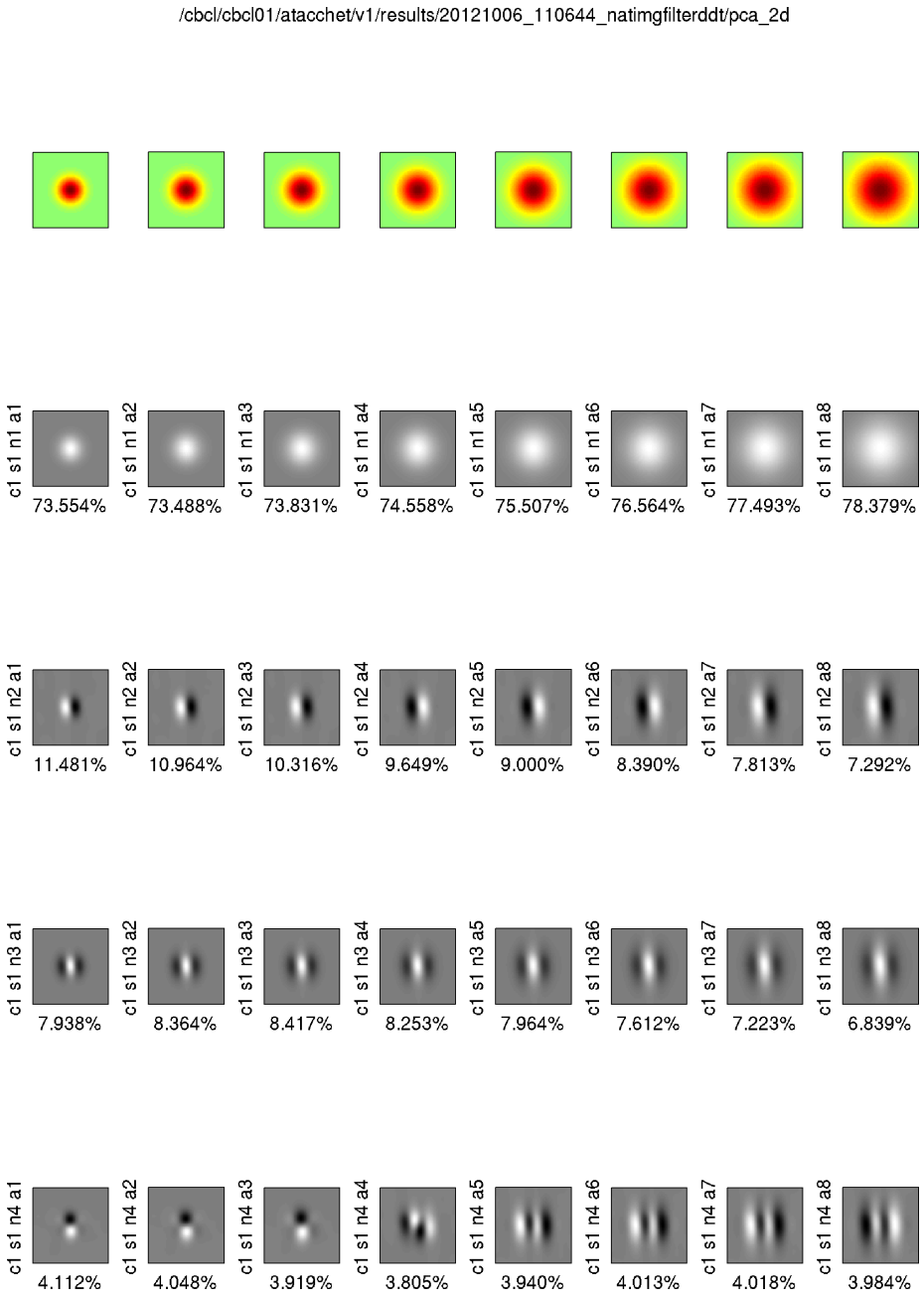


Figure 20: *Principal components of the template book. These are obtained observing* 40 *natural images translate through Gaussian apertures. The pipeline consists of a Gaussian blur, a DoG filter, a spatial low-pass filter* $1/\sqrt{\omega_x^2 + \omega_y^2}$ *and an imperfect temporal derivative.*

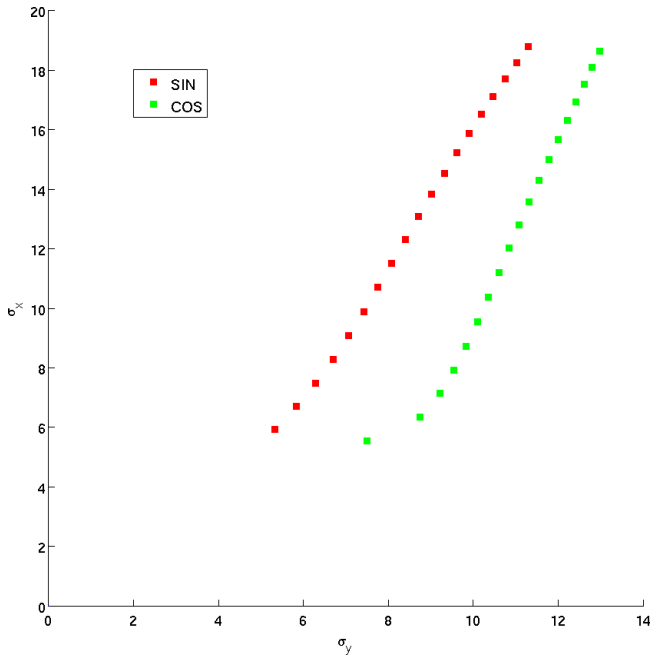


Figure 21: *Width of the Gaussian envelope for the modulated and unmodulated directions.*

esting 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 22) whether absence of motion implies that some λ go to infinity (see Figures 26 and 27). Physiology data (see Figure 19) 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 18).

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,

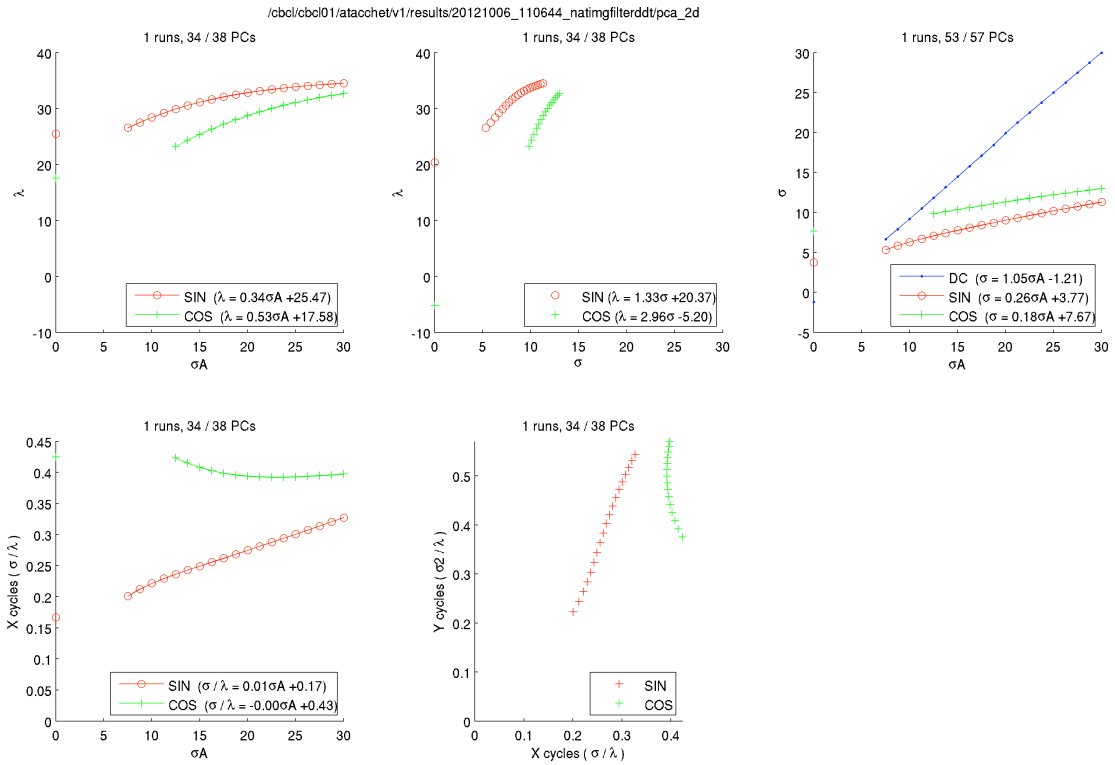


Figure 22: *Summary plots for 2D simulations. Figures from top left to bottom right: a) sinusoid wavelength (λ) vs. Gaussian aperture width (σ_α). b) Sinusoid wavelength (λ) vs. Gaussian envelope width on the modulated direction (σ). c) Gaussian envelope width for the modulated direction (σ) vs. Gaussian 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.*

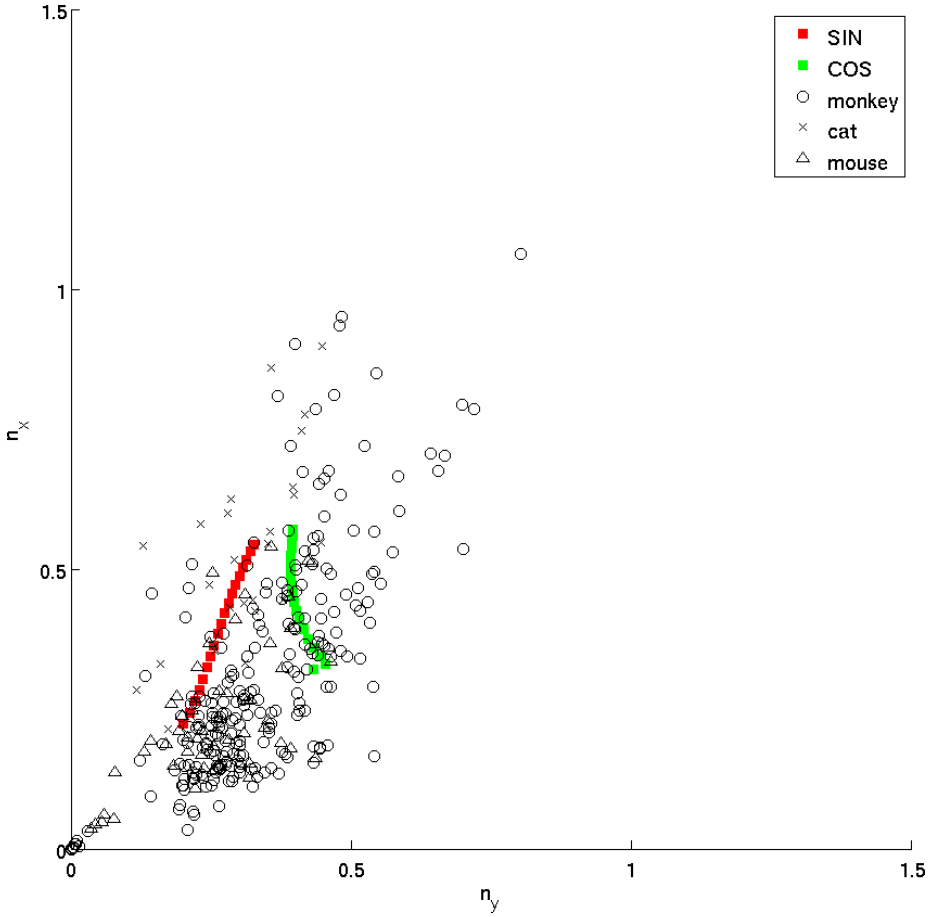


Figure 23: *This figure shows the σ/λ ratio for the modulated and unmodulated direction of the Gabor wavelet. Neurophysiology data from monkeys, cats and mice are reported together with our simulations*

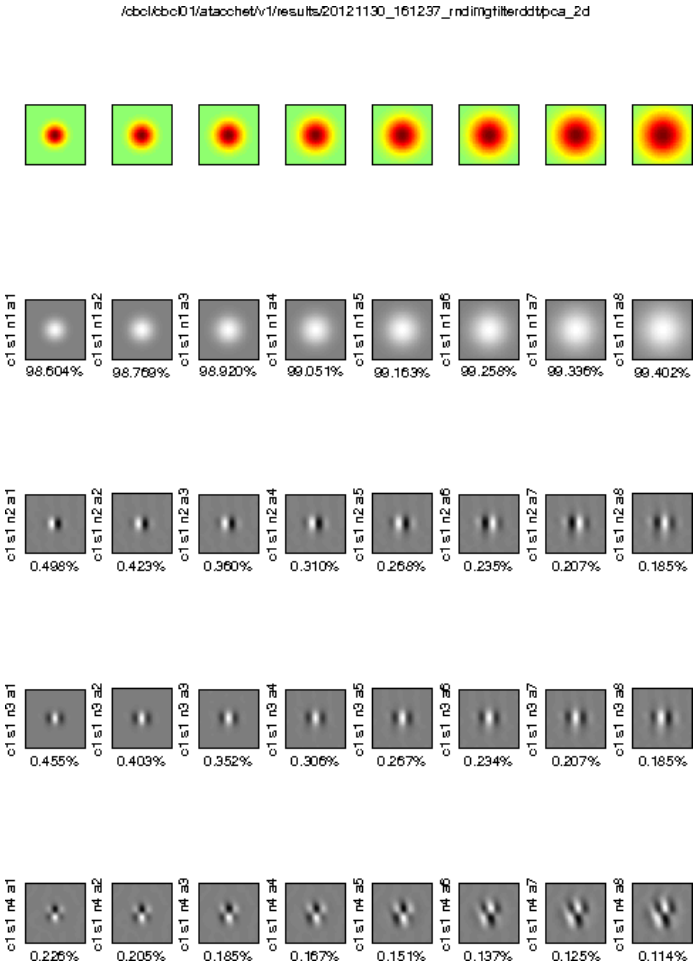


Figure 24: A white noise visual pattern is translated.

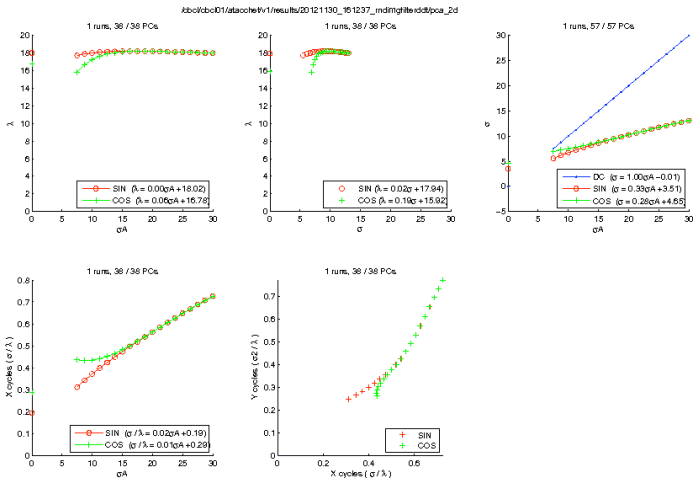


Figure 25: Properties of the eigenfunctions for translation of a white noise pattern.

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 [70]. 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 15) collects a few notes about transformations and 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 ([78, 10]) and made precise by Andreas Maurer.

- *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 11. *(Anselmi, Spigler, Poggio)*

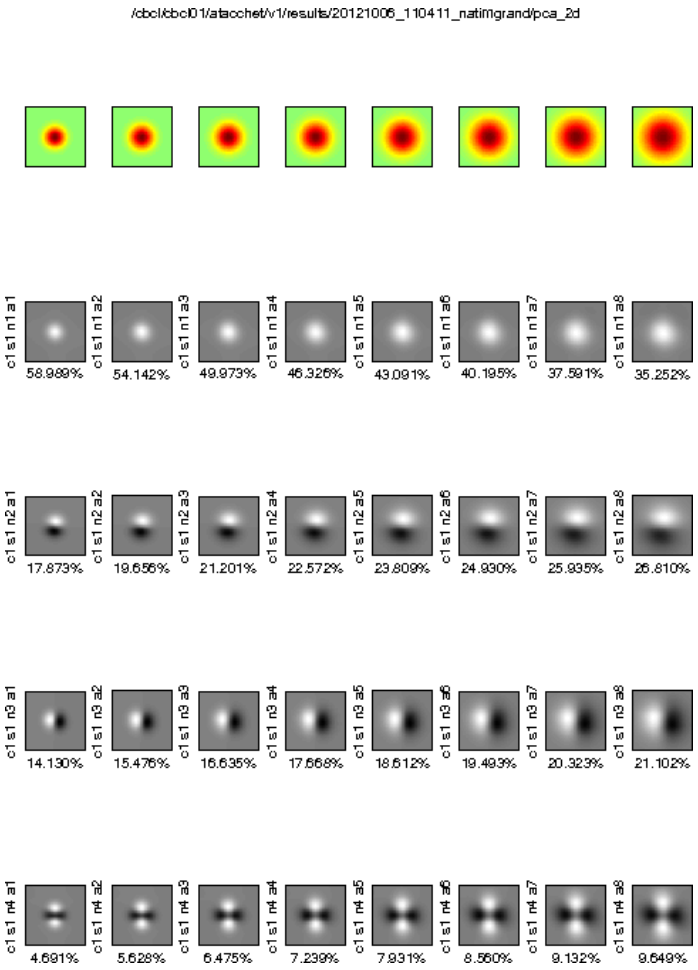


Figure 26: Eigenvectors of covariance matrix of scrambled set of images (same as in Figure 16 but scrambled). There is no continuous motion. The orientation of wavelets changes.

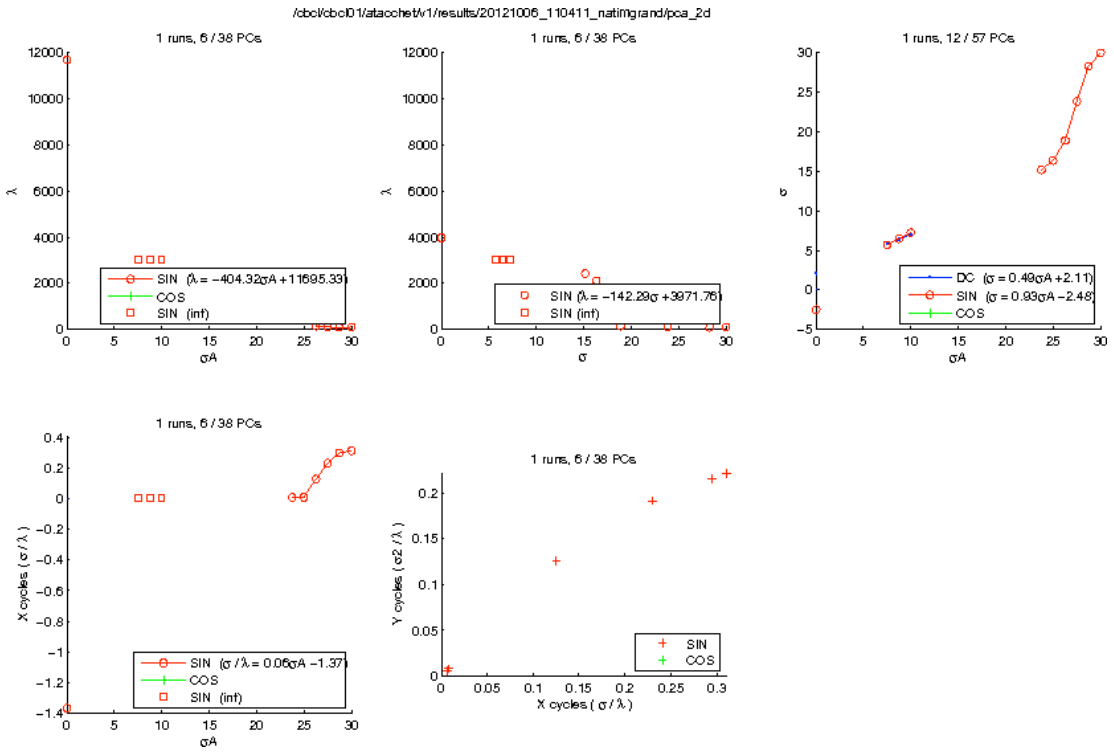


Figure 27: As in the previous Figure. λ can be infinite since orthogonality wrt to lower order eigenfunction is ensured by orthogonal orientation.

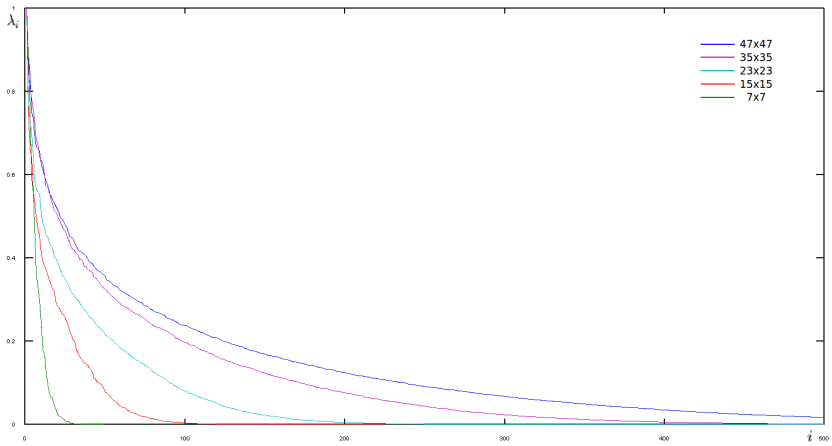


Figure 28: *Eigenvalues behavior as a function of the aperture for general Hilbert-Schmidt integral operators.*

- *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 with the first iteration of Mallat's scattering transform [42].

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 [42]. 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 ([42]). 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.

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 [18]), 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 5 (Affine Group)**.** *If $\mathcal{G}$ is the affine group and $\mathcal{X} = L^2(\mathbb{R})$, then under the admissibility condition*

$$\int |\langle T_g t, t \rangle|^2 < \infty,$$

[5] In fact it is easy to see that T is Hilbert Schmidt, $\mathrm{Tr}(T_t^* T_t) = \int dg \, \|T_g t\|$

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 9.4. 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 the 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 11.1)) – 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 11.1). 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 generates 4-dimensional wavelets. The prediction seems consistent with physiology data. Assuming that R is retinal eccentricity corresponds to assuming that most of the experienced and learned "rotations and looming 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 6. *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 7. *The affine group in dimension two...*

4.7.2 Independent shifts and commutators

(From [9])

Theorem 5. *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*

$$\mathcal{L}_a u = 1, \quad \mathcal{L}_b v = 0$$
$$\mathcal{L}_b u = 0, \quad \mathcal{L}_a v = 1$$

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 8. *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:

$$G = \mathbb{R}^2 \times \mathbb{S}_1 \times \mathbb{R} \tag{30}$$

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

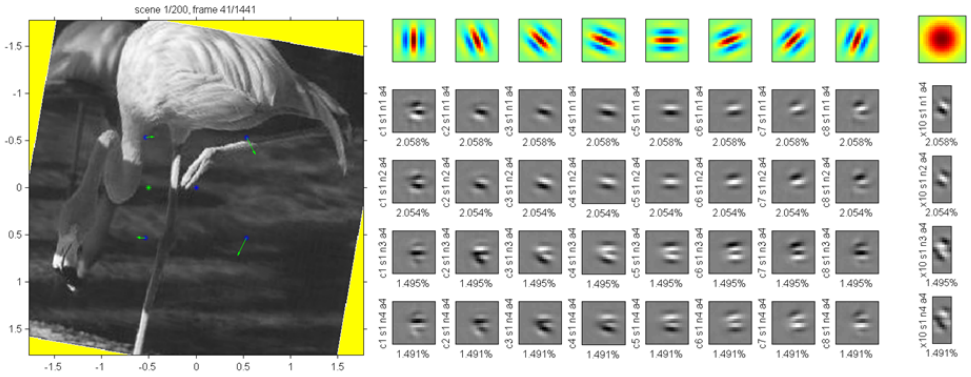


Figure 29: *Learning an S2 filter from C1 outputs (of a single scale only). Here the transformation is off-center rotation. The resulting S2 filters are Gabor filters in 3 dimensions: x; y, and orientation. Left: the receptive field center is in the middle (central blue asterisk) but the center of rotation 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).*

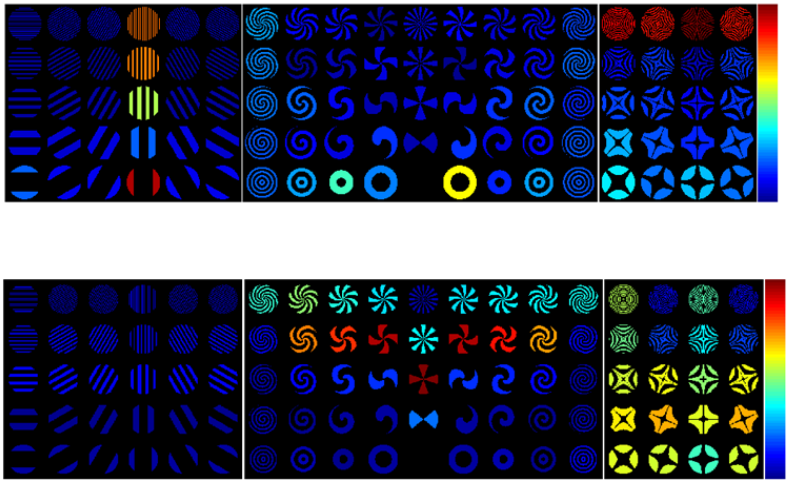


Figure 30: *The strength of the response of single cells in V4 is indicated by pseudocolor for each "non cartesian" pattern. Simulations compared to Gallant's data.*

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 in 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 [12]).

Remarks

- Mallat also considers wavelets of wavelets [42]. 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 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 [14].

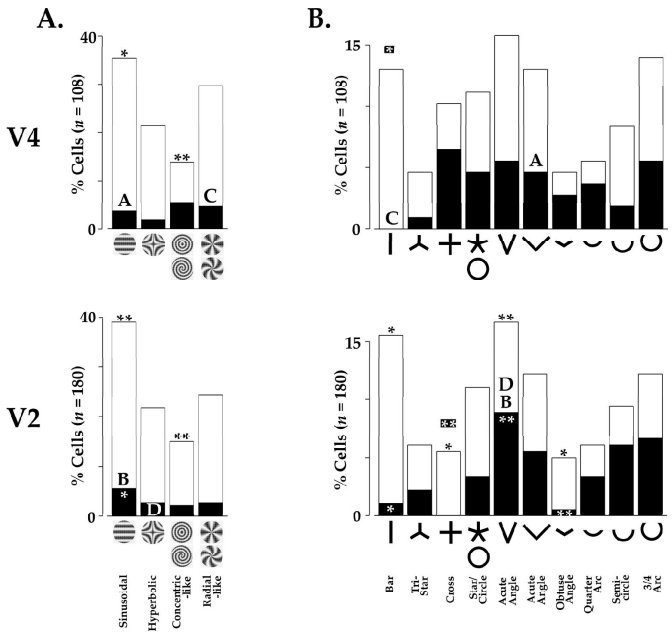


Figure 31: *The effectiveness of the various stimulus subclasses for V4 vs. V2. Each cell from either area was classified according the subclass to which its most effective grating or contour stimulus belonged. The resulting distributions are shown here for grating stimuli (panel A) or contour stimuli (panel B) for both V4 (top row) and V2. See [19].*

5 Part III: Class-specific transformations and modularity

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 it 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 is a function of 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 32). In homogenous coordinates this projection is given by

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{31}$$

In 3D homogenous coordinates a rotation around the y-axis is given by

$$R_\theta = \begin{pmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{32}$$

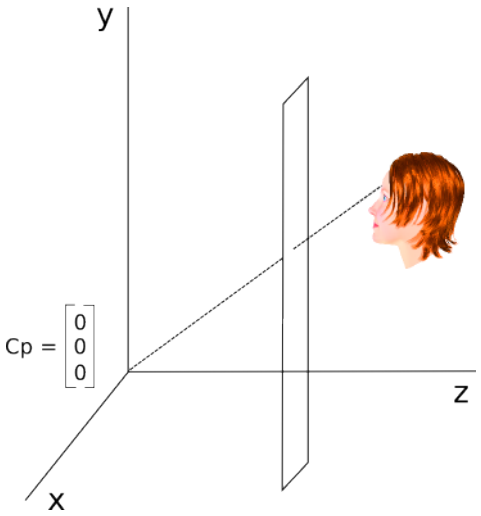


Figure 32: *Consider a situation where the center of projection Cp is at the origin in $\mathbb{R}^3$ and the projection is along the z-axis.*

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

$$PR_\theta X = \begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \\ 1 \end{pmatrix} \tag{33}$$

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.

$$t_{\theta,z} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \end{pmatrix} \tag{34}$$

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

$$\|t_{\theta,z}\begin{pmatrix}x\\y\end{pmatrix} - t_{\theta,z_\tau}\begin{pmatrix}x\\y\end{pmatrix}\| < (\epsilon\sin(\theta))^2 \tag{35}$$

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 $\forall v \in \mathbb{R}^2$:

$$\mathbb{P}_q[f](v) = \begin{cases} q^i & \text{if } \ v = Pf^i \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

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 $\vec{z} \in \mathbb{R}^N$.

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]] = \mathbb{P}_q[\{R_\theta g^i \ \ : i = 1, ..., N\}] \tag{37}$$

Where g^i is obtained by replacing the z-component of f^i with $\vec{z}^i$. Thus

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]] = \mathbb{P}_q[\{R_\theta f^i \ \ : i = 1, ..., N\}] \qquad \text{if } \ \vec{z}^i = \text{ the z-component of } f^i \ \ (\forall i) \tag{38}$$

$T_{\theta,\vec{z}}$ transforms individual points in the following way:

$$T_{\theta,\vec{z}}[\mathbb{P}_q[f]]\begin{pmatrix}x\\y\end{pmatrix} \ = \ \mathbb{P}_q[f]\begin{pmatrix} x\cos(\theta) + z\sin(\theta) \\ y \end{pmatrix} \tag{39}$$

We can bound the error arising from mapping the image using $\vec{z_\tau}$ obtained from a template object $\tau = \{(x_\tau^i, y_\tau^i, z_\tau^i, 1)^\intercal\}$—different from the test object f.

If $|z_\tau^i - z_f^i| < \epsilon \ (\forall i)$ then

$$\|T_{\theta,\vec{z_\tau}}[\mathbb{P}_q[f]] - \mathbb{P}_q[\{R_\theta f^i \ : i = 1, ..., N\}]\| \ < \ \Sigma_{i=1}^N |z_\tau^i \sin(\theta) - z_f^i \sin(\theta)|^2 \ = \ N(\epsilon \sin(\theta))^2 \tag{40}$$

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.

$$\Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[\tau]}(\mathbb{P}_q[f]) = \begin{pmatrix} < T_{-\pi,\vec{z_\tau}}[\mathbb{P}_q[\tau]] \ , \ \mathbb{P}_q[f] > \\ \vdots \\ < T_{\pi,\vec{z_\tau}}[\mathbb{P}_q[\tau]] \ , \ \mathbb{P}_q[f] > \end{pmatrix} \tag{41}$$

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. How close of an approximation it is depends on to what extent the template and test object share 3D structure. We have the following statement.

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

$$|z_\tau^i - z_f^i| < \epsilon \ (\forall i). \tag{42}$$

Then there exists a permutation function S such that

$$S(\Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[\tau]}(\mathbb{P}_q[f])) - \Delta_{T_{\theta,\vec{z_\tau}},\mathbb{P}_q[f]}(\mathbb{P}_q[\tau]) < N(\epsilon \sin(\theta))^2 \vec{1} \tag{43}$$

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). It does not impact empirical investigations of class-specific invariance.

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 33 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 34 and 35).

Figures 34 and 35 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.

The original HMAX model [66], 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 34I and 35I).

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 34I and 35I). When C3 features from the wrong category are used in this way, performance suffers (off-diagonal plots). In all

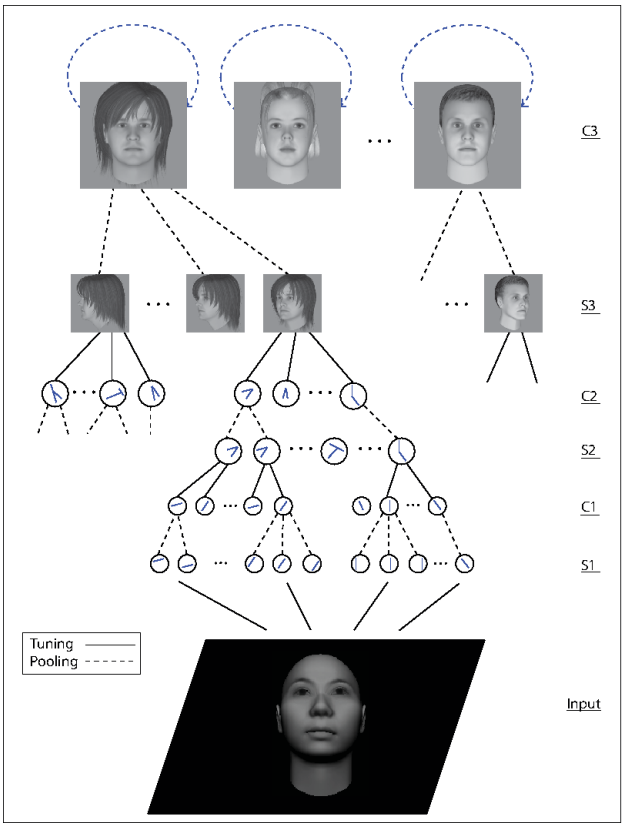


Figure 33: *Illustration of an extension to the HMAX model to incorporate class-specific invariance to face viewpoint changes. Note: All simulations with this model (Figures 34, 35) use a Gaussian radial basis function to compute the S2 and S3 layers as opposed to the normalized dot product that is used in its S1 layer and elsewhere in this report.*

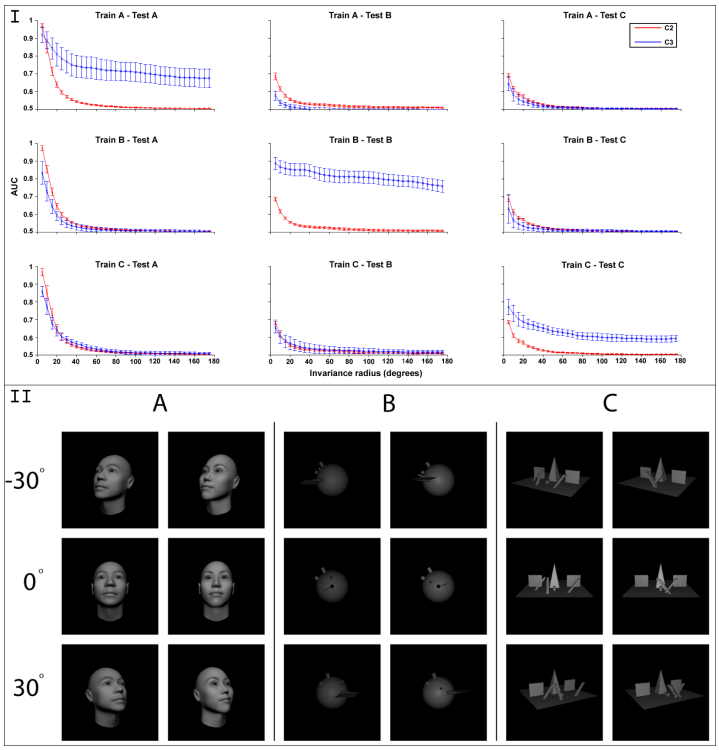


Figure 34: *Viewpoint invariance. Bottom panel (II): Example images from three classes of stimuli. Class A consists of faces produced using FaceGen (Singular Inversions). Class B is a set of synthetic objects produced using Blender (Stichting Blender Foundation). Each object in this class has a central 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.*

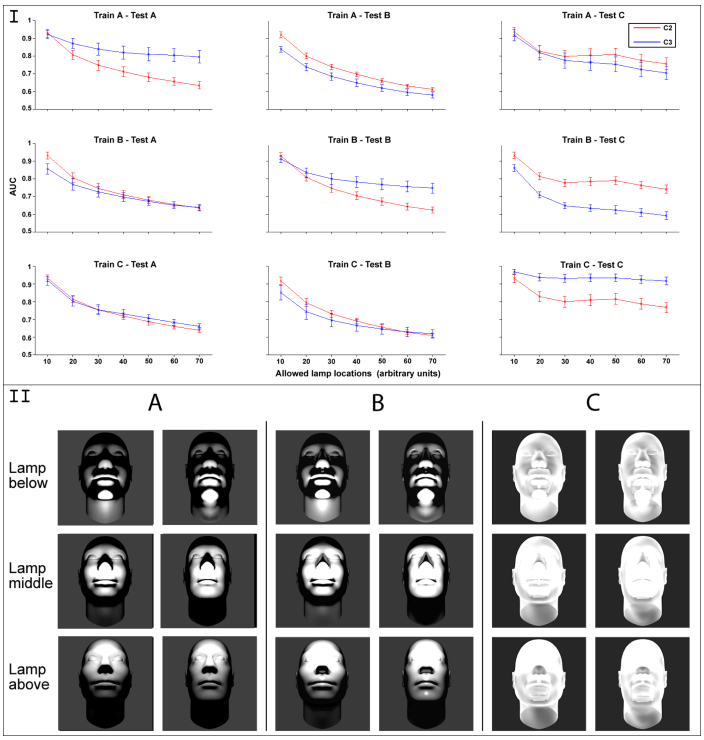


Figure 35: *Illumination invariance. Same organization as in figure 3. Bottom panel (II): Example images from three classes of stimuli. Each class consists of faces with different light reflectance properties, modeling different materials. Class A was opaque and non-reflective like wood. Class B was 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.*

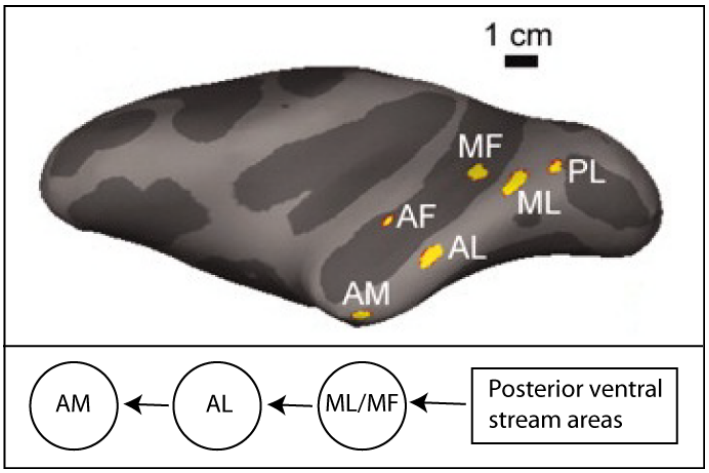


Figure 36: *Layout of face-selective regions in macaque visual cortex, adapted from [13] with permission.*

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 [7]) and reading (the visual word form area [4, 2]) 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 [71, 72]. 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 [13]. 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 [44]. In addition, spatial position invariance increases from ML/MF to AL, and increases further to AM [13] 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 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 37).

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.

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

[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.

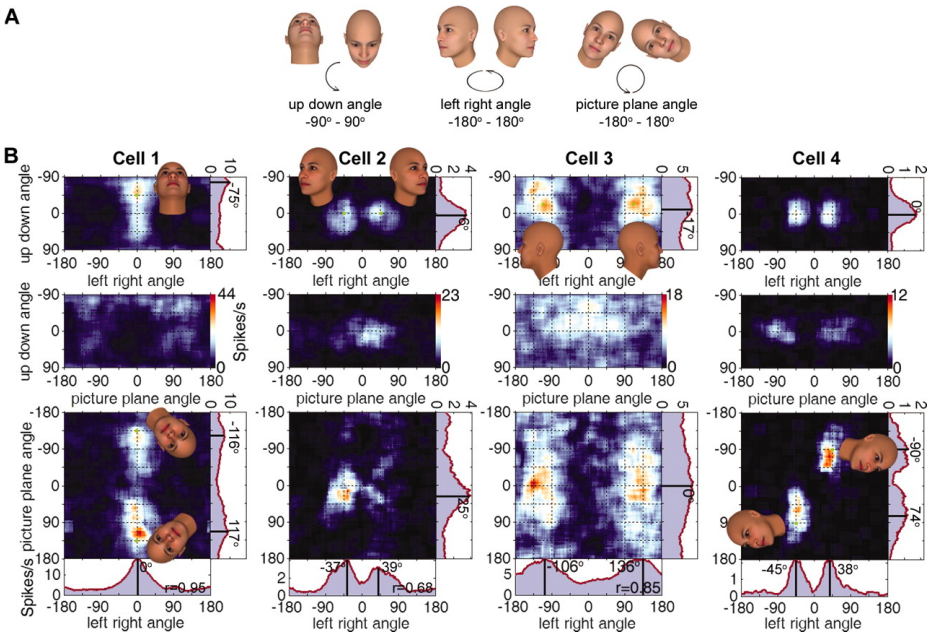


Figure 37: *Figure from [13]. Tuning of AL cells to a head randomly rotated in three dimensions. (A) Illustration of stimulus head and three axes of rotation. (B) View tuning in four typical cells. Tuning to up-down angle versus left-right angle (responses averaged across picture-plane angle). (Middle) 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).*

40).

Let R denote the reflection operator $R(\tau_{n,t}) = \tau_{n,-t}$. For simplicity, consider a templatebook that only contains one template τ and its reflection $R\tau$.

Let J denote the operator that takes a neural frame τ and returns the templatebook consisting of τ and its reflection.

$$J = (I, R) \tag{44}$$

$$J(\tau) = \mathbb{T} = \begin{pmatrix} \tau \\ R\tau \end{pmatrix} \tag{45}$$

$$\tag{46}$$

Thus

$$[J^\intercal J](R\tau) = \begin{pmatrix} R\tau + \tau \\ \tau + R\tau \end{pmatrix} = R([J^\intercal J](\tau)) \tag{47}$$

Therefore $R(J^\intercal J) = (J^\intercal J)R$, (they commute). Thus $J^\intercal J$ and R must have the same eigenfunctions. Since the eigenfunctions of R are even or odd functions, the principal components of $\mathbb{T}^\intercal\mathbb{T}$ must also be even and odd. Therefore, since we use the absolute value of the normalized dot product of the input with a template, both even and odd templates yield tuning curves that show identical tuning to symmetric face views.

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 38).

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

[7]Singular Inversions Inc.

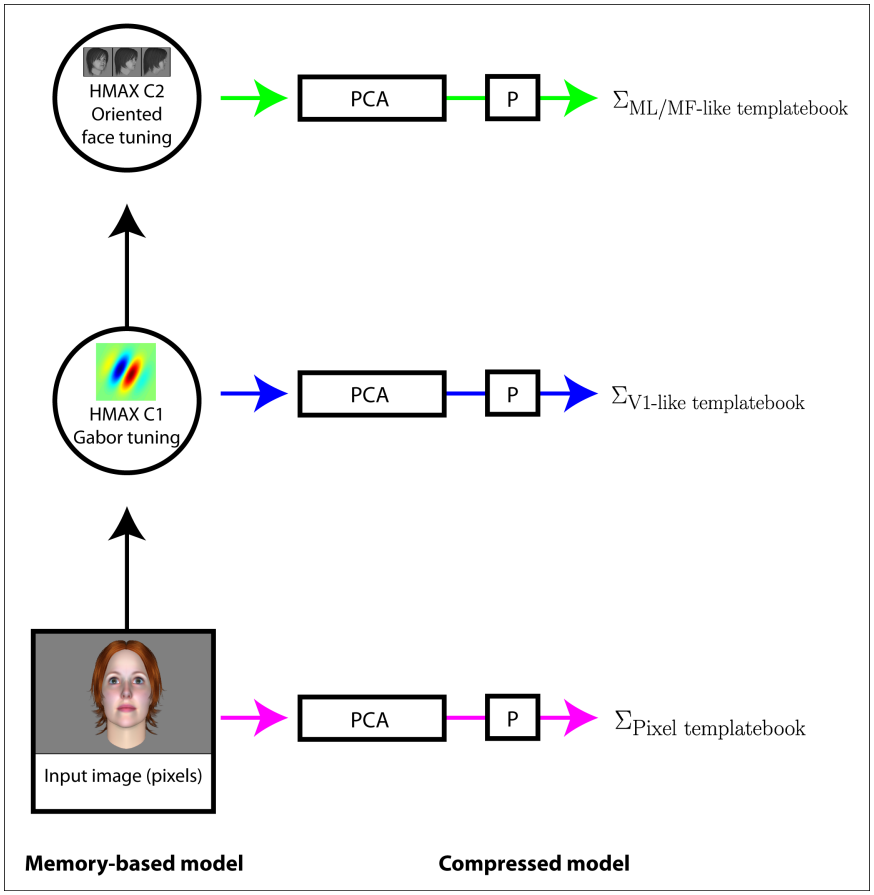


Figure 38: *Schematic of three possible models. Magenta: A model where the templatebooks were raw pixels with no preprocessing. Blue: A model where the templatebooks were encoded in an HMAX C1 layer (preprocessing with Gabor filtering and some limited pooling over position). Green: A model where the 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.*

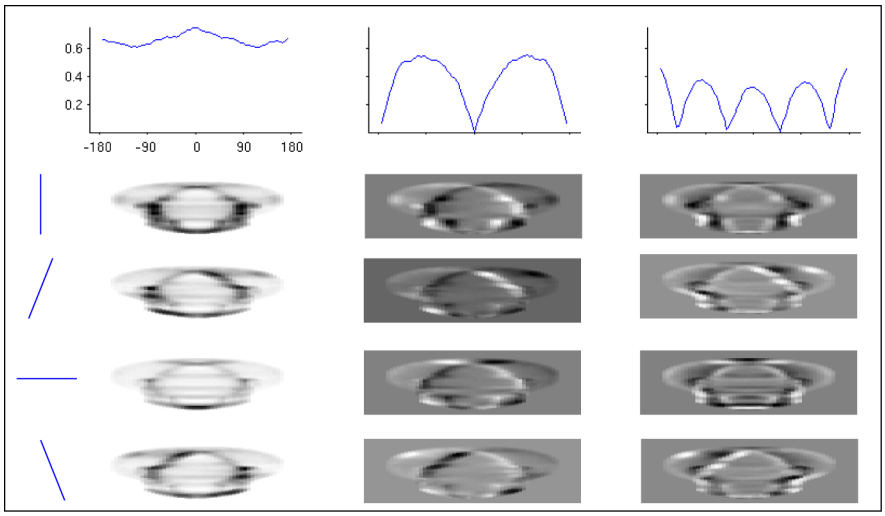


Figure 39: *Sample tuning curves and principal components for the model that encodes all inputs as HMAX C1 responses (the blue model in figures 38 and 41). Top row: the responses of S-PCA layer cells to systematically varying the orientation of a randomly-chosen test face. Below each tuning curve are 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.*

Blender[8]. Images of each face were rendered every 5 degrees, Each templatebook covered nearly the full range of orientations ($0 - 355°$).

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 41 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

[8]The Blender foundation

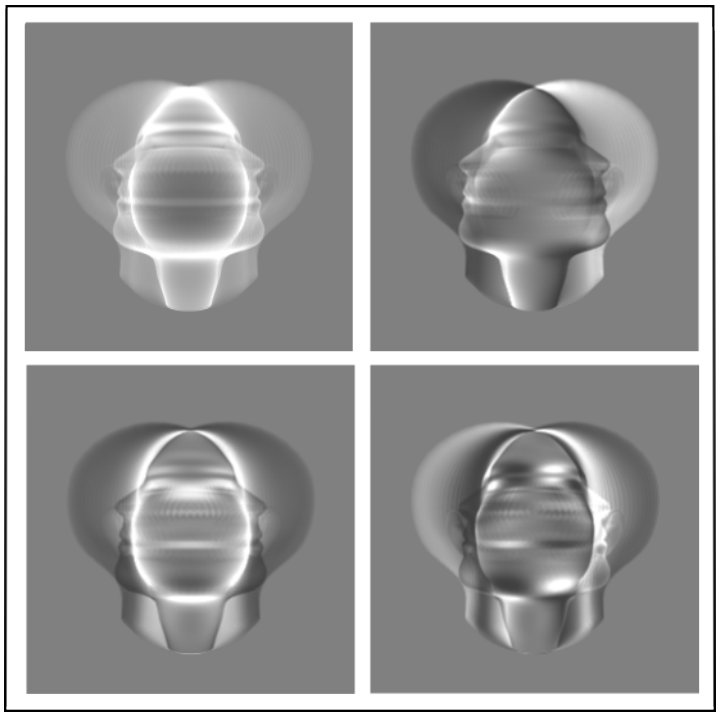


Figure 40: *More sample principal components. These were obtained from a model that does PCA directly on pixel inputs. They are the first 4 PCs obtained from the rotation of one head from* -90° *to* 90°.

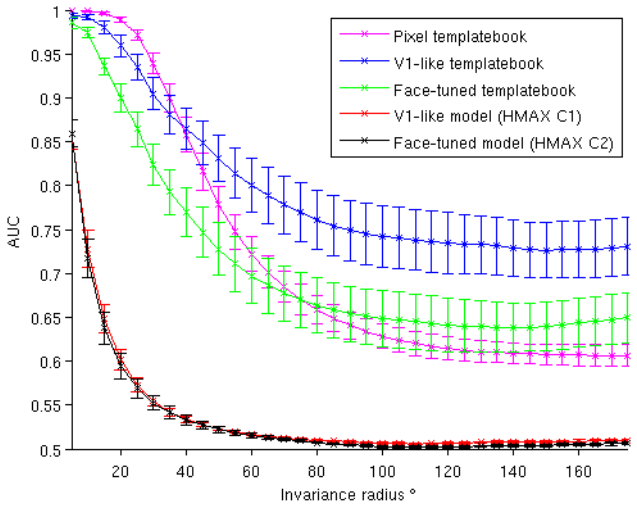


Figure 41: *Results from a test of viewpoint-invariant face identification. Test faces were presented on a black background. The task was to correctly categorize images by whether or not they depict the same person shown in a reference image—despite changes in viewpoint. This is a test of generalization 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 38. Red and black curves: Performance of the HMAX* $C1$ *and HMAX C2 layers on this task (included for comparison).*

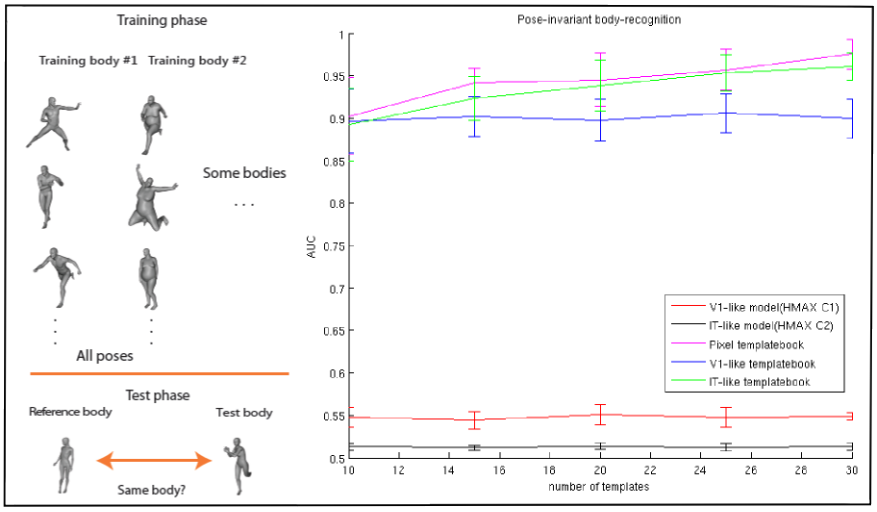


Figure 42: *Left: Images of human bodies in various poses were used to train and test the model. 1280 3D object models of human body were created with DAZ 3D Studio and one 256*256 pixel greyscale image was rendered from each object automatically with blender. The 1280 objects consisted of 40 differently 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) [8]—written words—the visual word form area (VWFA) [4], and bodies—the extrastriate body area (EBA) and the fusiform body area (FBA) [7, 50]. Many of these regions were shown to be necessary for recognition tasks with the objects they process by lesion studies ([34, 45]) and TMS ([75, 52]. 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: [27]).

Figures 42 and 43 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

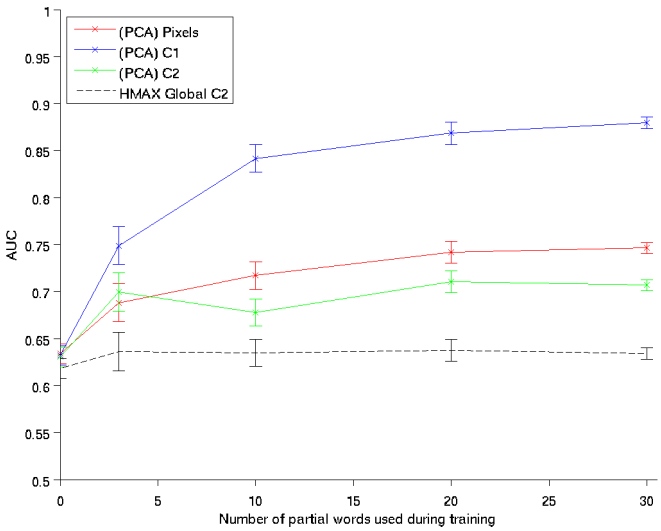


Figure 43: *Words (4-grams) were chosen from a fixed alphabet of 4 letters. A nearest-neighbor classifier ranked each image—of a word in a particular font—by its similarity to the image of a reference word. Templatebooks were obtained from translated and font-transformed images of single letters, bigrams 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 [22] and model simulations [64]. 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 44 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.

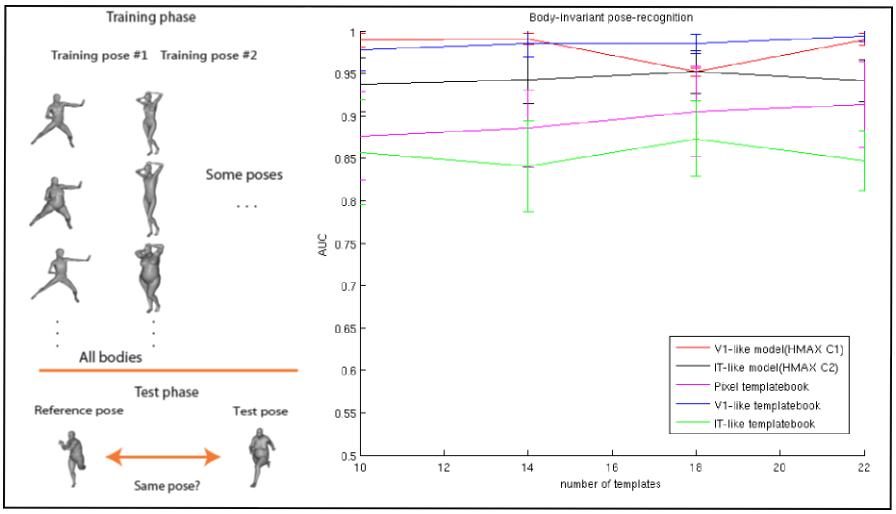


Figure 44: *Left: These simulations used the same images as the one in figure 42. Right: Performance of class-specific models and HMAX control models on a body-invariant pose recognition task. 10 poses were used for testing. The abscissa is the number of poses used to train the model. Performance was 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 [30]) 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. Another idea is that memory-based invariances determine the spectral properties of samples of transformed images and thus of a set of 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;
- 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

[9] There may be physiological evidence (from Motter and Poggio) suggesting invariance of several minutes of arc at the level of V1 and above.

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:

[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.

- "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.

- 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 the *signature vector* is input to an associative memory so that 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.

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 [49, 54]). 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, 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 Visual abstractions

- *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 abstract the concept of parallel lines from a single specific example of two parallel lines*. The argument is that according to the theorems in the paper, the signature of the single image of the parallel lines will be invariant to any affine transformations.

- *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.

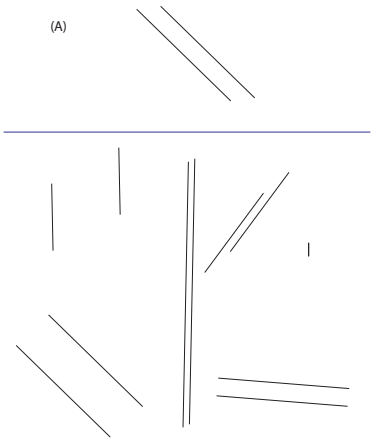


Figure 45: For a system which is invariant to affine transformations a single training example (A) allows recognition of all other instances of parallel lines – never seen before.

.

6.4.3 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.

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 a' la 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).

6.4.4 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.

6.4.5 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 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] J. Antoine, R. Murenzi, P. Vandergheynst, and S. Ali. *Two-dimensional wavelets and their relatives*. Cambridge University Press, Cambridge, UK, 2004.

[2] 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.

[3] Y. Bengio and Y. LeCun. Scaling learning algorithms towards ai. *Large-Scale Kernel Machines*, 34, 2007.

[4] L. Cohen, S. Dehaene, and L. Naccache. The visual word form area. *Brain*, 123(2):291, 2000.

[5] D. Cox, P. Meier, N. Oertelt, and J. DiCarlo. 'Breaking'position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–1147, 2005.

[6] 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.

[7] P. Downing and Y. Jiang. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470, 2001.

[8] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[9] 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.

[10] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

[11] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, 23(9):2289–2323, 2011.

[12] J. Freeman and E. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14:11951201, 2011.

[13] W. Freiwald and D. Tsao. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, 330(6005):845, 2010.

[14] J. Gallant, J. Braun, and D. V. Essen. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, 1993.

[15] L. Galli and L. Maffei. Spontaneous impulse activity of rat retinal ganglion cells in prenatal life. *Science (New York, NY)*, 242(4875):90, 1988.

[16] L. Glass and R. Perez. Perception of Random Dot Interference Patterns. *Nature*, 31(246):360–362, 1973.

[17] K. Groechenig. Multivariate gabor frames and sampling of entire functions of several variables. *Appl. Comp. Harm. Anal.*, pages 218 – 227, 2011.

[18] 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.

[19] J. Hegde and D. Van Essen. Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, 20(5):61, 2000.

[20] G. Hinton and R. Memisevic. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 22:14731492, 2010.

[21] W. Hoffman. The Lie algebra of visual perception. *Journal of Mathematical Psychology*, 3(1):65–98, 1966.

[22] 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.

[23] A. Hyvrinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[24] 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.

[25] 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.

[26] J. Karhunen. Stability of Oja's PCA subspace rule. *Neural Comput.*, 6:739–747, July 1994.

[27] 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.

[28] J. Koenderink. The brain a geometry engine. *Psychological Research*, 52(2):122–127, 1990.

[29] M. Kouh and T. Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural computation*, 20(6):1427–1451, 2008.

[30] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[31] 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.

[32] Y. LeCun. Learning invariant feature hierarchies. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 496–505. Springer, 2012.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] J. Z. Leibo, J. Mutch, S. Ullman, and T. Poggio. From primal templates to invariant recognition. *MIT-CSAIL-TR-2010-057, CBCL-293*, 2010.

[41] N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–7, Sept. 2008.

[42] 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.

[43] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[44] 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.

[45] 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.

[46] 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.

[47] C. Niell and M. Stryker. Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience*, 28(30):7520–7536, 2008.

[48] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[49] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992.

[50] G. Palm. On associative memory. *Biological Cybernetics*, 36(1):19–31, 1980.

[51] M. Peelen and P. Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005.

[52] 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.

[53] 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.

[54] 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.

[55] T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19(4):201–209, 1975.

[56] T. Poggio. The computational magic of the ventral stream: Supplementary Material. *CBCL Internal Memo*, 2011.

[57] 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.

[58] R. Rao and D. Ruderman. Learning Lie groups for invariant visual perception. *Advances in neural information processing systems*, pages 810–816, 1999.

[59] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, Nov. 1999.

[60] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3(11), 2000.

[61] D. Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1):455–463, 2002.

[62] D. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1994.

[63] T. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.

[64] A. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Ng. On random weights and unsupervised feature learning. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1089–1096, New York, NY, USA, June 2011. ACM.

[65] 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.

[66] 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.

[67] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007.

[68] C. F. Stevens. Preserving properties of object shape by computations in primary visual cortex. *PNAS*, 101(11):15524–15529, 2004.

[69] G. Strang and N. Troug. *Wavelets and filter Banks* . welleslay*cambridge press, wellesley, ma, 1996.

[70] S. Stringer and E. Rolls. Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation*, 14(11):2585–2596, 2002.

[71] A. Torralba and A. Oliva. Statistics of natural image categories. In *Network: Computation in Neural Systems*, pages 391–412, 2003.

[72] D. Tsao and W. Freiwald. Faces and objects in macaque cerebral cortex. *Nature . . .* , 6(9):989–995, 2003.

[73] D. Tsao, W. Freiwald, and R. Tootell. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670, 2006.

[74] G. Turrigiano and S. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2):97–107, 2004.

[75] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 992–1006, 1991.

[76] C. Urgesi, G. Berlucchi, and S. Aglioti. Magnetic stimulation of extrastriate body area impairs visual processing of nonfacial body parts. *Current Biology*, 14(23):2130–2134, 2004.

[77] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*, pages 0–61, 2011.

[78] 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.

[79] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

[80] R. Wong, M. Meister, and C. Shatz. Transient period of correlated bursting activity during development of the mammalian retina. *Neuron*, 11(5):923–938, 1993.

[81] G. Yu and J.-M. Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*, 2011.